YOU ARE NOT THE API I USED TO KNOW.

4. READER

Reading Salon

Tuesday 25 June

Location:
Rooms: 0.04 (Salon #1) and 0.013 (Salon #2)
University of Amsterdam, Media Studies
Turfdraagsterpad 9
1012 XT Amsterdam

Location:
Rooms: E 012 OHMP (Salon #3) and EK 0.2 OMHP (Salon #4)
Oudemanhuispoort 4-6
1012 CN Amsterdam

Table of Contents

Puschmann, Cornelius, and Jean Burgess. 2013. “The Politics of Twitter Data”. SSRN Scholarly Paper ID 2206225. Rochester, NY: Social Science Research Network.

Rieder, Bernhard. 2012. “The Refraction Chamber: Twitter as Sphere and Network.” First Monday 17 (11) (November 4). Twitter data.

3. Reading Salon #3: We Take All (Network) Shapes and Sizes

Elmer, Greg, and Ganaele Langlois. 2013. “Networked Campaigns: Traf@ic Tags and Cross Platform Analysis on the Web.” Information Polity 18 (1) (January 1): 43–56.

High@ield, Tim. 2012. “Talking of Many Things: Using Topical Networks to Study Discussions in Social Media.” Journal of Technology in Human Services 30 (3-4): 204–218.

Manjoo, Farhad. 2012. “The End of the Echo Chamber.” Slate, January 17.

Ruppert, E, J Law, and M Savage. “Reassembling Social Science Methods: the Challenge of Digital Devices.” Theory, Culture & Society (May 14, 2013).

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The Role of Social Networks in Information Diffusion. Proceedings of the 21st International Conference on the World Wide Web (WWW ’12) (pp. 1–10). New York, New York, USA: ACM Press.

4. Reading Salon #4: Known Knowns (More or Less)

Borra, Erik, and Ingmar Weber. 2012. “Political Insights: Exploring Partisanship in Web Search Queries.” First Monday 17 (7) (June 23). 9 months of Yahoo!’s US web search query logs.

boyd, D.M. & Ellison, N.B., 2008. Social network sites: DeXinition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), pp. 210–230.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” Proceedings of the National Academy of Sciences (March 11).

Rieder, B., 2013. Studying Facebook via Data Extraction: The Netvizz Application. Proceedings of ACM Web Science 2013, Paris, May 2-4.

Rogers, Richard. 2009. “Post-demographic Machines.”

Zimmer, M., 2010. “But the data is already public”: on the ethics of research in Facebook. *Ethics and information technology*, 12(4), pp.313–325.

Reading Salon #1: Bigger, Faster, Lighter

Moderators: Carolin Gerlitz and Bernhard Rieder

boyd, danah, and Kate Crawford. 2011. “Six Provocations for Big Data”. SSRN Scholarly Paper ID 1926431. Rochester, NY: Social Science Research Network.

Dash, Anil. 2012. “The Web We Lost.” December 13.

Elmer, Greg. 2013. “Live Research: Twittering an Election Debate.” New Media & Society 15 (1) (February 1): 18–30. doi: 10.1177/1461444812457328.

Gerlitz, Carolin and Bernhard Rieder. 2013. Mining One Percent of Twitter: Collections, Baselines, Sampling. M/C Journal 16 (2).

Giglietto, Fabio, Luca Rossi, and Davide Bennato. 2012. “The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source.” Journal of Technology in Human Services 30 (3-4): 145–159.

Ramsay, S. 2010. “The Hermeneutics of Screwing Around; or What You Do with a Million Books.” Unpublished Presentation Delivered at Brown University, Providence, RI 17.

Paper to be presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.

Six Provocations for Big Data

danah boyd
Microsoft Research
dmb@microsoft.com

Kate Crawford
University of New South Wales
k.crawford@unsw.edu.au

Technology is neither good nor bad; nor is it neutral...technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves.

Melvin Kranzberg (1986, p. 545)

We need to open a discourse – where there is no effective discourse now – about the varying temporalities, spatialities and materialities that we might represent in our databases, with a view to designing for maximum flexibility and allowing as possible for an emergent polyphony and polychrony. Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care.

Geoffrey Bowker (2005, p. 183-184)

The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and many others are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing information from Twitter, Google, Verizon, 23andMe, Facebook, Wikipedia, and every space where large groups of people leave digital traces and deposit data. Significant questions emerge. Will large-scale analysis of DNA help cure diseases? Or will it usher in a new wave of medical inequality? Will data analytics help make people's access to information more efficient and effective? Or will it be used to track protesters in the streets of major cities? Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means? Some or all of the above?

Big Data is, in many ways, a poor term. As Lev Manovich (2011) observes, it has been used in the sciences to refer to data sets large enough to require supercomputers, although now vast sets of data can be analyzed on desktop computers with standard software. There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem. Big Data is notable not because of its size, but because of its relationality to other data. Due to efforts to mine and aggregate data, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.

Furthermore, Big Data is important because it refers to an analytic phenomenon playing out in academia and industry. Rather than suggesting a new term, we are using Big Data here because of its popular salience and because it is the phenomenon around Big Data that we want to address. Big Data tempts some researchers to believe that they can see everything at a 30,000-foot view. It is the kind of data that encourages the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions. Due to this, it is crucial to begin asking questions about the analytic assumptions, methodological frameworks, and underlying biases embedded in the Big Data phenomenon.

While databases have been aggregating data for over a century, Big Data is no longer just the domain of actuaries and scientists. New technologies have made it possible for a wide range of people – including humanities and social science academics, marketers, governmental organizations, educational institutions, and motivated individuals – to produce, share, interact with, and organize data. Massive data sets that were once obscure and distinct are being aggregated and made easily accessible. Data is increasingly digital air: the oxygen we breathe and the carbon dioxide that we exhale. It can be a source of both sustenance and pollution.

How we handle the emergence of an era of Big Data is critical: while it is taking place in an environment of uncertainty and rapid change, current decisions will have considerable impact in the future. With the increased automation of data collection and analysis – as well as algorithms that can extract and inform us of massive patterns in human behavior – it is necessary to ask which systems are driving these practices, and which are regulating them. In *Code*, Lawrence Lessig (1999) argues that systems are regulated by four forces: the market, the law, social norms, and architecture – or, in the case of technology, code. When it comes to Big Data, these four forces are at work and, frequently, at odds. The market sees Big Data as pure opportunity: marketers use it to target advertising, insurance providers want to optimize their offerings, and Wall Street bankers use it to read better readings on market temperament. Legislation has already been proposed to curb the collection and retention of data, usually over concerns about privacy (for example, the Do Not Track Online Act of 2011 in the United States). Features like personalization allow rapid access to more relevant information, but they present difficult ethical questions and fragment the public in problematic ways (Pariser 2011).

There are some significant and insightful studies currently being done that draw on Big Data methodologies, particularly studies of practices in social network sites like Facebook and Twitter. Yet, it is imperative that we begin asking critical questions about what all this data means, who gets access to it, how it is deployed, and to what ends. With Big Data come big responsibilities. In this essay, we are offering six provocations that we hope can spark conversations about the issues of Big Data. Social and cultural researchers have a stake in the computational culture of Big Data precisely because many of its central questions are fundamental to our disciplines. Thus, we believe that it is time to start critically interrogating this phenomenon, its assumptions, and its biases.

1. Automating Research Changes the Definition of Knowledge.

In the early decades of the 20th century, Henry Ford devised a manufacturing system of mass production, using specialized machinery and standardized products. Simultaneously, it became the dominant vision of technological progress. Fordism meant automation and assembly lines, and for decades onward, this became the orthodoxy of manufacturing: out with skilled craftspeople and slow work, in with a new machine-made era (Baca 2004). But it was more than just a new set of tools. The 20th century was marked by Fordism at a cellular level: it produced a new understanding of labor, the human relationship to work, and society at large.

Big Data not only refers to very large data sets and the tools and procedures used to manipulate and analyze them, but also to a *computational turn* in thought and research (Burkholder 1992). Just as Ford changed the way we made cars – and then transformed work itself – Big Data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community. 'Change the instruments, and you will change the entire social theory that goes with them,' Latour reminds us (2009, p. 9).

We would argue that Bit Data creates a radical shift in how we think about research. Commenting on computational social science, Lazer *et al* argue that it offers 'the capacity to collect and analyze data with an unprecedented breadth and depth and scale' (2009, p. 722). But it is not just a matter of scale. Neither is enough to consider it in terms of proximity, or what Moretti (2007) refers to as distant or close analysis of texts. Rather, it is a profound change at the levels of epistemology and ethics. It reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality. Just as du Gay and Pryke note that 'accounting tools...do not simply aid the measurement of economic activity, they shape the reality they measure' (2002, pp. 12-13), so Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life.

Speaking in praise of what he terms 'The Petabyte Age', Chris Anderson, Editor-in-Chief of *Wired*, writes:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (2008)

Do numbers speak for themselves? The answer, we think, is a resounding 'no'. Significantly, Anderson's sweeping dismissal of all other theories and disciplines is a tell: it reveals an arrogant undercurrent in many Big Data debates where all other forms of analysis can be sidelined by production lines of numbers, privileged as having a direct line to raw knowledge. Why people do things, write things, or make things is erased by the sheer volume of numerical repetition and large patterns. This is not a space for reflection or the older forms of intellectual craft. As David Berry (2011, p. 8) writes, Big Data provides 'destablising amounts of knowledge and information that lack the regulating force of philosophy.' Instead of philosophy – which Kant saw as the rational basis for all institutions – 'computationality might then be understood as an ontotheology, creating a new ontological "epoch" as a new historical constellation of intelligibility' (Berry 2011, p. 12).

We must ask difficult questions of Big Data's models of intelligibility before they crystallize into new orthodoxies. If we return to Ford, his innovation was using the assembly line to break down interconnected, holistic tasks into simple, atomized, mechanistic ones. He did this by designing specialized tools that strongly predetermined and limited the action of the worker. Similarly, the specialized tools of Big Data also have their own inbuilt limitations and restrictions. One is the issue of time. 'Big Data is about exactly right now, with no historical context that is predictive,' observes Joi Ito, the director of the MIT Media Lab (Bollier 2010, p. 19). For example, Twitter and Facebook are examples of Big Data sources that offer very poor archiving and search functions, where researchers are much more likely to focus on something in the present or immediate past – tracking reactions to an election, TV finale or natural disaster – because of the sheer difficulty or impossibility of accessing older data.

If we are observing the automation of particular kinds of research functions, then we must consider the inbuilt flaws of the machine tools. It is not enough to simply ask, as Anderson suggests 'what can science learn from Google?', but to ask how Google and the other harvesters of Big Data might change the *meaning* of learning, and what new possibilities and new limitations may come with these systems of knowing.

2. Claims to Objectivity and Accuracy are Misleading

'Numbers, numbers, numbers,' writes Latour (2010). 'Sociology has been obsessed by the goal of becoming a quantitative science.' Yet sociology has never reached this goal, in Latour's view, because of where it draws the line between what is and is not quantifiable knowledge in the social domain.

Big Data offers the humanistic disciplines a new way to claim the status of quantitative science and objective method. It makes many more social spaces quantifiable. In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth – particularly when considering messages from social media sites. But there remains a mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing facts. In this way, Big Data risks reinscribing established divisions in the long running debates about scientific method.

The notion of objectivity has been a central question for the philosophy of science and early debates about the scientific method (Durkheim 1895). Claims to objectivity suggest an adherence to the sphere of objects, to things as they exist in and for themselves. Subjectivity, on the other hand, is viewed with suspicion, colored as it is with various forms of individual and social conditioning. The scientific method attempts to remove itself from the subjective domain through the application of a dispassionate process whereby hypotheses are proposed and tested, eventually resulting in improvements in knowledge. Nonetheless, claims to objectivity are necessarily made by subjects and are based on subjective observations and choices.

All researchers are interpreters of data. As Lisa Gitelman (2011) observes, data needs to be imagined as data in the first instance, and this process of the imagination of data entails an interpretative base: 'every discipline and disciplinary institution has its own norms and standards for the imagination of data.' As computational scientists have started engaging in acts of social science, there is a tendency to claim their work as the business of facts and not interpretation. A model may be mathematically sound, an experiment may seem valid, but as soon as a researcher seeks to understand what it means, the process of interpretation has begun. The design decisions that determine what will be measured also stem from interpretation.

For example, in the case of social media data, there is a 'data cleaning' process: making decisions about what attributes and variables will be counted, and which will be ignored. This process is inherently subjective. As Bollier explains,

> As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned?' (2010, p. 13)

In addition to this question, there is the issue of data errors. Large data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together. Social scientists have a long history of asking critical questions about the collection of data and trying to account for any biases in their data (Cain & Finch, 1981; Clifford & Marcus, 1986). This requires understanding the properties and limits of a dataset, regardless of its size. A dataset may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a dataset, we need to know where data is coming from; it is similarly important to know and account for the weaknesses in that data. Furthermore, researchers must be able to account for the biases in their interpretation of the data. To do so requires recognizing that one's identity and perspective informs one's analysis (Behar & Gordon, 1996).

Spectacular errors can emerge when researchers try to build social science findings into technological systems. A classic example arose when Friendster chose to implement Robin Dunbar's (1998) work. Analyzing gossip practices in humans and grooming habits in monkeys, Dunbar found that people could only actively maintain 150 relationships at any time and argued that this number represented the maximum size of a person's personal network. Unfortunately, Friendster believed that people were replicating their pre-existing personal networks on the site, so they inferred that no one should have a friend list greater than 150. Thus, they capped the number of 'Friends' people could have on the system (boyd, 2006).

Interpretation is at the center of data analysis. Regardless of the size of a data set, it is subject to limitation and bias. Without those biases and limitations being understood and outlined, misinterpretation is the result. Big Data is at its most effective when researchers take account of the complex methodological processes that underlie the analysis of social data.

3. Bigger Data are Not Always Better Data

Social scientists have long argued that what makes their work rigorous is rooted in their systematic approach to data collection and analysis (McClosky, 1985). Ethnographers focus on reflexively accounting for bias in their interpretations. Experimentalists control and standardize the design of their experiment. Survey researchers drill down on sampling mechanisms and question bias. Quantitative researchers weigh up statistical significance. These are but a few of the ways in which social scientists try to assess the validity of each other's work. Unfortunately, some who are embracing Big Data presume the core methodological issues in the social sciences are no longer relevant. There is a problematic underlying ethos that bigger is better, that quantity necessarily means quality.

Twitter provides an example in the context of a statistical analysis. First, Twitter does not represent 'all people', although many journalists and researchers refer to 'people' and 'Twitter users' as synonymous. Neither is the population using Twitter representative of the global population. Nor can we assume that accounts and users are equivalent. Some users have multiple accounts. Some accounts are used by multiple people. Some people never establish an account, and simply access Twitter via the web. Some accounts are 'bots' that produce automated content without involving a person. Furthermore, the notion of an 'active' account is problematic. While some users post content frequently through Twitter, others participate as 'listeners' (Crawford 2009, p. 532). Twitter Inc. has revealed that 40 percent of active users sign in just to listen (Twitter, 2011). The very meanings of 'user' and 'participation' and 'active' need to be critically examined.

Due to uncertainties about what an account represents and what engagement looks like, it is standing on precarious ground to sample Twitter accounts and make claims about people and users. Twitter Inc. can make claims about all accounts or all tweets or a random sample thereof as they have access to the central database. Even so, they cannot easily account for lurkers, people who have multiple accounts or groups of people who all access one account. Additionally, the central database is also prone to outages, and tweets are frequently lost and deleted.

Twitter Inc. makes a fraction of its material available to the public through its APIs[1]. The 'firehose' theoretically contains all public tweets ever posted and explicitly excludes any tweet that a user chose to make private or 'protected.' Yet, some publicly accessible tweets are also missing from the firehose. Although a handful of companies and startups have access to the firehose, very few researchers have this level of access. Most either have access to a 'gardenhose' (roughly 10% of public tweets), a 'spritzer' (roughly 1% of public tweets), or have used 'white-listed' accounts where they could use the APIs to get access to different subsets of content from the public stream.[2] It is not clear what tweets are included in these different data streams or sampling them represents. It could be that the API pulls a random sample of tweets or that it pulls the first few thousand tweets per hour or that it only pulls tweets from a particular segment of the network graph. Given uncertainty, it is difficult for researchers to make claims about the quality of the data that they are analyzing. Is the data representative of all tweets? No, because it excludes tweets from protected accounts.[3] Is the data representative of all public tweets? Perhaps, but not necessarily.

These are just a few of the unknowns that researchers face when they work with Twitter data, yet these limitations are rarely acknowledged. Even those who provide a mechanism for how they sample from the firehose or the gardenhose rarely reveal what might be missing or how their algorithms or the architecture of Twitter's system introduces biases into the dataset. Some scholars simply focus on the raw number of tweets: but big data and whole data are not the same. Without taking into account the sample of a dataset, the size of the dataset is meaningless. For example, a researcher may seek to understand the topical frequency of tweets, yet if Twitter removes all tweets that contain problematic words or content – such as references to pornography – from the stream, the topical frequency would be wholly inaccurate. Regardless of the number of tweets, it is not a representative sample as the data is skewed from the beginning.

Twitter has become a popular source for mining Big Data, but working with Twitter data has serious methodological challenges that are rarely addressed by those who embrace it. When researchers approach a dataset, they need to understand – and publicly account for – not only the limits of the dataset, but also the limits of which questions they can ask of a dataset and what interpretations are appropriate.

[1] API stands for application programming interface; this refers to a set of tools that developers can use to access structured data.

[2] Details of what Twitter provides can be found at https://dev.twitter.com/docs/streaming-api/methods White-listed accounts were a common mechanism of acquiring access early on, but they are no longer available.

[3] The percentage of protected accounts is unknown. In a study of Twitter where they attempted to locate both protected and public Twitter accounts, Meeder et al (2010) found that 8.4% of the accounts they identified were protected.

This is especially true when researchers combine multiple large datasets. Jesper Anderson, co-founder of open financial data store FreeRisk, explains that combining data from multiple sources creates unique challenges: 'Every one of those sources is error-prone…I think we are just magnifying that problem [when we combine multiple data sets]' (Bollier 2010, p. 13). This does not mean that combining data doesn't have value – studies like those by Alessandro Acquisti and Ralph Gross (2009), which reveal how databases can be combined to reveal serious privacy violations are crucial. Yet, it is imperative that such combinations are not without methodological rigor and transparency.

Finally, in the era of the computational turn, it is increasingly important to recognize the value of 'small data'. Research insights can be found at any level, including at very modest scales. In some cases, focusing just on a single individual can be extraordinarily valuable. Take, for example, the work of Tiffany Veinot (2007), who followed one worker - a vault inspector at a hydroelectric utility company - in order to understand the information practices of blue-collar worker. In doing this unusual study, Veinot reframed the definition of 'information practices' away from the usual focus on early-adopter, white-collar workers, to spaces outside of the offices and urban context. Her work tells a story that could not be discovered by farming millions of Facebook or Twitter accounts, and contributes to the research field in a significant way, despite the smallest possible participant count. The size of data being sampled should fit the research question being asked: in some cases, small is best.

4. Not All Data Are Equivalent

Some researchers assume that analyses done with small data can be done better with Big Data. This argument also presumes that data is interchangeable. Yet, taken out of context, data lose meaning and value. Context matters. When two datasets can be modeled in a similar way, this does not mean that they are equivalent or can be analyzed in the same way. Consider, for example, the rise of interest in social network analysis that has emerged alongside the rise of social network sites (boyd & Ellison 2007) and the industry-driven obsession with the 'social graph'. Countless researchers have flocked to Twitter and Facebook and other social media to analyze the resultant social graphs, making claims about social networks.

The study of social networks dates back to early sociology and anthropology (e.g., Radcliffe-Brown 1940), with the notion of a 'social network' emerging in 1954 (Barnes) and the field of 'social network analysis' emerging shortly thereafter (Freeman 2006). Since then, scholars from diverse disciplines have been trying to understand people's relationships to one another using diverse methodological and analytical approaches. As researchers began interrogating the connections between people on public social media, there was a surge of interest in social network analysis. Now, network analysts are turning to study networks produced through mediated communication, geographical movement, and other data traces.

However, the networks produced through social media and resulting from communication traces are not necessarily interchangeable with other social network data. Just because two people are physically co-present – which may be made visible to cell towers or captured through photographs – does not mean that they know one another. Furthermore, rather than indicating the presence of predictable objective patterns, social network sites facilitate connectedness across structural boundaries and act as a dynamic source of change: taking a snapshot, or even witnessing a set of traces over time does not capture the complexity of all social relations. As Kilduff and Tsai (2003, p. 117) note, 'network research tends to proceed from a naive ontology that takes as unproblematic the objective existence and persistence of patterns, elementary parts and social systems.' This approach can yield a particular kind of result when analysis is conducted only at a fixed point in time, but quickly unravels as soon as broader questions are asked (Meyer et al. 2005).

Historically speaking, when sociologists and anthropologists were the primary scholars interested in social networks, data about people's relationships was collected through surveys, interviews, observations, and experiments. Using this data, social scientists focused on describing one's 'personal networks' – the set of relationships that individuals develop and maintain (Fischer 1982). These connections were evaluated based on a series of measures developed over time to identify personal connections. Big Data introduces two new popular types of social networks derived from data traces: 'articulated networks' and 'behavioral networks.'

Articulated networks are those that result from people specifying their contacts through a mediating technology (boyd 2004). There are three common reasons in which people articulate their connections: to have a list of contacts for personal use; to publicly display their connections to others; and to filter content on social media. These articulated networks take the form of email or cell phone address books, instant messaging buddy lists, 'Friends' lists on social network sites, and 'Follower' lists on other social media genres. The motivations that people have for adding someone to each of these lists vary widely, but the result is that these lists can include friends, colleagues, acquaintances, celebrities, friends-of-friends, public figures, and interesting strangers.

Behavioral networks are derived from communication patterns, cell coordinates, and social media interactions (Meiss *et al.* 2008; Onnela *et al.* 2007). These might include people who text message one another, those who are tagged in photos together on Facebook, people who email one another, and people who are physically in the same space, at least according to their cell phone.

Both behavioral and articulated networks have great value to researchers, but they are not equivalent to personal networks. For example, although often contested, the concept of 'tie strength' is understood to indicate the importance of individual relationships (Granovetter, 1973). When a person chooses to list someone as their 'Top Friend' on MySpace, this may or may not be their closest friend; there are all sorts of social reasons to not list one's most intimate connections first (boyd, 2006). Likewise, when mobile phones recognize that a worker spends more time with colleagues than their spouse, this does not necessarily mean that they have stronger ties with their colleagues than their spouse. Measuring tie strength through frequency or public articulation is a common mistake: tie strength – and many of the theories built around it – is a subtle reckoning in how people understand and value their relationships with other people.

Fascinating network analysis can be done with behavioral and articulated networks. But there is a risk in an era of Big Data of treating every connection as equivalent to every other connection, of assuming frequency of contact is equivalent to strength of relationship, and of believing that an absence of connection indicates a relationship should be made. Data is not generic. There is value to analyzing data abstractions, yet the context remains critical.

5. Just Because it is Accessible Doesn't Make it Ethical

In 2006, a Harvard-based research project started gathering the profiles of 1,700 college-based Facebook users to study how their interests and friendships changed over time (Lewis et al. 2008). This supposedly anonymous data was released to the world, allowing other researchers to explore and analyze it. What other researchers quickly discovered was that it was possible to de-anonymize parts of the dataset: compromising the privacy of students, none of whom were aware their data was being collected (Zimmer 2008).

The case made headlines, and raised a difficult issue for scholars: what is the status of so-called 'public' data on social media sites? Can it simply be used, without requesting permission? What constitutes best ethical practice for researchers? Privacy campaigners already see this as a key battleground where better privacy protections are needed. The difficulty is that privacy breaches are hard to make specific – is there damage done at the time? What about twenty years hence? 'Any data on human subjects inevitably raise privacy issues, and the real risks of abuse of such data are difficult to quantify' (*Nature*, cited in Berry 2010).

Even when researchers try to be cautious about their procedures, they are not always aware of the harm they might be causing in their research. For example, a group of researchers noticed that there was a correlation between self-injury ('cutting') and suicide. They prepared an educational intervention seeking to discourage people from engaging in acts of self-injury, only to learn that their intervention prompted an increase in suicide attempts. For some, self-injury was a safety valve that kept the desire to attempt suicide at bay. They immediately ceased their intervention (Emmens & Phippen 2010).

Institutional Review Boards (IRBs) – and other research ethics committees – emerged in the 1970s to oversee research on human subjects. While unquestionably problematic in implementation (Schrag, 2010), the goal of IRBs is to provide a framework for evaluating the ethics of a particular line of research inquiry and to make certain that checks and balances are put into place to protect subjects. Practices like 'informed consent' and protecting the privacy of informants are intended to empower participants in light of earlier abuses in the medical and social sciences (Blass, 2004; Reverby, 2009). Although IRBs cannot always predict the harm of a particular study – and, all too often, prevent researchers from doing research on grounds other than ethics – their value is in prompting scholars to think critically about the ethics of their research.

With Big Data emerging as a research field, little is understood about the ethical implications of the research being done. Should someone be included as a part of a large aggregate of data? What if someone's 'public' blog post is taken out of context and analyzed in a way that the author never imagined? What does it mean for someone to be spotlighted or to be analyzed without knowing it? Who is responsible for making certain that individuals and communities are not hurt by the research process? What does consent look like?

It may be unreasonable to ask researchers to obtain consent from every person who posts a tweet, but it is unethical for researchers to justify their actions as ethical simply because the data is accessible. Just because content is publicly accessible doesn't mean that it was meant to be consumed by just anyone (boyd & Marwick, 2011). There are serious issues involved in the ethics of online data collection and analysis (Ess, 2002). The process of evaluating the research ethics cannot be ignored simply because the data is seemingly accessible. Researchers must keep asking themselves – and their colleagues – about the ethics of their data collection, analysis, and publication.

In order to act in an ethical manner, it is important that scholars reflect on the importance of accountability. In the case of Big Data, this means both accountability to the field of research, and accountability to the research subjects. Academic researchers are held to specific professional standards when working with human participants in order to protect their rights and well-being. However, many ethics boards do not understand the processes of mining and anonymizing Big Data, let alone the errors that can cause data to become personally identifiable. Accountability to the field and to human subjects required rigorous thinking about the ramifications of Big Data, rather than assuming that ethics boards will necessarily do the work of ensuring people are protected. Accountability here is used as a broader concept that privacy, as Troshynski *et al.* (2008) have outlined, where the concept of accountability can apply even when conventional expectations of privacy aren't in question. Instead, accountability is a multi-directional relationship: there may be accountability to superiors, to colleagues, to participants and to the public (Dourish & Bell 2011).

There are significant questions of truth, control and power in Big Data studies: researchers have the tools and the access, while social media users as a whole do not. Their data was created in highly context-sensitive spaces, and it is entirely possible that some social media users would not give permission for their data to be used elsewhere. Many are not aware of the multiplicity of agents and algorithms currently gathering and storing their data for future use. Researchers are rarely in a user's imagined audience, neither are users necessarily aware of all the multiple uses, profits and other gains that come from information they have posted. Data may be public (or semi-public) but this does not simplistically equate with full permission being given for all uses. There is a considerable difference between being in public and being public, which is rarely acknowledged by Big Data researchers.

6. Limited Access to Big Data Creates New Digital Divides

In an essay on Big Data, Scott Golder (2010) quotes sociologist George Homans (1974): 'The methods of social science are dear in time and money and getting dearer every day.' Historically speaking, collecting data has been hard, time consuming, and resource intensive. Much of the enthusiasm surrounding Big Data stems from the perception that it offers easy access to massive amounts of data.

But who gets access? For what purposes? In what contexts? And with what constraints? While the explosion of research using data sets from social media sources would suggest that access is straightforward, it is anything but. As Lev Manovich (2011) points out, 'only social media companies have access to really large social data - especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not.' Some companies restrict access to their data entirely; other sell the privilege of access for a high fee; and others offer small data sets to university-based researchers. This produces considerable unevenness in the system: those with money – or those inside the company – can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access.

It is also important to recognize that the class of the Big Data rich is reinforced through the university system: top-tier, well-resourced universities will be able to buy access to data, and students from the top universities are the ones most likely to be invited to work within large social media companies. Those from the periphery are less likely to get those invitations and develop their skills. The result is that the divisions between those who went to the top universities and the rest will widen significantly.

In addition to questions of access, there are questions of skills. Wrangling APIs, scraping and analyzing big swathes of data is a skill set generally restricted to those with a computational background. When computational skills are positioned as the most valuable, questions emerge over who is advantaged and who is disadvantaged in such a context. This, in its own way, sets up new hierarchies around 'who can read the numbers', rather than recognizing that computer scientists and social scientists both have valuable perspectives to offer. Significantly, this is also a gendered division. Most researchers who have computational skills at the present moment are male and, as feminist historians and philosophers of science have demonstrated, who is asking the questions determines which questions are asked (Forsythe 2001; Harding 1989). There are complex questions about what kinds of research skills are valued in the future and how those skills are taught. How can students be educated so that they are equally comfortable with algorithms and data analysis as well as with social analysis and theory?

Finally, the difficulty and expense of gaining access to Big Data produces a restricted culture of research findings. Large data companies have no responsibility to make their data available, and they have total control over who gets to see it. Big Data researchers with access to proprietary data sets are less likely to choose questions that are contentious to a social media company, for example, if they think it may result in their access being cut. The chilling effects on the kinds of research questions that can be asked - in public or private - are something we all need to consider when assessing the future of Big Data.

The current ecosystem around Big Data creates a new kind of digital divide: the Big Data rich and the Big Data poor. Some company researchers have even gone so far as to suggest that academics shouldn't bother studying social media - as in-house people can do it so much better.[4] Such explicit efforts to demarcate research 'insiders' and 'outsiders' – while by no means new – undermine the utopian rhetoric of those who evangelize about the values of Big Data. 'Effective democratisation can always be measured by this essential criterion,' Derrida claimed, 'the participation in and access to the archive, its constitution, and its interpretation' (1996, p. 4). Whenever inequalities are explicitly written into the system, they produce class-based structures. Manovich writes of three classes of people in the realm of Big Data: 'those who create data (both consciously and by leaving digital footprints), those who have the means to collect it, and those who have expertise to analyze it' (2011). We know that the last group is the smallest, and the most privileged: they are also the ones who get to determine the rules about how Big Data will be used, and who gets to participate. While institutional inequalities may be a forgone conclusion in academia, they should nevertheless be examined and questioned. They produce a bias in the data and the types of research that emerge.

By arguing that the Big Data phenomenon is implicated in some much broader historical and philosophical shifts is not to suggest it is solely accountable; the academy is by no means the sole driver behind the computational turn. There is a deep government and industrial drive toward gathering and extracting maximal value from data, be it information that will lead to more targeted advertising, product design, traffic planning or criminal policing. But we do think there are serious and wide-ranging implications for the operationalization of Big Data, and what it will mean for future research agendas. As Lucy Suchman (2011) observes, via Levi Strauss, 'we are our tools.' We should consider how they participate in shaping the world with us as we use them. The era of Big Data has only just begun, but it is already important that we start questioning the assumptions, values, and biases of this new wave of research. As scholars who are invested in the production of knowledge, such interrogations are an essential component of what we do.

[4] During his keynote talk at the International Conference on Weblogs and Social Media (ICWSM) in Barcelona on July 19, 2011, Jimmy Lin – a researcher at Twitter – discouraged researchers from pursuing lines of inquiry that internal Twitter researchers could do better given their preferential access to Twitter data.

Acknowledgements

We wish to thank Heather Casteel for her help in preparing this article. We are also deeply grateful to Eytan Adar, Tarleton Gillespie, and Christian Sandvig for inspiring conversations, suggestions, and feedback.

References

Acquisti, A. & Gross, R. (2009) 'Predicting Social Security Numbers from Public Data', Proceedings of the National Academy of Science, vol. 106, no. 27, pp. 10975-10980.

Anderson, C. (2008) 'The End of Theory, Will the Data Deluge Makes the Scientific Method Obsolete?', Edge, <http://www.edge.org/3rd_culture/anderson08/ anderson08_index.html>. [25 July 2011]

Baca, G. (2004) 'Legends of Fordism: Between Myth, History, and Foregone Conclusions', Social Analysis, vol. 48, no.3, pp. 169-178.

Barnes, J. A. (1954) 'Class and Committees in a Norwegian Island Parish', Human Relations, vol. 7, no. 1, pp. 39–58.

Barry, A. and Born, G. (2012) Interdisciplinarity: reconfigurations of the Social and Natural Sciences. Taylor and Francis, London.

Behar, R. and Gordon, D. A., eds. (1996) *Women Writing Culture.* University of California Press, Berkeley, California.

Berry, D. (2011) 'The Computational Turn: Thinking About the Digital Humanities', Culture Machine. vol 12. <http://www.culturemachine.net/index.php/cm/article/view/440/470>. [11 July 2011].

Blass, T. (2004) *The Man Who Shocked the World: The Life and Legacy of Stanley Milgram*. Basic Books, New York, New York.

Bollier, D. (2010) 'The Promise and Peril of Big Data', <http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf>. [11 July 2011].

boyd, d. (2004) 'Friendster and Publicly Articulated Social Networks', Conference on Human Factors and Computing Systems (CHI 2004). ACM, April 24-2, Vienna.

boyd, d. (2006) 'Friends, Friendsters, and Top 8: Writing community into being on social network sites', First Monday vol. 11, no. 12, article 2.

boyd, d. and Ellison, N. (2007) 'Social Network Sites: Definition, History, and Scholarship', Journal of Computer-Mediated Communication, vol. 13, no.1, article 11.

boyd, d. and Marwick, A. (2011) 'Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies,' paper given at Oxford Internet Institute Decade in Time Conference. Oxford, England.

Bowker, G. C. (2005) Memory Practices in the Sciences. MIT Press, Cambridge, Massachusetts.

Burkholder, L, ed. (1992) Philosophy and the Computer, Boulder, San Francisco, and Oxford: Westview Press.

Cain, M. and Finch, J. (1981) Towards a Rehabilitation of Data. In: P. Abrams, R. Deem, J. Finch, & P. Rock (eds.), Practice and Progress: British Sociology 1950-1980, George Allen and Unwin, London.

Clifford, J. and Marcus, G. E., eds. (1986) *Writing Culture: The Poetics and Politics of Ethnography*. University of California Press, Berkeley, California.

Crawford, K. (2009) 'Following you: Disciplines of listening in social media', Continuum: Journal of Media & Cultural Studies vol. 23, no. 4, 532-33.

Du Gay, P. and Pryke, M. (2002) Cultural Economy: Cultural Analysis and Commercial Life, Sage, London.

Dunbar, R. (1998) Grooming, Gossip, and the Evolution of Language, Harvard University Press, Cambridge.

Derrida, J. (1996) Archive Fever: A Freudian Impression. Trans. Eric Prenowitz, University of Chicago Press, Chicago & London.

Emmens, T. and Phippen, A. (2010) 'Evaluating Online Safety Programs', Harvard Berkman Center for Internet and Society, <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Emmens_Phippen_Evaluating-Online-Safety-Programs_2010.pdf>. [23 July 2011].

Ess, C. (2002) 'Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee,' Association of Internet Researchers, <http://aoir.org/reports/ethics.pdf >. [12 September 2011].

Fischer, C. (1982) To Dwell Among Friends: Personal Networks in Town and City. University of Chicago, Chicago.

Forsythe, D. (2001) Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence, Stanford University Press, Stanford.

Freeman, L. (2006) The Development of Social Network Analysis, Empirical Press, Vancouver.

Gitelman, L. (2011) Notes for the upcoming collection 'Raw Data' is an Oxymoron, <https://files.nyu.edu/lg91/public/>. [23 July 2011].

Golder, S. (2010) 'Scaling Social Science with Hadoop', Cloudera Blog, <http://www.cloudera.com/blog/2010/04/scaling-social-science-with-hadoop/>. [June 18 2011].

Granovetter, M. S. (1973) 'The Strength of Weak Ties,' *American Journal of Sociology* vol. 78, issue 6, pp. 1360-80.

Harding, S. (2010) 'Feminism, science and the anti-Enlightenment critiques', in Women, knowledge and reality: explorations in feminist philosophy, eds A. Garry and M. Pearsall, Boston: Unwin Hyman, 298–320.

Homans, G.C. (1974) Social Behavior: Its Elementary Forms, Harvard University Press, Cambridge, MA.

Isbell, C., Kearns, M., Kormann, D., Singh, S. & Stone, P. (2000) 'Cobot in LambdaMOO: A Social Statistics Agent', paper given at the 17th National Conference on Artificial Intelligence (AAAI-00). Austin, Texas.

Kilduff, M. and Tsai, W. (2003) Social Networks and Organizations, Sage, London.

Kranzberg, M. (1986) 'Technology and History: Kranzberg's Laws', Technology and Culture vol. 27, no. 3, pp. 544-560.

Latour, B. (2009). 'Tarde's idea of quantification', in The Social After Gabriel Tarde: Debates and Assessments, ed M. Candea, London: Routledge, pp. 145-162.< http:// www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf>. [19 June 2011].

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D.,Christakis, N., Contractor, N., Fowler, J.,Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). 'Computational Social Science'. Science vol. 323, pp. 721-3.

Lewis, K., Kaufman, J., Gonzalez, M.,Wimmer, A., & Christakis, N. (2008) 'Tastes, ties, and time: A new social network dataset using Facebook.com', Social Networks vol. 30, pp. 330-342.

Manovich, L. (2011) 'Trending: The Promises and the Challenges of Big Social Data', Debates in the Digital Humanities, ed M.K.Gold. The University of Minnesota Press, Minneapolis, MN <http://www.manovich.net/DOCS/Manovich_trending_paper.pdf>.[15 July 2011].

McCloskey, D. N. (1985) 'From Methodology to Rhetoric', In The Rhetoric of Economics au D. N. McCloskey, University of Wisconsin Press, Madison, pp. 20-35.

Meeder, B., Tam, J., Gage Kelley, P., & Faith Cranor, L. (2010) 'RT @IWantPrivacy: Widespread Violation of Privacy Settings in the Twitter Social Network', Paper presented at Web 2.0 Security and Privacy, W2SP 2011, Oakland, CA.

Meiss, M.R., Menczer, F., and A. Vespignani. (2008) 'Structural analysis of behavioral networks from the Internet', Journal of Physics A: Mathematical and Theoretical, vol. 41, no. 22, pp. 220-224.

Meyer D, Gaba, V., Colwell, K.A., (2005) 'Organizing Far from Equilibrium: Nonlinear Change in Organizational Fields', Organization Science, vol. 16, no. 5, pp.456-473.

Moretti, F. (2007) Graphs, Maps, Trees: Abstract Models for a Literary History. Verso, London.

Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., & Kertész, J., Barabási, A.L. (2007) 'Structure and tie strengths in mobile communication networks', Proceedings from the National Academy of Sciences, vol.104, no.18, pp. 7332-7336.

Pariser, E. (2011) The Filter Bubble: What the Internet is Hiding from You. Penguin Press, New York, NY.

Radcliffe-Brown, A.R. (1940) 'On Social Structure', The Journal of the Royal Anthropological Institute of Great Britain and Ireland vol.70, no.1, pp.1–12.

Reverby, S. M. (2009) *Examining Tuskegee: The Infamous Syphilis Study and Its Legacy*. University of North Carolina Press.

Schrag, Z. M. (2010) Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965-2009. Johns Hopkins University Press, Baltimore, Maryland.

Suchman, L. (2011) 'Consuming Anthropology', in Interdisicpinarity: Reconfigurations of the social and natural sciences, eds Andrew Barry and Georgina Born, Routledge, London and New York.

Twitter. (2011) 'One hundred million voices', Twitter blog, <http://blog.twitter.com/2011/09/one-hundred-million-voices.html>. [12 September 2011]

Veinot, T. (2007) 'The Eyes of the Power Company: Workplace Information Practices of a Vault Inspector', The Library Quarterly, vol.77, no.2, pp.157-180.

Zimmer, M. (2008) 'More on the 'Anonymity' of the Facebook Dataset – It's Harvard College', MichaelZimmer.org Blog, <http://www.michaelzimmer.org/2008/01/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/>. [20 June 2011].

dashes.com

The Web We Lost

The tech industry and its press have treated the rise of billion-scale social networks and ubiquitous smartphone apps as an unadulterated win for regular people, a triumph of usability and empowerment. They seldom talk about what we've lost along the way in this transition, and I find that younger folks may not even know how the web used to be.

So here's a few glimpses of a web that's mostly faded away:

- Five years ago, most social photos were uploaded to Flickr, where they could be tagged by humans or even by apps and services, using machine tags. Images were easily discoverable on the public web using simple RSS feeds. And the photos people uploaded could easily be licensed under permissive licenses like those provided by Creative Commons, allowing remixing and reuse in all manner of creative ways by artists, businesses, and individuals.
- A decade ago, Technorati let you search most of the social web in real-time (though the search tended to be awful slow in presenting results), with tags that worked as hashtags do on Twitter today. You could find the sites that had linked to your content with a simple search, and find out who was talking about a topic regardless of what tools or platforms they were using to publish their thoughts. At the time, this was so exciting that when Technorati failed to keep up with the growth of the blogosphere, people were so disappointed that even the usually-circumspect Jason Kottke flamed the site for letting him down. At the first blush of its early success, though, Technorati elicited effusive praise from the likes of John Gruber:

> [Y]ou could, in theory, write software to examine the source code of a few hundred thousand weblogs, and create a database of the links between these weblogs. If your software was clever enough, it could refresh its information every few hours, adding new links to the

> database nearly in real time. This is, in fact, exactly what Dave Sifry has created with his amazing Technorati. At this writing, Technorati is watching over 375,000 weblogs, and has tracked over 38 million links. If you haven’t played with Technorati, you’re missing out.

- Ten years ago, you could allow people to post links on your site, or to show a list of links which were driving inbound traffic to your site. Because Google hadn't yet broadly introduced AdWords and AdSense, links weren't about generating revenue, they were just a tool for expression or editorializing. The web was an interesting and different place before links got monetized, but by 2007 it was clear that Google had changed the web forever, and for the worse, by corrupting links.

- In 2003, if you introduced a single-sign-in service that was run by a company, even if you documented the protocol and encouraged others to clone the service, you'd be described as introducing a tracking system worthy of the PATRIOT act. There was such distrust of consistent authentication services that even Microsoft had to give up on their their attempts to create such a sign-in. Though their user experience was not as simple as today's ubiquitous ability to sign in with Facebook or Twitter, the TypeKey service introduced then had much more restrictive terms of service about sharing data. And almost every system which provided identity to users allowed for pseudonyms, respecting the need that people have to not always use their legal names.

- In the early part of this century, if you made a service that let users create or share content, the expectation was that they could easily download a full-fidelity copy of their data, or import that data into other competitive services, with no restrictions. Vendors spent years working on interoperability around data exchange purely for the benefit of their users, despite theoretically lowering the barrier to entry for competitors.

- In the early days of the social web, there was a broad expectation that regular people might own their own identities by having their own websites, instead of being dependent on a few big sites to host their online identity. In this vision, you would own your own domain name and have complete control over its contents, rather than having a handle tacked on to the end of a huge company's site. This was a sensible reaction to the realization that big sites rise and fall in popularity, but that regular people need an identity that persists longer than those sites do.

- Five years ago, if you wanted to show content from one site or app on your own site or app, you could use a simple, documented format to do so, without requiring a business-development deal or contractual agreement between the sites. Thus, user experiences weren't subject to the vagaries of the political battles between different companies, but instead were consistently based on the extensible architecture of the web itself.
- A dozen years ago, when people wanted to support publishing tools that epitomized all of these traits, they'd crowd-fund the costs of the servers and technology needed to support them, even though things cost a lot more in that era before cloud computing and cheap bandwidth. Their peers in the technology world, though ostensibly competitors, would even contribute to those efforts.

This isn't our web today. We've lost key features that we used to rely on, and worse, we've abandoned core values that used to be fundamental to the web world. To the credit of today's social networks, they've brought in hundreds of millions of new participants to these networks, and they've certainly made a small number of people rich.

But they haven't shown *the web itself* the respect and care it deserves, as a medium which has enabled them to succeed. And they've now narrowed the possibilites of the web for an entire generation of users who don't realize how much more innovative and meaningful their experience could be.

Back To The Future

When you see interesting data mash-ups today, they are often still using Flickr photos because Instagram's feeble metadata sucks, and the app is only reluctantly on the web at all. We get excuses about why we can't search for old tweets or our own relevant Facebook content, though we got more comprehensive results from a Technorati search that was cobbled together on the feeble software platforms of its era.

We get bullshit turf battles like Tumblr not being able to find your Twitter friends or Facebook not letting Instagram photos show up on Twitter because of giant companies pursuing their agendas instead of collaborating in a way that would serve users. And we get a generation of entrepreneurs encouraged to make more narrow-minded, web-hostile products like these because it continues to make a small number of wealthy people even more wealthy, instead of letting lots of people build innovative new opportunities for themselves on top of the web itself.

We'll fix these things; I don't worry about that. The technology industry, like all industries, follows cycles, and the pendulum is swinging back to the broad, empowering philosophies that underpinned the early social web. But we're going to face a big challenge with re-educating a billion people about what the web *means*, akin to the years we spent as everyone moved off of AOL a decade ago, teaching them that there was so much more to the experience of the Internet than what they know.

This isn't some standard polemic about "those stupid walled-garden networks are bad!" I know that Facebook and Twitter and Pinterest and LinkedIn and the rest are *great* sites, and they give their users a lot of value. They're amazing achievements, from a pure software perspective. But they're based on a few assumptions that aren't necessarily correct. The primary fallacy that underpins many of their mistakes is that user flexibility and control necessarily lead to a user experience complexity that hurts growth. And the second, more grave fallacy, is the thinking that exerting extreme control over users is the best way to maximize the profitability and sustainability of their networks.

The fist step to disabusing them of this notion is for the people creating the next generation of social applications to learn a little bit of history, to *know your shit*, whether that's about Twitter's business model or Google's social features or anything else. We have to know what's been tried and failed, what good ideas were simply ahead of their time, and what opportunities have been lost in the current generation of dominant social networks.

So what did I miss? What else have we lost on the social web?

Original URL:
http://dashes.com/anil/2012/12/the-web-we-lost.html

Live research: Twittering an election debate

Greg Elmer

New Media Society 2013 15: 18 originally published online 23 September 2012
DOI: 10.1177/1461444812457328

Article

new media & society
15(1) 18–30
© The Author(s) 2012
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461444812457328
nms.sagepub.com
SAGE

Live research: Twittering an election debate

Greg Elmer
Ryerson University, Canada

Abstract

This paper questions how vertical tickers on leading social media platforms (blogs, Facebook, and in particular the Twitter micro-blogging platform) pose new challenges to research that focuses on political communications campaigns. Vertical looped tickers highlight the fleeting nature of contemporary networked and socially mediated communications, since they provide an intensely compressed space (interface) and time to have posts viewed by friends and followers. This article draws upon a research collaboration with the news division of the Canadian Broadcasting Corporation (CBC) to understand how Canadian political parties increasingly worked to strategically intervene, in real time on Twitter, during a broadcast political debate.

Keywords

Election debate, political communications, politics 2.0, social media, Twitter

The rapid growth of networked, handheld, virtual, embedded, and locative information and communication technologies raises important questions about methods of studying processes, objects, actors and technological platforms that are by design or dysfunction constantly in flux. Mediated life has so vastly multiplied its forms and sites of communication and storytelling that the ability to recall where one heard or viewed a news report, a rumor about a friend, or even the source of an urgent work-related request now requires a panoply of aggregate remediators – smartphones, RSS feed managers, personalized search engines, live social network feeds and so forth. In an age of meta-information such technologies serve to collapse and focus time – which is increasingly socially mediated time – to a window of approximately ten minutes. This occurs both in the past, through interfaces like Facebook or Twitter that bury ten-minute-old communications,

Corresponding author:
Greg Elmer, Rogers Communications Centre, Ryerson University, Room 309, 350 Victoria Street, Toronto, ON M5B 2K3, Canada.
Email: gelmer@ryerson.ca

and in the future through anticipatory buzzing and pinging reminders of duties to come in ten minutes' time. Visually, such *interface time* literally hypermediates a window in time – what can fit on the interface before being pushed off (or typically down) to make way for the next ten minutes.

Unlike Facebook, Myspace, Cyworld, Bebo and other social networking sites that offer a vast array of interfaces and functions for users and their networked friends, microblogging platforms like Twitter offer a decidedly trimmed-down interface focused on a vertical ticker of short (140 characters maximum) bursts of text. Such an interface maintains a concise focus on a very small window of time. Unlike horizontal stock or sports tickers that communicate incremental changes in prices and scores in a constant loop, Twitter's vertical ticker relies upon friends and contacts to actively repost or 'retweet' a post back to the top of its vertical-ticker interface. Unlike looped horizontal feeds and tickers, initial research has found that only 6 percent of all tweets are reposted back to the top of Twitter's vertical interface.[1] Duncan Geere (2010) summarizes this point nicely: '92 percent of … retweets occur within the first hour. Multiplying those probabilities together means that fewer than one in 200 messages get retweeted after an hour's gone by. Essentially, once that hour's up, your message is ancient history.' Such findings thus question the means by which individuals or, as we shall see in this article, political campaigns might sustain and expand the readership of their posts across Twitter's social networks. Architecturally speaking, un-retweeted or reposted comments on Twitter resemble a hyperactive blog interface, in which newer posts simply push older ones down in short order off a user's PC, tablet or smartphone interface. Older posts are in effect buried into the interface depths of the infinite downward scroll or pushed off onto additional hyperlinked pages (i.e., the indefinite 'next page' click).

The 'live' form of research discussed in this article seeks to understand the techniques, technologies and user dynamics that attempt to expand this intensely time- and interface-compressed platform during a live broadcast political debate. The article argues that the emergence of vertical tickers and other forms of hyper-immediate, time-compressed social media interfaces highlight the need for real-time forms of Internet research. The article investigates how political forms of communication – particularly during heightened periods of partisan conflict such as elections, scandals and political/economic crises – are being expanded onto 'second screens' (typically PCs and smartphones running social media interfaces) that enable socially mediated and networked commentary and conversations on live broadcast events. This live form of research thus requires an understanding of the networked affordances and technological encodings (e.g., meta-tags) of discrete digital bursts or *objects* (Schneider and Foot, 2010), particularly tweets, blog posts, comments posted on online newspapers' web pages, images and videos from their specific platforms, or from larger aggregators such as personalized feed (RSS) managers, social networking sites (e.g. Facebook) or search engines like Google. Such components of social networking sites consequently form the basis for software code-focused media research, the platform upon which researchers can attempt to determine the tactics, conventions, functions and dysfunctions of real-time political discourse on Twitter, or across mediated screens, platforms and interfaces (Rogers, 2006). Given the rapid development of social media platforms, conventions on these platforms, and the ever-changing sets of rules and regulations that govern sites like

Twitter (as manifested through their programming interface or API), this article's discussion of 'live research' seeks to account for the always already shifting dynamics in online communication flows. While some may impart a Latourian (Latour and Weibel 2005) motive at work here, particularly with regard to his 'object-oriented' philosophy (see Harman, 2009), this study extends well beyond the tweet-as-object to an appreciation of the temporalities of interfaces, information architectures and the political tactics deployed on social media platforms like Twitter. Thus, what is suggested here is a more hyper-immediate and immersed form of research, not one that merely 'tracks the object', as Lash (2007) argues, but rather a reflexive, empirical approach to understanding media flows in social media's increasingly compressed interface time.

A focus on in-the-moment communications and networking attempts to build upon broader discussions, theories and methods of understanding open-ended networked, non-hierarchical or distributed forms of communications (Fuller, 2003; Galloway, 2004), to one that attempts to understand the strategic (politically speaking) deployment of political campaigns and communications in terms of compressed and socially mediated interface time (Cunningham, 2008). The question of 'live research' in Internet studies, and consequently in ICT-enabled studies of political communications (Chadwick and Howard, 2009; Kluver et al., 2007) continues to develop an important methodological debate within the broader field of Internet studies. Andrew Chadwick's (2011) recent study of shifting political information cycles are of particular importance to this form of live research. In attempting to determine the new roles and opportunities that social media afford in the political process, Chadwick investigates the temporality and flow of political news, much like Norris (2000) before him, so as to better understand how social media actors intervene and disrupt political and mainstream media tempos and schedules in real time, in effect producing a new tempo of mediated political life, or a new 'political information cycle'.

Methods of real-time research, however, have a much longer history in Internet studies. Annette Markham's (1998) study of virtual chat rooms, for example, offered an auto-ethnographic approach to the study of computer-mediated communication, a distinctly participatory form of real-time or live research. Markham's study sought to enumerate the complex literacies involved in navigating a virtual chat room in the moment by logging the challenges she faced as they occurred in real time onscreen. Markham's study highlighted important conventions that occur in online environments, a process that was made all the more apparent by her recollections of being immersed in live interactions with other users and the software and interface itself. Christine Hine (2007) similarly suggested a 'connective ethnographic' approach to understanding how various forms of computer-mediated communication connected the user to their 'offline' life. It is this connective, networked approach that informs the present work. This article is an attempt to understand how political campaigns and communications seek to reconnect political communication (e.g. images, blog posts, excerpts from speeches) across social media interfaces, and in doing so we hope to redress the temporal limits of communication and subsequent limited attention span of new media audiences and social media interfaces. Recent examples of live or real-time research have also emphasized the act of always being ready to conduct research, of being in a position to capture a political crisis or a live-mediated event on the Net. Andreas Jungherr and Pascal Jurgens' study of Twitter in

Germany (2011), for instance, builds upon Allan's (2002) notion of 'topic detection', an attempt to continuously collect and analyze social media content feeds and flows of information for signs of increased activity. While the project presented here similarly developed a method of data collection and content analysis of tweets in advance of the televised election debate under study, this article seeks to understand the tactical forms of political communication deployed in real time on Twitter and other Web platforms during the debate broadcast.

Overall, the 'live research' paradigm discussed in this article places greater emphasis on the relationship between the rules and regulations of social media platforms as we move from a 'news cycle' paradigm to one defined by a new media-enabled 'political information cycle' (Chadwick, 2011). At the center of this shift in mediated temporalities is a set of tactics that seeks to sustain networked and fleeting/time-compressed communications across new and old media platforms (e.g. TV, the Web, social media, hand-held devices) and, of course, mediated political dialogue, debate and commentary (Gurevitch et al., 2009). Methodologically speaking, the question to be answered is: why is there a need to study and analyze such dynamics in real time? One answer relates to the contingencies of interface time as a space that requires various strategies for communication (political communication in this instance) to be re-posted (or 're-tweeted' on the Twitter platform, although similar dynamics can be found on many other vertical feed-like social media platforms), so as to recursively spread across social networks and push the limits of socially mediated interface time. In the context of political campaigns, crisis management public relations or environmental disasters, such efforts to expand interface time take on an even greater significance in the form of the emergent use of second screens and interfaces. The interactive appendage to the broadcast sphere of political life (e.g. 24-hour news channels and live political programming) becomes an increasingly important space to view immediate reactions to live events from a host of online political actors (e.g. media pundits, political bloggers, politicians and their staff). Such 'live' or near real-time reactions in the Twittersphere have consequently emerged as sites from which to support, ridicule and/or refute the statements and claims made by public figures on live television. In political terms, micro-blogging sites like Twitter have become key sites of 'rapid response' to live political events and other particularly time-sensitive news stories.[2]

The effort to develop a 'live research' paradigm in new media studies also attempts to take into consideration the speed of communications. Publishing one's political opinion online, for example on a blog, is no longer subject to editorial delay. User-generated content can be posted in real time at the click of a mouse. Does it not make sense then to build such limited media time (or interface time, in the case of the Twitter ticker) into research methods to understand the effects of such media platforms and networks? Social media are structured to visualize only near real-time contributions; as such, their temporalities, flows and interfaces set the context in which political communications and campaigns are enabled, deployed and represented through the introduction of real-time architectures (back-end code) and interfaces (e.g. feeds and tickers). The fleeting nature of not only networked communication but also the ever-changing software code, interfaces and APIs that facilitate such micro-blogging activity require a temporal rethinking of what it means to conduct research on contemporary political communications and campaigning.

Networked (or ‘2.0’) communications and interactivity are over-determined by conventions of the present. Whether uploading, sharing, commenting, downloading, re-naming, importing, embedding or seeding, all such networked forms of communication and interactivity are enacted or published in the moment with little or no delay. Likewise, the very language of networked life, political or otherwise, amplifies the immediate while clearly ex-distancing the technological, political and economic underpinnings of such networks. It is this latter phenomenon that needs to be understood through the lens of the ‘live’.

Twittering a debate

In order to better understand the link between social media’s compressed interface time and second-screen interactivity in their aggregate role as re-mediator of live political discourse, the example of live research discussed here focuses on a collaboration between Ryerson’s Infoscape Lab and the news division of the Canadian Broadcasting Corporation (CBC) during the 2008 federal election in Canada.[3] This study focuses specifically on the development and execution of a near real-time analysis of political tweets posted during the CBC’s live English-language broadcast of the federal leaders’ debate on television,[4] a key moment in Canada’s national election. The Infoscape Lab’s live approach to the election night study was designed to capture an early-adopter moment in ICT-enabled political communications[5] – one that sought to determine the influence of Net-savvy political operatives, and also the degree to which the platform served as an interactive space for real-time commentary on a live broadcast event.[6]

Given the minority status of the governing Conservative Party in the Westminster-style Canadian House of Commons, a series of potential election-inducing showdowns had occurred over the previous 12 months. During this period, we developed a series of research methods and tools that tracked the growing importance and impact of the Canadian political blogosphere and published our findings. After receiving substantial media coverage of our research during the Ontario provincial election 2007,[7] producers in the news division of the Canadian Broadcasting Corporation (CBC News) invited the Infoscape Lab to extend our collaboration to the federal election. Dubbed ‘Ormiston Online’ (for the lead reporter on the project, Susan Ormiston), the CBC brought together staff from all their key news divisions (radio, new media, local, national and 24-hour TV) so as to better disseminate the news stories produced by the team for the CBC’s myriad news-focused programs and platforms. Unbeknownst to the Infoscape Lab at the time, the CBC had designed the project as a dry run for their subsequent multi-platform news realignment. The Infoscape Lab was approached to assist in the development of a public Web portal, Internet campaigning research, on-air interviews, and other advice related to developing news stories during the campaign. While we anticipated some analysis would need to be conducted on a daily basis during the campaign, our methods of collecting data (for blog posts and YouTube videos of the main party leaders) had been established, tested and refined over many months prior to our collaboration with the CBC. On a routine basis (three times per week), our team produced a ranking and short qualitative analysis of the most cited (linked to) blog posts from a sample of all the self-defined partisan political bloggers in Canada,[8] and a similar ranking of the week’s most-viewed YouTube-hosted videos related to the federal party leaders during the campaign.[9]

Data was collected and analyzed each morning and formatted for publication on the CBC's website (cbc.ca). One or two paragraphs were written in accessible language to provide context for the findings, which typically involved providing analysis for why certain posts or videos were receiving such attention online.

Our research into the impact of blogging and YouTube videos on the election campaign process served as the backbone of our contribution to the CBC's coverage of the Internet-based aspects of the campaign. The first half of the official campaign period had witnessed a series of Internet-based scandals, missteps and other campaign-related shenanigans that our collaborative project helped shed light upon through our social media research and its subsequent dissemination through the CBC's website and broadcast platforms. Executives at the CBC were reportedly pleased with our work and subsequently pushed for more content analysis, research and coverage of Internet-bound, campaign-related goings-on.

The most challenging live research aspect of the CBC collaboration concerns the use of the micro-blogging platform Twitter during the campaign's nationally televised leaders' debate. Days before the debate, we met with the producers of the Ormiston Online project at the CBC's corporate offices to discuss how we might cover the forthcoming televised event. Our discussions focused on converging the broadcast and social media screens so as to highlight the real-time discussions and debates initiated on Twitter that we believed would be responding to the comments, barbs, guffaws and poignant zingers served up by the party leaders during the televised debate.

Collecting the tweets

Unlike our research on Canadian partisan blogs (Elmer et al., 2009) that restricted its sample to opt-in, self-described partisan members of one of Canada's political party-branded blogrolls, the Twitter debate night project was a decidedly open-ended affair that called into question the means by which we would filter or otherwise collect micro-blogging posts. Recognizing the limits of Twitter's compressed interface time, and its real-time use as a form of audience debate and dialogue, our project not only sought to analyze the content but also the context – the time – it was posted. Axel Bruns' (2010) initial research on the use of Twitter during the 2010 Australian televised leaders' debate was similarly designed to compare trends with those attributed to a popular cooking show, implicitly questioning the social media activity of contrasting social interests. In this context, Bruns' use of hashtags (#) – the most common form of creating new feeds or thread-like vertical posts of tweets on similar topics – to filter and collect relevant posts for two simultaneously televised programs served as a helpful comparative approach to data collection. By contrast, partly due to the infancy of Twitter use in Canada at the time of our collaborative project, and in particular the conventions and practices associated with hashtagging content, no one hashtag could capture a representative sample of posts during the Canadian televised debate. In other words, the use of specific hashtags has emerged over time after much conversation, debate and adoption.

Unlike Bruns' study of the Australian debate night, our live research project also sought to merge two sets of data to pose both qualitative and quantitative questions.

We were not solely driven by the goal of determining the quantity of tweets during the debate broadcast, nor their numbers in context to other live events.[10] Rather, the project sought to determine the interplay between broadcast comments by the leaders and reactions on Twitter. After determining whether or not there was a correlation between specific rhetorical flourishes, issues or lively exchanges among the party leaders during the debate and audience members' Twitter posts, we also sought to determine how such exchanges were deployed tactically to expand Twitter's limited interface time and the subsequent reach of fleeting posts.[11]

We decided to cast a wide net to collect our micro-blogging posts related to the live broadcast debate. Forty-eight hours before the debate, the project staff – both academic and CBC-based – promoted the use of the #ormistondebate hashtag. Since both the project and the debate were being broadcast by the CBC, they were keen to cross-promote and otherwise brand their coverage. Overall, our research deployed a mixed hashtag, a Twitter account name, and formal party leader name search term 'basket' to cull as large a sample as possible.[12]

In addition to these meta-tag and formal name search terms, the project also made important use not only of the content of the tweets, but the time stamp or log that accompanied each post. Such time stamps afforded the ability to cross-reference Twitter posts with the time-stamped transcripts of the leaders' televised comments. While it took mere seconds to collect the tweets during the broadcast, our analysis was delayed by about ten minutes as we waited for the delivery of the transcripts from the CBC via email.

Debate night proved to be incredibly hectic as we collected the data, and subsequently produced charts (see Figure 1) that depicted the minute-by-minute activity in the Twittersphere (the chart was broadcast later that evening on CBC). While preparing such charts for broadcast our research team also referred to the transcript of the debate to correlate jumps in Twitter posts to specific moments in the televised debate. While we did

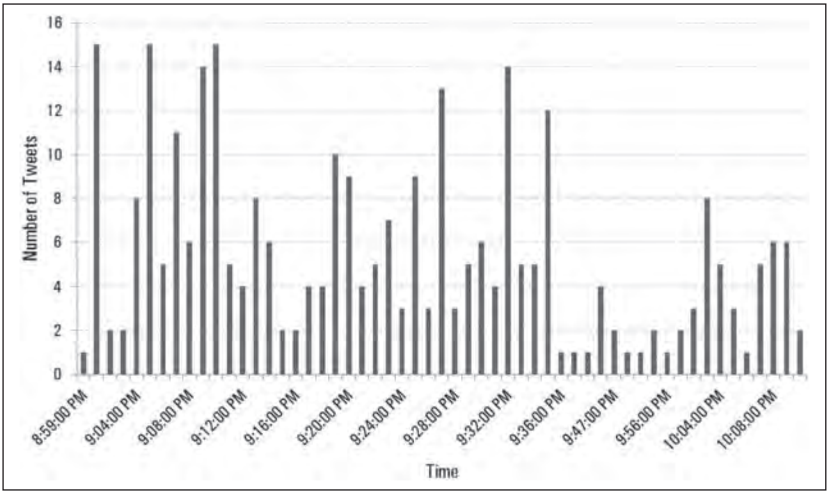


Figure 1. Twittering the debate.

not have enough time or space to include representative tweets on the charts to demonstrate these findings, we provided these tweets to our head reporter who used them on the live broadcast to qualify the spikes in Twitter activity shown on our chart. The most active moment over the first hour of the debate on Twitter, for example, occurred at 9:32 p.m., immediately after the left-of-center New Democratic Party (NDP) leader Jack Layton turned to the Prime Minister and let loose the first zinger of the debate: ‘Where is your platform? Under the sweater?’[13] The subsequent set of tweets clearly demonstrate a largely phatic or parrot-like use of the micro-blogging platform, meaning users either let out ‘wow’-like exclamations or simply reposted Layton’s one-liner, or both.

Reactions to the NDP leader’s jibe also demonstrate a distinct partisan moment between political parties. The succession of 12 posts that repeated or otherwise exalted the witty one-liner over the next minute was only briefly interrupted by one tweet from the centrist Liberal Party of Canada’s campaign account:

> (9: 32 pm) Liberal feed: Debates prove Jack Layton just doesn’t get it.

Over the course of the evening, however, the Liberal Party was not the most active political party on Twitter. While all the parties’ known bloggers and online activists took turns supporting their leader and taking apart the responses of their foes, only the NDP actively prepared a rapid-response approach to Twitter on the debate night. Using the @JackLayton account (the name of the NDP’s leader), the NDP sent out a series of ‘fact check’ posts over the course of the two-hour debate, with periodic links to more extensive rebuttals posted on the party’s election website. The party, in short, used the medium to respond to their opponents’ live statements in near real time, adding a whole new temporality to the media spin that typically erupts at the conclusion of televised debates:

> (9:53 pm) jacklayton: FACT CHECK: Harper says he is making important investments in science and technology in Canada #ormistondebate.
>
> (9:56 pm) jacklayton: FACT CHECK: Bloc not the only party with a Buy Canada policy − http://www.ndp.ca/page/7136.

While a number of users picked up on the tactic and lauded the party for its innovative use of Twitter, other comments suggested that viewers/Twitterers thought that Layton himself was posting such notes live on set:

> (10:10 pm) @jacklayton, stop texting from under the table!
>
> (10:57 pm) @jacklayton, explain to me how you are tweeting while the debate is on.

Such confusion might be explained by the early adopters’ lack of established social media conventions, but such strategic use of social media by a political party also highlights one aspect of media personalization deployed during campaign events. Given that social media are built upon a lexicon and architecture of friendship networks, the use of a personalized account by the NDP served to normalize partisan communications within the conventions of social media, while at the same time extending Twitter’s limited interface time onto their campaign website where additional ‘fact checks’, policies and the party’s campaign platform could be found.

A content analysis of the total number of mentions of the party leaders on the night of the debate concluded that while the NDP leader received substantial attention on Twitter (27 percent of all tweets mentioned Jack Layton), during the course of the live broadcast, it was the first-time participation of the Green Party in Canadian debates that topped the discussion on Twitter. As we see in Figure 2, Elizabeth May, the Green Party’s leader, was mentioned in almost one third (29 percent) of all the tweets during the debate night.

Upon reviewing our data 12 months after the live research project concluded, a series of other findings emerged – evidence that again supports and further qualifies the manner in which Twitter was used tactically by political parties, partisans and other online viewers/users on the debate night. The multi-mediated nature of the debate evening, and in particular the interplay between viewership, social media commentary and partisan campaigning, is also further amplified in a number of posts made during the debate evening. The Canadian federal leaders’ debate happened to coincide with the live broadcast of the debate between US vice-presidential candidates, which, it should be noted, included the controversial yet media-friendly Republican nominee Sarah Palin. At the very outset of the Canadian debate a number of users posted tweets referring to the use of multiple screens, online video streams and the switching of TV sets to catch one or the other debate:

> (9:19 pm) Watching #vpdebate on CNN and #cdbdeb08 on CBC live stream #ormistondebate.
>
> (9:21 pm) Just changed to the US VP Debate because so far it’s better than watching Jack Layton and Elizabeth May attack @pmharper. Will go back soon.

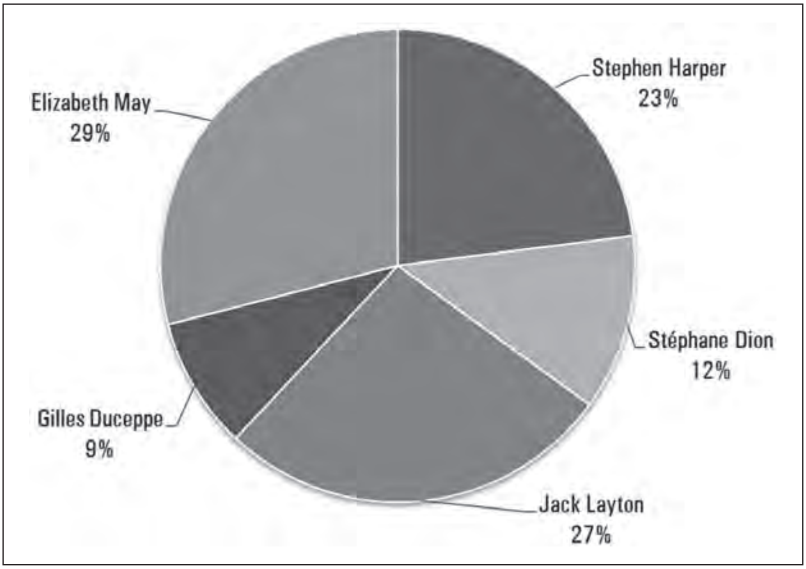


Figure 2. ‘Twittering about the leaders’ (broadcast on 2 October 2008 on CBC national news).

Other users engaged yet more social media platforms, in this instance the digital-photo hosting site Flickr, to capture and share their experiences of watching the Canadian leaders debate.

> (10:07 pm) The 5 leaders as they appear on my TV set. Elizabeth May http://flickr.com/photos/sarahroger/2908875540/.

After retrospectively reviewing data from the debate night, curiously absent is an expansion of Twitter's interface time onto other Web-based political documents. Apart from the previously noted NDP hyperlinks to their campaign website, of all the tweets posted during the live debate only two include links to other relevant documents. Such a finding seems counterintuitive given Twitter's predominate convention today of sharing links to articles, YouTube videos, Wikipedia and the like. While one of these posted links is rather whimsical, using a Web link to lyrics of a popular song to lampoon the NDP leader's choice of words,[14] the other is more tactically relevant in terms of expanding the sphere of the debate. At the outset of the debate, upon hearing the Green Party's leader cite a report on the economy, a user finds the document and shares it online:

> (9:18 pm) Here's a link to the OECD report Elizabeth May's talking about: http://www.oecd.org/dataoecd/33/55/40912642.pdf.

Lastly, efforts to tactically manage – as opposed to perhaps simply expand – Twitter's interface time were also clearly evident in the hours leading up to the televised debate. A debate over an appropriate hashtag for the event quickly degenerated into partisan bickering and balkanization, with online Conservatives promoting the use of the hashtag #cdndeb08. There were, in short, decidedly partisan and institutional elements to various attempts at promoting specific hashtags, including the one used by the CBC's Ormiston Online project. Indeed, from the outset some Conservative bloggers took offense to the CBC's promotion of the #ormistondebate hashtag, with some partisans questioning my own role in this process:

> @greg_elmer … Did you play a part in setting up Ormiston to monitor the following twitter tag #ormistondebate?

Such after-the-fact findings, while further qualifying the expansion of both the time and space (screens and platforms) of micro-blogging during a live broadcast event, also highlight the limits of real-time research, and in particular the inability to conduct expansive, time-consuming reviews of data. Real-time or 'live' research is a bit of a misnomer in that it requires the pre-setting of a research agenda, a method of data collection, and, in this instance at least, a heavy reliance upon other forms of near real-time comparative data (e.g. the CBC's debate transcripts). Live research should therefore be viewed and understood as an effort at developing methods of collecting and analyzing data flows on platforms that hyper-accentuate the present, rather than simply enacting research and analysis in real time.

Conclusions

The collaboration discussed in this article offered a number of researchers the ability to intervene in public debates about the role that new media platforms play in important social and political issues of our day; or in this instance, in the very discourse enacted by our country's political leaders. Scholars of new media suffer perhaps more than most in their frustrations at seeing their work – particularly time-sensitive research – delayed for many months and sometimes years. This, however, is not a call to do away with established forms of peer review and scholarly publishing, but rather to question how new theories, methods and venues for publishing and otherwise making research findings public can begin to address the growing importance of real-time media as a distinct event into itself (e.g. a debate or media event such as a weather-related disaster), or a series of micro-events that in sum offer researchers insight into the structure and effect of 'political cycles', as Chadwick (2011) notes. Live research, as such, serves not only to question and understand the interface time of social media practices and platforms, but also challenges the time-compressed and space-delimited sphere of academic scholarship.

Moving forward, live research needs to distinguish itself as a research project from certain strands of information design – projects that seek to creatively visualize complex datasets and flows in the search for intuitive iconography and dynamic flux (Abrams and Hall, 2006). Live research, in other words, should not only be concerned with re-presenting the world of things or their imprints, but rather work to offer concepts, theories and methods that might critically understand how users mobilize and sustain texts and other digital objects (by uploading, sharing, remixing and downloading) across the field of networked communication. Live research, as such, could serve as an important contingent step in recognizing the ever-shifting social media plane and the tactics deployed to sustain meaningful communication in a socially networked media age.

Funding

Funding for this article was provided by a grant from the Social Science and Humanities Research Council of Canada and from the Korean government's World Class University project.

Notes

1. See Sysomos' September 2010 social media marketing study. Available at http://sysomos.com/insidetwitter/engagement.
2. For an early insider's view of the emergence of rapid-response political tactics in the context of new information and communication technologies (ICTs) see Myers (1993).
3. The case study focused on Canada's fortieth general election. The campaign officially began on 7 September 2008 and ended on voting day, 14 October 2008. More details can be found on the Elections Canada website (www.elections.ca/content.aspx?section=ele&document=index&dir=pas/40ge&lang=e).
4. Canadian convention for televised debates is typically to broadcast in both of Canada's official languages, English and French. This study focuses exclusively on the English-language debate broadcast on 2 October 2008, although a dry run of our methods was informally tested during the French-language debate held the day earlier.
5. The platform launched worldwide in July 2006.

6. A number of projects have since investigated how more established Twitter conventions can help understand the nature of audience feedback and interaction during live broadcast debates. See Anstead and O'Loughlin (2011: 7). See also www.infoscapelab.ca/ontarioelection2007.
8. An archive of the Ormiston Online project can be found at www.cbc.ca/news/canadavotes/campaign2/ormiston/.
9. At the time, YouTube provided only total cumulative views of videos. Working with the platform's API, we wrote a software script that determined on a weekly basis how many views a video received.
10. A number of the posted tweets made reference to switching back and forth between the Canadian party leaders' and American vice-presidential televised debates.
11. The search terms and hashtags included #ormistondebate, the Twitter account names for the Canadian party leaders and campaigns ('jacklayton', 'LiberalTour', 'Pmharper', 'ElizabethMay', 'gillesduceppe') and the search terms 'jack layton', 'elizabeth may', 'gilles duceppe', 'stephane dion' and 'stephen harper'. The total sample included 558 tweets.
12. The search terms and hashtags included #ormistononline, the Twitter account names for the party leaders and campaigns (i.e. jacklayton, LiberalTour, Pmharper, ElizabethMay, gillesduceppe), and the formal names of the federal party leaders ('jack layton', 'elizabeth may','gilles duceppe', 'stephane dion', and 'stephen harper'). The total sample included 558 tweets.
13. The comment was made in reference to the Conservatives' lack of a formal party platform and an advertisement depicting the Conservative Prime Minister in an atypically informal sweater.
14. 'I'm sure it's a coincidence but Jack Layton just paraphrased a Propagandhi song.'

References

Abrams J and Hall P (2006) *Else/Where: Mapping New Cartographies of Networks and Territories*. Minneapolis, MN: University of Minnesota Design Institute.

Allan J (ed.) (2002) *Topic Detection and Tracking: Event-based Information Organization*. Norwell, MA: Kluwer Academic Publishers.

Anstead N and O'Loughlin B (2011) The emerging viewertariat and BBC question time: television debate and real-time commenting online. *International Journal of Press-Politics* 16(4): 440–462.

Bruns A (2010) Politics vs. Masterchef: the view from Twitter. Available at: www.mappingonlinepublics.net/2010/07/26/politics-vs-masterchef-the-view-from-twitter/.

Chadwick A (2011) Britain's first live televised party leaders' debate: from the news cycle to the political information cycle. *Parliamentary Affairs* 64(1): 24–44.

Chadwick A and Howard P (2009) *Handbook of Internet Politics*. London: Routledge.

Cunningham SD (2008) Political and media leadership in the age of YouTube. In: Hart P and Uhr J (eds) *Public Leadership: Perspectives and Practices*. Canberra: Australian National University E Press, pp. 177–186.

Elmer G, Curlew B, Devereaux Z, et al. (2009) Blogs I read: partisanship and party loyalty in the Canadian blogosphere. *Journal of Information Technology & Politics* 6(2): 156–165.

Fuller M (2003) *Behind the Blip: Essays on the Culture of Software*. New York: Autonomedia.

Galloway A (2004) *Protocol: How Control Exists after Decentralization*. Cambridge, MA: MIT Press.

Geere D (2010) It's not just you: 71 percent of tweets are ignored. Available at: www.wired.com/epicenter/2010/10/its-not-just-you-71-percent-of-tweets-are-ignored/.

Gurevich M, Coleman S and Blumler JG (2009) Political communication: old and new media relationships. *Annals of the American Academy of Political and Social Science* 625(1): 164–181.

Harman G (2009) *Prince of Networks: Bruno Latour and Metaphysics*. Melbourne, VIC: Re.Press.

Hine C (2007) Connective ethnography for the exploration of e-science. *Journal of Computer-Mediated Communication* 12(2). Available at: http://jcmc.indiana.edu/vol12/issue2/hine.html.

Jungherr A and Pascal J (2011) One tweet at a time: Mapping political campaigns through social media data. Paper presented at the 6th ECPR General Conference, Reykjavik, Iceland.

Kluver R, Foot K and Jankowski N, et al. (eds) (2007) *The Internet and National Elections: A Comparative Study of Web Campaigning*. London: Routledge.

Lash S (2007) Objects that judge: Latour's parliament of things, *Transversal*. Available at: http://eipcp.net/transversal/0107/lash/en.

Latour B and Weibel P (2005) From realpolitik to dingpolitik: or how to make things public. In: Latour B and Weibel B (eds) *Making Things Public: Atmospheres of Democracy*. Cambridge: MIT Press, pp. 4–31.

Markham AN (1998) *Life Online: Researching Real Experience in Virtual Space*. Lanham, MD and Oxford: Rowman & Littlefield.

Myers D (1993) New technology and the 1992 Clinton presidential campaign. *American Behavioral Scientist* 37: 181–184.

Norris P (2000) *A Virtuous Circle: Political Communications in Postindustrial Societies*. Cambridge: Cambridge University Press.

Rogers R (2006) *Information Politics on the Web*. Cambridge, MA: MIT Press.

Schneider S and Foot K (2010) Object-oriented web historiography. In: Brugger N (ed.) *Web History*. New York: Peter Lang, pp. 61–82.

Greg Elmer (PhD, University of Massachusetts, Amherst) is Associate Professor in the School of Media and the graduate program in Communication and Culture at Ryerson University. He is co-author of *The Permanent Campaign: New Media, New Politics* (forthcoming).

M/C Journal, Vol. 16, No. 2 (2013) - 'mining'

Mining One Percent of Twitter: Collections, Baselines, Sampling

http://journal.media-culture.org.au/index.php/mcjournal/article/view/620

Carolin Gerlitz, Bernhard Rieder

Introduction

Social media platforms present numerous challenges to empirical research, making it different from researching cases in offline environments, but also different from studying the "open" Web. Because of the limited access possibilities and the sheer size of platforms like Facebook or Twitter, the question of *delimitation*, i.e. the selection of subsets to analyse, is particularly relevant. Whilst sampling techniques have been thoroughly discussed in the context of social science research (Uprichard; Noy; Bryman; Gilbert; Gorard), sampling procedures in the context of social media analysis are far from being fully understood. Even for Twitter, a platform having received considerable attention from empirical researchers due to its relative openness to data collection, methodology is largely emergent. In particular the question of how smaller collections relate to the entirety of activities of the platform is quite unclear. Recent work comparing case based studies to gain a broader picture (Bruns and Stieglitz) and the development of graph theoretical methods for sampling (Papagelis, Das, and Koudas) are certainly steps in the right direction, but it seems that truly large-scale Twitter studies are limited to computer science departments (e.g. Cha *et al*.; Hong, Convertino, and Chi), where epistemic orientation can differ considerably from work done in the humanities and social sciences.

The objective of the paper is to reflect on the affordances of different techniques for making Twitter collections and to suggest the use of a random sampling technique, made possible by Twitter's Streaming API (Application Programming Interface), for baselining, scoping, and contextualising practices and issues. We discuss this technique by analysing a one percent sample of all tweets posted during a 24-hour period and introduce a number of analytical directions that we consider useful for qualifying some of the core elements of the platform, in particular hashtags. To situate our proposal, we first discuss how platforms propose particular affordances but leave considerable margins for the emergence of a wide variety of practices. This argument is then related to the question of how medium and sampling technique are intrinsically connected.

Indeterminacy of Platforms

A variety of new media research has started to explore the material-technical conditions of platforms (Rogers`; Gillespie; Hayles), drawing attention to the performative capacities of platform protocols to enable and structure specific activities; in the case of Twitter that refers to elements such as tweets, retweets, @replies, favourites, follows, and lists. Such features and conventions have been both a subject and a starting point for researching platforms, for instance by using hashtags to demarcate topical conversations (Bruns and Stieglitz), @replies to trace interactions, or following relations to establish social networks (Paßmann, Boeschoten, and Schäfer). The emergence of platform studies (Gillespie; Montfort and Bogost; Langlois *et al*.) has drawn attention to platforms as interfacing infrastructures that offer blueprints for user activities through technical and interface affordances that are pre-defined yet underdetermined, fostering sociality in the front end whilst mining for data in the back end (Stalder). Doing so, they cater to a variety of actors, including users, developers, advertisers, and third-party services, and allow for a variety of distinct use practices to emerge. The use practices of platform features on Twitter are, however, not solely produced by users themselves, but crystallise in relation to wider ecologies of platforms, users, other media, and third party services (Burgess and Bruns), allowing for sometimes unanticipated vectors of development. This becomes apparent in the case of the retweet function, which was initially introduced by users as verbatim operation, adding "retweet" and later "RT" in front of copied content, before Twitter officially offered a retweet button in 2009 (boyd, Golder, and Lotan). Now, retweeting is deployed for a series of objectives, including information dissemination, promotion of opinions, but also ironic commentary.

Gillespie argues that the capacity to interface and create relevance for a variety of actors and use practices is, in fact, the central characteristic of platforms (Gillespie). Previous research for instance addresses Twitter as medium for public participation in specific societal issues (Burgess and Bruns; boyd, Golder, and Lotan), for personal conversations (Marwick and boyd; boyd, Golder, and Lotan), and as facilitator of platform-specific communities (Paßmann, Boeschoten, and Schäfer). These case-based studies approach and demarcate their objects of study by focussing on particular hashtags or use practices such as favoriting and retweeting.

But using these elements as basis for building a collection of tweets, users, etc. to be analysed has significant epistemic weight: these sampling methods come with specific notions of use scenarios built into them or, as Uprichard suggests, there are certain "a priori philosophical assumptions intrinsic to any sample design and the subsequent validity of the sample criteria themselves" (Uprichard 2). Building collections by gathering tweets containing specific hashtags, for example, assumes that a) the conversation is held together by hashtags and b) the chosen hashtags are indeed the most relevant ones. Such assumptions go beyond the statistical question of *sampling bias* and concern the fundamental problem of how to go fishing in a pond that is big, opaque, and full of quickly evolving populations of fish. The classic information retrieval concepts of *recall* (How many of the relevant fish did I get?) and *precision* (How many fish caught are relevant?) fully apply in this context. In a next step, we turn more directly to the question of sampling Twitter, outlining which methods allow for accessing which practices – or not – and what the role of medium-specific features is.

Sampling Twitter

Sampling, the selection of subsets from a larger set of elements (the population), has received wide attention especially in the context of empirical sociology (Uprichard; Noy; Bryman; Gilbert; Gorard; Krishnaiah and Rao). Whilst there is considerable overlap in sampling practices between quantitative sociology and social media research, some key differences have to be outlined: first, social media data, such as tweets, generally pre-exist their collection rather than having to be produced through surveys; secondly, they come in formats specific to platforms, with analytical features, such as counts, already built into them (Marres and Weltevrede); and third, social media assemble very large populations, yet selections are rarely related to full datasets or grounded in baseline data as most approaches follow a case study design

(Rieder).

There is a long history to sampling in the social sciences (Krishnaiah and Rao), dating back to at least the 19th century. Put briefly, modern sampling approaches can be distinguished into probability techniques, emphasising the representative relation between the entire population and the selected sample, and non-probability techniques, where inference on the full population is problematic (Gilbert). In the first group, samples can either be based on a fully random selection of cases or be stratified or cluster-based, where units are randomly selected from a proportional grid of known subgroups of a population. Non-probability samples, on the contrary, can be representative of the larger population, but rarely are. Techniques include accidental or convenience sampling (Gorard), based on ease of access to certain cases. Purposive non-probability sampling however, draws on expert sample demarcation, on quota, case-based or snowball sampling techniques – determining the sample via *a priori* knowledge of the population rather than strict representational relations. Whilst the relation between sample and population, as well as access to such populations (Gorard) is central to all social research, social media platforms bring to the reflection of how samples can function as "knowable objects of knowledge" (Uprichard 2) the role of medium-specific features, such as built-in markers or particular forms of data access.

Ideally, when researching Twitter, we would have access to a *full sample*, the subject and phantasy of many *big data* debates (boyd and Crawford; Savage and Burrows), which in practice is often limited to platform owners. Also, growing amounts of daily tweets, currently figuring around 450 million (Farber), require specific logistic efforts, as a project by Cha *et al*. indicates: to access the tweets of 55 million user accounts, 58 servers to collect a total amount of 1.7 billion tweets (Cha *et al*.). Full samples are particularly interesting in the case of *exploratory data analysis* (Tukey) where research questions are not set before sampling occurs, but emerge in engagement with the data.

The majority of sampling approaches on Twitter, however, follow a non-probabilistic, non-representative route, delineating their samples based on features specific to the platform.

The most common Twitter sampling technique is *topic-based sampling* that selects tweets via hashtags or search queries, collected through API calls (Bruns and Stieglitz, Burgees and Bruns; Huang, Thornton, and Efthimiadis) Such sampling techniques rest on the idea that content will group around the shared use of hashtags or topical words. Here, hashtags are studied with an interest in the emergence and evolution of topical concerns (Burgees and Bruns), to explore brand communication (Stieglitz and Krüger), during public unrest and events (Vis), but also to account for the multiplicity of hashtag use practices (Bruns and Stieglitz). The approach lends itself to address issue emergence and composition, but also draws attention to medium-specific use practices of hashtags.

Snowball sampling, an extension of topic-based sampling, builds on predefined lists of user accounts as starting points (Rieder), often defined by experts, manual collections or existing lists, which are then extended through "snowballing" or triangulation, often via medium-specific relations such as following. Snowball sampling is used to explore national spheres (Rieder), topic- or activity-based user groups (Paßmann, Boeschoten, and Schäfer), cultural specificity (Garcia-Gavilanes, Quercia, and Jaimes) or dissemination of content (Krishnamurthy, Gill, and Arlitt). Recent attempts to combine random sampling and graph techniques (Papagelis, Das, and Koudas) to throw wider nets while containing technical requirements are promising, but conceptually daunting.

Marker-based sampling uses medium-specific metadata to create collections based on shared language, location, Twitter client, nationality or other elements provided in user profiles (Rieder). This sampling method can be deployed to study the language or location specific use of Twitter. However, an increasing amount of studies develop their own techniques to detect languages (Hong, Convertino, and Chi).

Non-probability selection techniques, topic-, marker-, and basic graph-based sampling struggle with representativeness (Are my results generalisable?), exhaustiveness (Did I capture all the relevant units?), cleanness (How many irrelevant units did I capture?), and scoping (How "big" is my set compared to others?), which does – of course – not invalidate results. It does, however, raise questions about the generality of derived claims, as case-based approaches only allow for sense-making from inside the sample and not in relation to the entire population of tweets. Each of these techniques also implies commitments to *a priori* conceptualisations of Twitter practices: snowball sampling presupposes coherent network topologies, marker-based sampling has to place a lot of faith in Twitter's capacity to identify language or location, and topic-based samples consider words or hashtags to be *sufficient* identifiers for issues. Further, specific sampling techniques allow for studying issue *or* medium dynamics, and provide insights to the negotiation of topical concerns versus the specific use practices and medium operations on the platform.

Following our interest in relations between sample, population and medium-specificity, we therefore turn to *random sampling*, and ask whether it allows to engage Twitter without commitments – or maybe *different* commitments? – to particular *a priori* conceptualisations of practices. Rather than framing the relation between this and other sampling techniques in oppositional terms, we explore in what way it might serve as baseline foil, investigating the possibilities for relating non-probability samples to the entire population, thereby embedding them in a "big picture" view that provides context and a potential for inductive reasoning and exploration. As we ground our arguments in the analysis of a concrete random sample, our approach can be considered *experimental*.

Random Sampling with the Streaming API

While much of the developer API features Twitter provides are "standard fare", enabling third party applications to offer different interfaces to the platform, the so-called Streaming API is unconventional in at least two ways. First, instead of using the common query-response logic that characterises most REST-type implementations, the Streaming API requires a persistent connection with Twitter's server, where tweets are then pushed in near real-time to the connecting client. Second, in addition to being able to "listen" to specific keywords or usernames, the logic of the *stream* allows Twitter to offer a form of data access that is circumscribed in quantitative terms rather than focussed on particular entities. The so called *statuses/firehose* endpoint provides the full stream of tweets to selected clients; the *statuses/sample* endpoint, however, "returns a small random sample of all public statuses" with a size of one percent of the full stream. (In a forum post, Twitter's senior partner engineer, Taylor Singletary, states: "The sample stream is a random sample of 1% of the tweets being issues [*sic*] publicly.") If we estimate a daily tweet volume of 450 million tweets (Farber), this would mean that, in terms of standard sampling theory, the 1% endpoint would provide a representative and high resolution sample with a maximum margin of error of 0.06 at a confidence level of 99%, making the study of even relatively small subpopulations within that sample a realistic option.

While we share the general prudence of boyd and Crawford when it comes to the validity of this sample stream, a technical analysis of the Streaming API indicates that some of their caveats are unfounded: because tweets appear in near real-time in the queue (our tests show that tweets are delivered via the API approx. 2 seconds after they are sent), it is clear that the system does not pull only "the first few thousand tweets per hour" (boyd and Crawford 669); because the sample is most likely a simple filter on the *statuses/firehose* endpoint, it would be technically impractical to include only "tweets from a particular segment of the network graph" (ibid.). Yet, without access to the complete stream, it is difficult to fully assess the selection bias of the different APIs (González-Bailón, Wang, and Rivero). A series of tests in which we compared the sample to the full output of high volume bot accounts can serve as an indicator: in particular, we looked into the activity of *SportsAB*, *Favstar_Bot*, and *TwBirthday*, the three most active accounts in our sample (respectively 38, 28, and 27 tweets captured). Although Twitter communicates a limit of 1000 tweets per day and account, we found that these bots consistently post over 2500 messages in a 24 hour period. *SportsAB* attempts to post 757 tweets every three hours, but runs into *some* limit every now and then. For every successful peak, we captured between five and eight messages, which indicates a pattern consistent with a random selection procedure. While more testing is needed, various elements indicate that the *statuses/sample* endpoint provides data that are indeed representative of all public tweets.

Using the soon to be open-sourced *Digital Methods Initiative Twitter Capture and Analysis Toolset* (DMI-TCAT) we set out to test the method and the insights that could be derived from it by capturing 24 hours of Twitter activity, starting on 23 Jan. 2013 at 7 p.m. (GMT). We captured 4,376,230 tweets, sent from 3,370,796 accounts, at an average rate of 50.65 tweets per second, leading to about 1.3GB of uncompressed and unindexed MySQL tables. While a truly robust approach would require a longer period of data capture, our main goal – to investigate how the Streaming API can function as a "big picture" view of Twitter and as baseline for other sampling methods – led us to limit ourselves to a manageable corpus. We do not propose our 24-hour dataset to function as a baseline in itself, but to open up reflections about representative metrics and the possibilities of baseline sampling in general. By making our scripts public, we hope to facilitate the creation of (background) samples for other research projects. (DMI-TCAT is developed by Erik Borra and Bernhard Rieder. The stream capture scripts are already available at https://github.com/bernorieder/twitterstreamcapture.)

A Day of Twitter

Exploring how the Twitter one percent sample can provide us with a contrast foil against other collection techniques, we suggest that it might allow to create relations between entire populations, samples and medium-specific features in different ways; as illustration, we explore four of them.

a) Tweet Practices Baseline:

Figure 1 shows the temporal baseline, giving indications for the pace and intensity of activity during the day. The temporal pattern features a substantial dip in activity, which corresponds with the fact that around 60% of all tweets have English language settings, which might indicate sleeping time for English-speaking users.

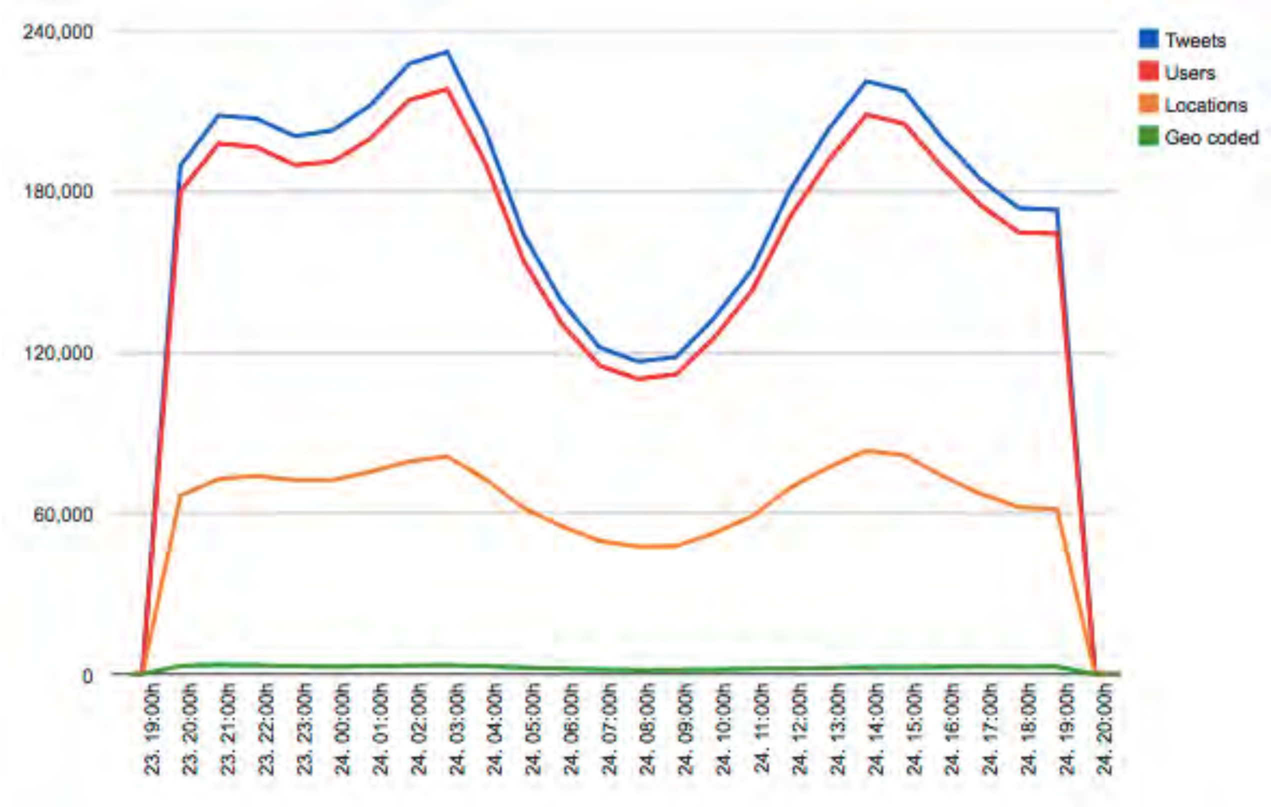


Figure 1: temporal patterns

Exploring the composition of users, the sample shows how "communicative" Twitter is; the 3,370,796 unique users we captured mentioned (all "@username" variants) 2,034,688 user accounts. Compared to the random sample of tweets retrieved by boyd *et al*. in 2009, our sample shows differences in use practices (boyd, Golder, and Lotan): while the number of tweets with hashtags is significantly higher (yet small in relation to all tweets), the frequency of URL use is lower. While these averages gloss over significant variations in use patterns between subgroups and languages (Poblete *et al*.), they do provide a baseline to relate to when working with a case-based collection.

Tweets containing	boyd *et al*. 2010	our findings
a hashtag	5%	13.18%
a URL	22%	11.7%

an @user mention	36%	57.2%
tweets beginning with @user	86%	46.8%

Table 1: Comparison between boyd *et al*. and our findings

b) Hashtag Qualification:

Hashtags have been a focus of Twitter research, but reports on their use vary. In our sample, 576,628 tweets (13.18%) contained 844,602 occurrences of 227,029 unique hashtags. Following the typical power law distribution, only 25.8% appeared more than once and only 0.7% (1,684) more than 50 times. These numbers are interesting for characterising Twitter as a platform, but can also be useful for situating individual cases against a quantitative baseline. In their hashtag metrics, Bruns and Stieglitz suggest a categorisation derived from *a priori* discussions of specific use cases and case comparison in literature (Bruns and Stieglitz). The random sample, however, allows for alternative, *a posteriori* qualifying metrics, based on emergent topic clusters, co-appearance and proximity measures.

Beyond purely statistical approaches, co-word analysis (Callon *et al*.) opens up a series of perspectives for characterising hashtags in terms of how they appear together with others. Based on the basic principle that hashtags mentioned in the same tweet can be considered *connected*, networks of hashtags can be established via graph analysis and visualisation techniques – in our case with the help of Gephi.

Our sample shows a high level of connectivity between hashtags: 33.8% of all unique hashtags are connected in a giant component with an average degree (number of connections) of 6.9, a diameter (longest distance between nodes) of 15, and an average path length between nodes of 12.7. When considering the 10,197 hashtags that are connected to at least 10 others, the network becomes much denser, though: the diameter shrinks to 9 and the average path length of 3.2 indicates a "small world" of closely related topic spaces.

Looking at how hashtags relate to this connected component, we detect that out of the 1,684 hashtags with a frequency higher than 50, 96.6% are part of it, while the remaining 3.4% are spam hashtags that are deployed by a single account only. In what follows, we focus on the 1,627 hashtags that are part of the giant component.

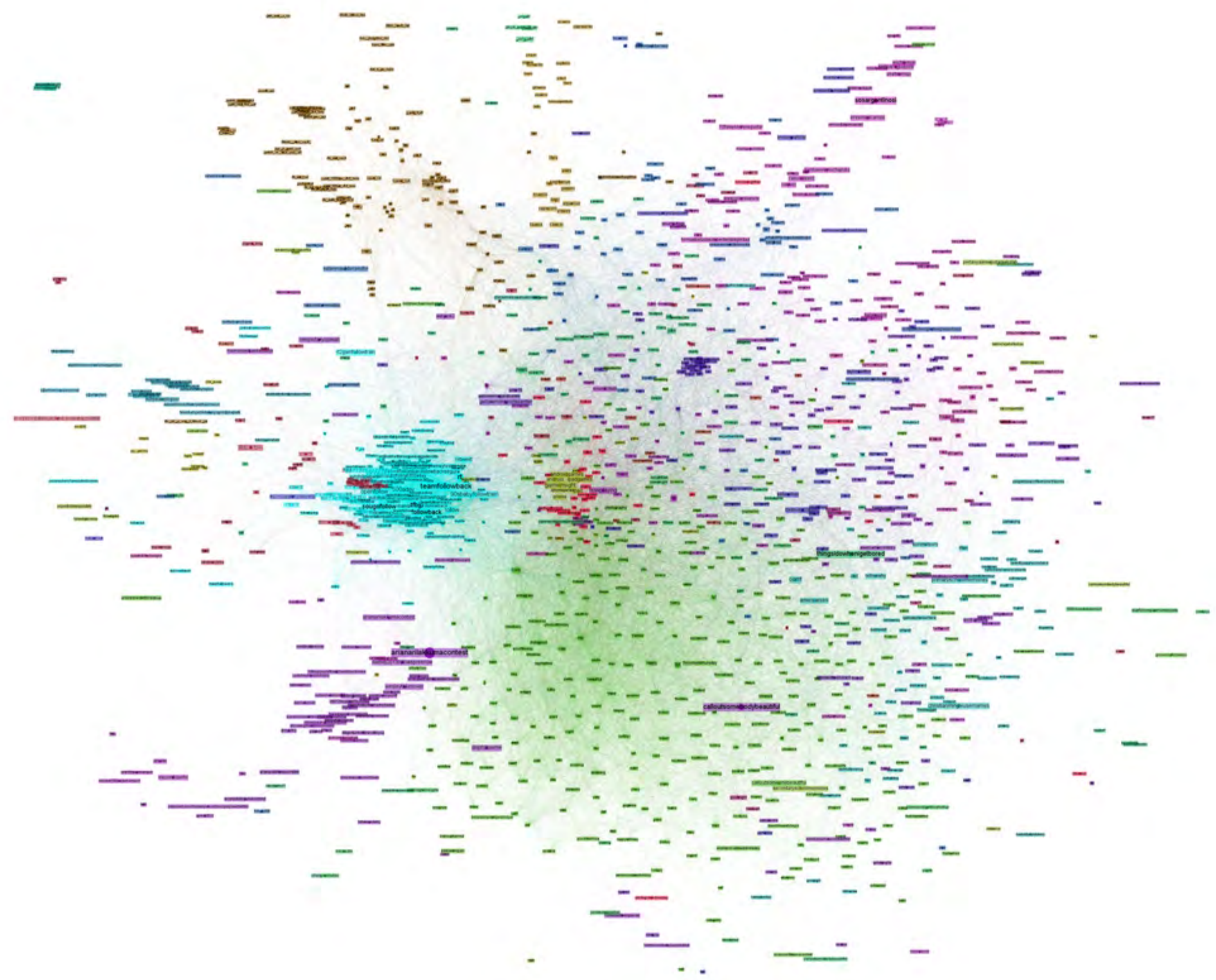


Figure 2: Co-occurrence map of hashtags
(spatialisation: Force Atlas 2; size: frequency of occurrence; colour: communities detected by modularity)

As shown in Figure 2, the resulting network allows us to identify topic clusters with the help of “community” detection techniques such as the *Gephi modularity* algorithm. While there are clearly identifiable topic clusters, such as a dense, high frequency cluster dedicated to following in turquoise (#teamfollowback, #rt, #followback and #sougofollow), a cluster concerning Arab countries in brown or a pornography cluster in bright red, there is a large, diffuse zone in green that one could perhaps most fittingly describe as “everyday life” on Twitter, where food, birthdays, funny images, rants, and passion can coexist. This *zone* – the term cluster suggesting too much coherence – is pierced by celebrity excitement (#arianarikkumacontest) or moments of social banter (#thingsidowhenigetbored, #calloutsomeonebeautiful) leading to high tweet volumes.

Figures 3 and 4 attempt to show how one can use network metrics to qualify – or even classify – hashtags based on how they connect to others. A simple metric such as a node's *degree*, i.e. its number of connections, allows us to distinguish between "combination" hashtags that are not topic-bound (#love, #me, #lol, #instagram, the various "follow" hashtags) and more specific topic markers (#arianarikkumacontest, #thingsidowhenigetbored, #calloutsomeonebeautiful, #sosargentinosi).

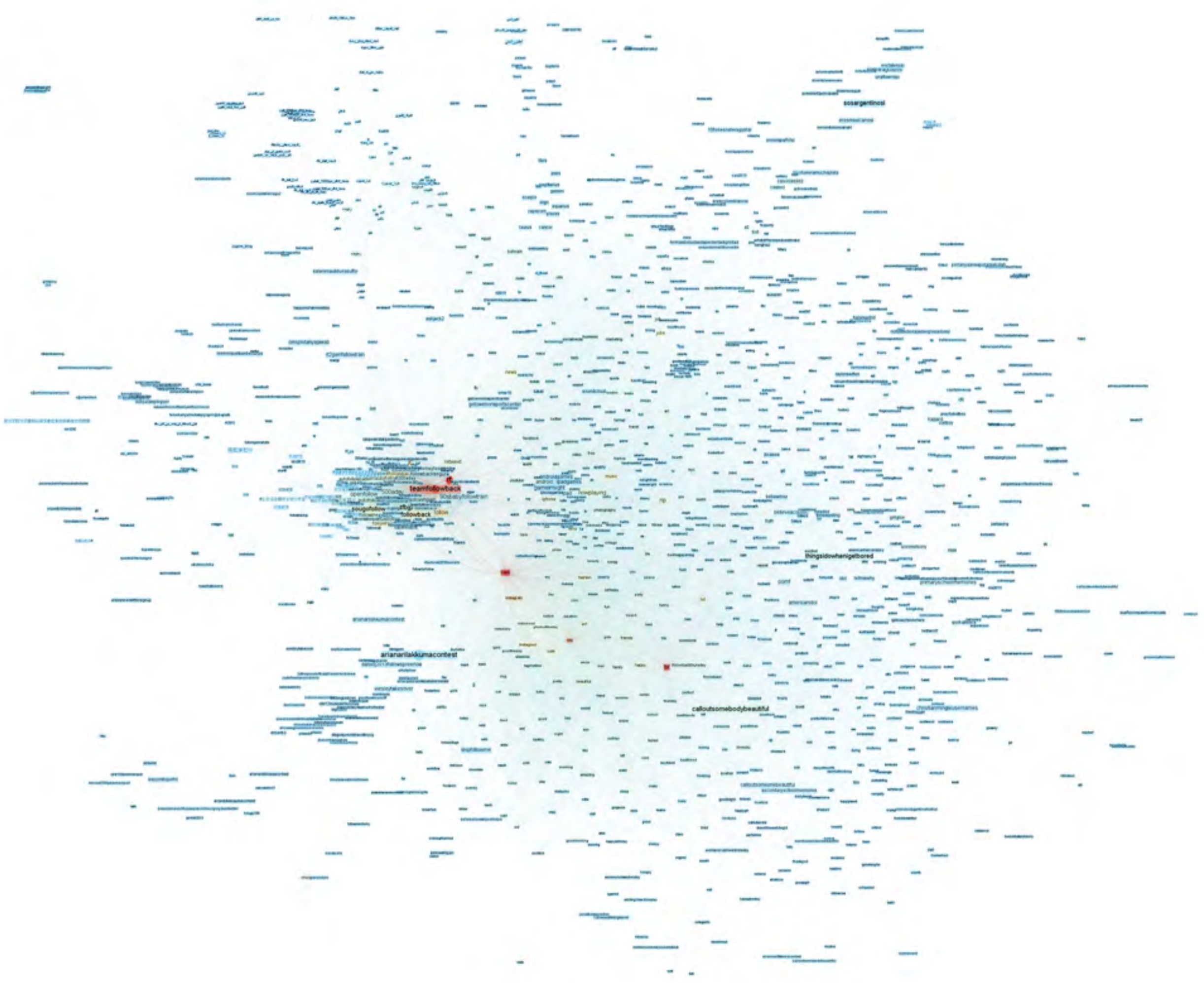


Figure 3: Co-occurrence map of hashtags
(spatialisation: Force Atlas 2; size: frequency of occurrence; colour (from blue to yellow to red): degree)

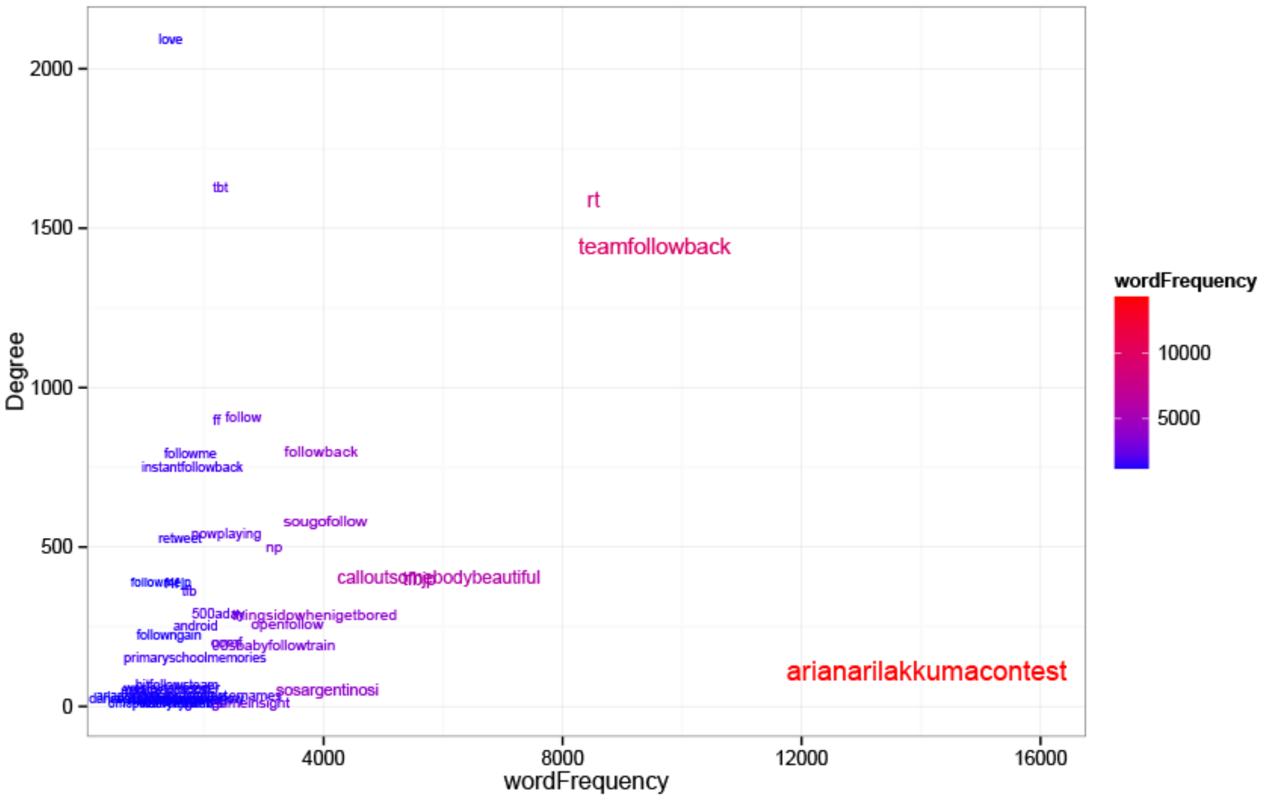


Figure 4: Hashtag co-occurrence in relation to frequency

Another metric, which we call “user diversity”, can be derived by dividing the number of unique users of a hashtag by the number of tweets it appears in, normalised to a percentage value. A score of 100 means that no user has used the hashtag twice, while a score of 1 indicates that the hashtag in question has been used by a single account. As Figures 5 and 6 show, this allows us to distinguish hashtags that have a “shoutout” character (#thingsidowhenigetbored, #calloutsomeonebeautiful, #love) from terms that become more “insisting”, moving closer to becoming spam.

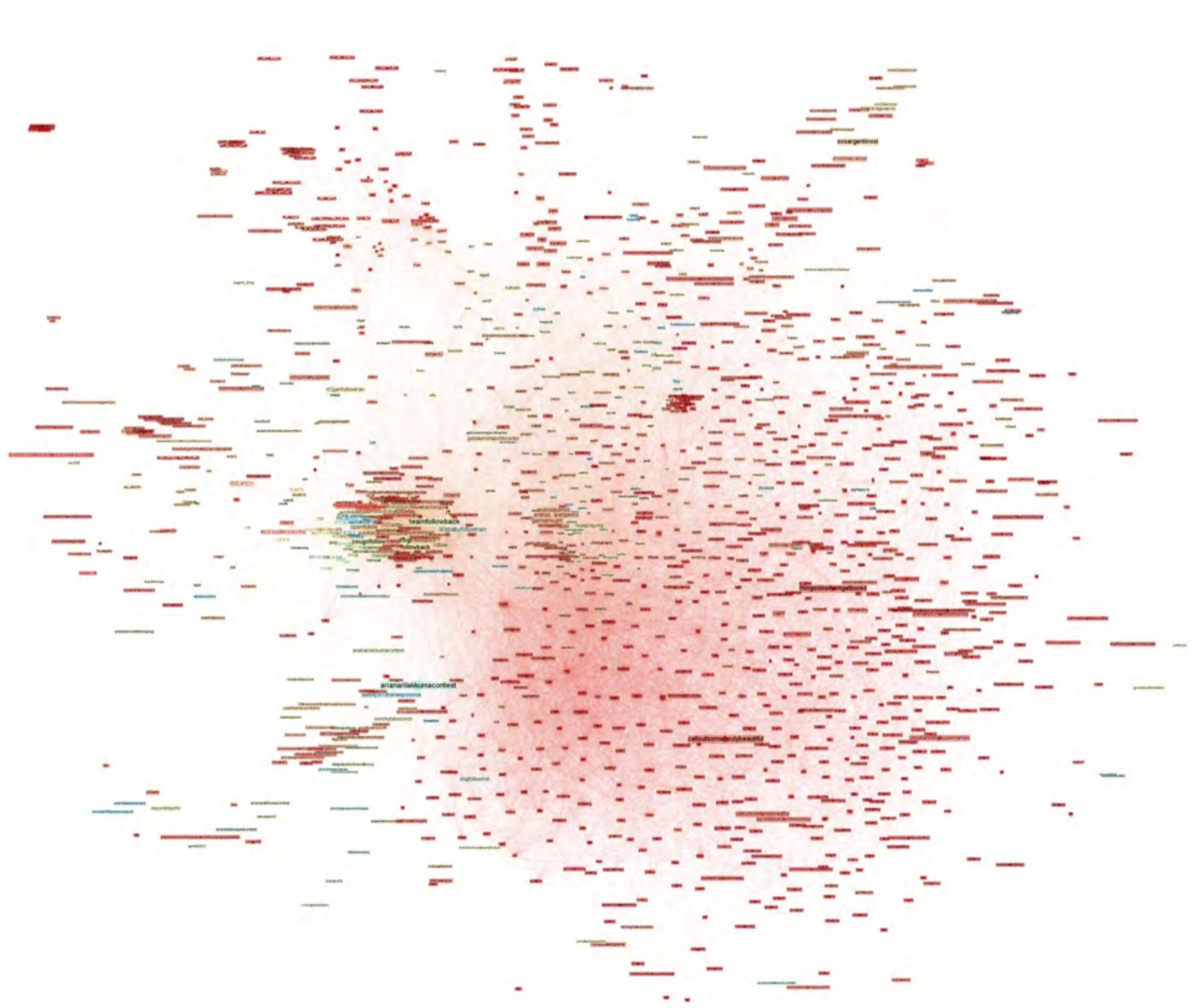


Figure 5: Co-occurrence map of hashtags
(spatialisation: Force Atlas 2; size: frequency of occurrence; colour (from blue to yellow to red): user diversity)

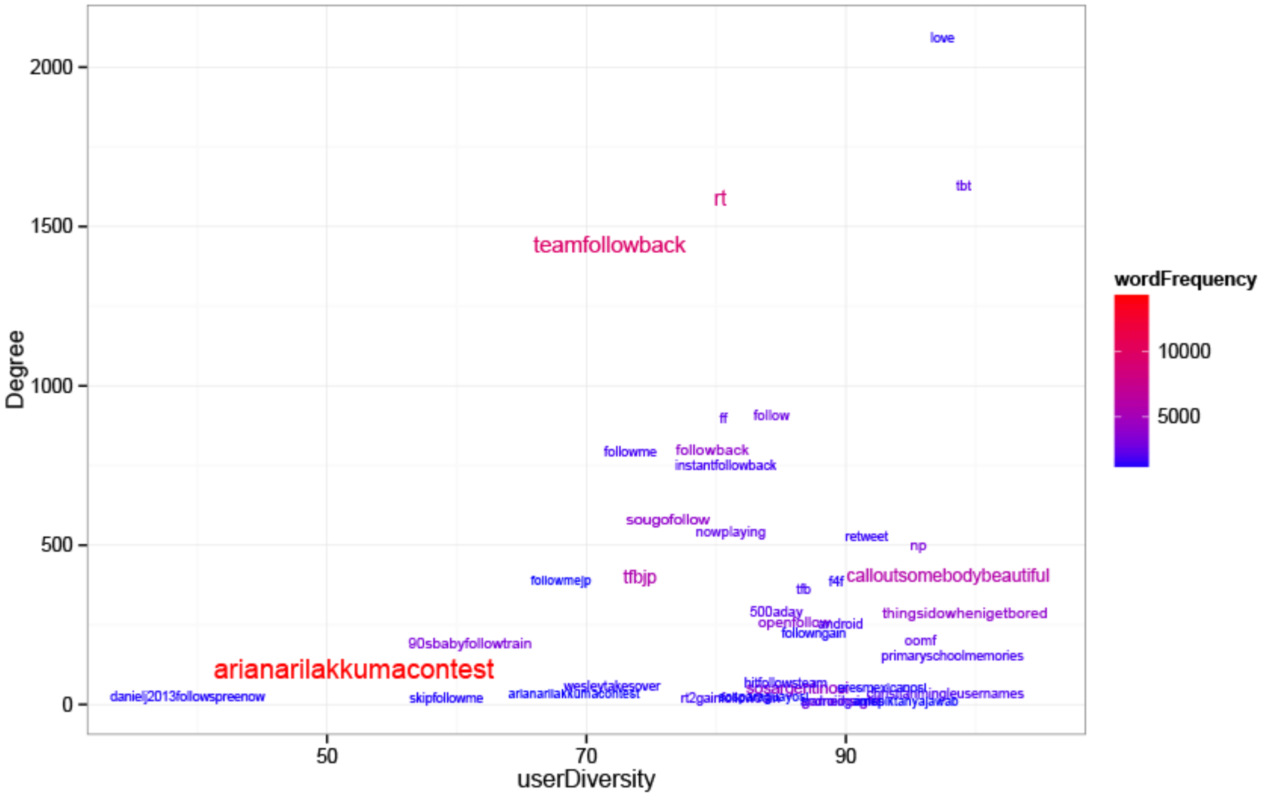
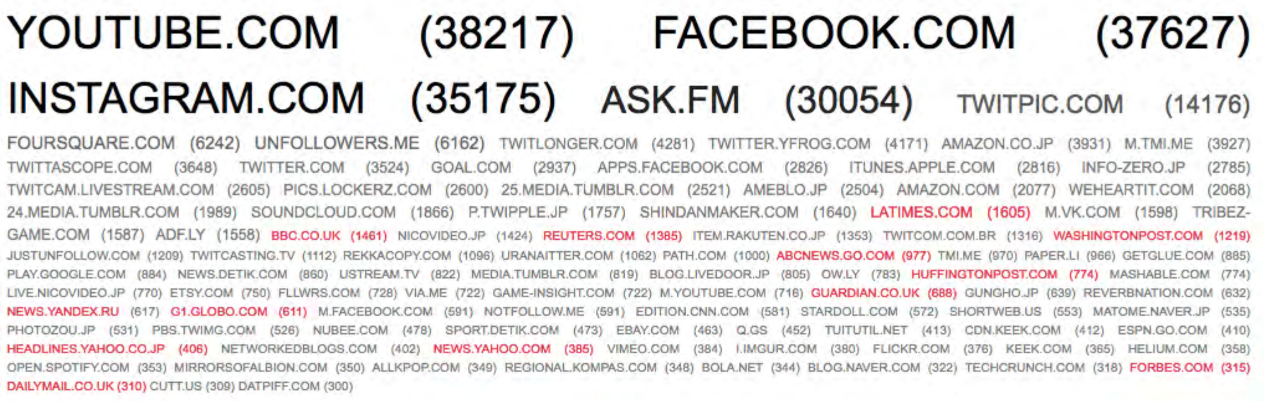


Figure 6: Hashtag user diversity in relation to frequency

All of these techniques, beyond leading to findings in themselves, can be considered as a useful backdrop for other sampling methods. Keyword- or hashtag-based sampling is often marred by the question of whether the "right" queries have been chosen; here, co-hashtag analysis can easily find further related terms – the same analysis is possible for keywords also, albeit with a much higher cost in computational resources.

c) Linked Sources:

Only 11% of all tweets contained URLs, and our findings show a power-law distribution of linked sources. The highly shared domains indicate that Twitter is indeed a predominantly "social" space, with a high presence of major social media, photo-sharing (Instagram and Twitpic) and Q&A platforms (ask.fm). News sources, indicated in red in figure 7, come with little presence – although we acknowledge that this might be subject to daily variation.



YOUTUBE.COM (38217) FACEBOOK.COM (37627) INSTAGRAM.COM (35175) ASK.FM (30054) TWITPIC.COM (14176) FOURSQUARE.COM (6242) UNFOLLOWERS.ME (6162) TWITLONGER.COM (4281) TWITTER.YFROG.COM (4171) AMAZON.CO.JP (3931) M.TMI.ME (3927) TWITTASCOPE.COM (3648) TWITTER.COM (3524) GOAL.COM (2937) APPS.FACEBOOK.COM (2826) ITUNES.APPLE.COM (2816) INFO-ZERO.JP (2785) TWITCAM.LIVESTREAM.COM (2605) PICS.LOCKERZ.COM (2600) 25.MEDIA.TUMBLR.COM (2521) AMEBLO.JP (2504) AMAZON.COM (2077) WEHEARTIT.COM (2068) 24.MEDIA.TUMBLR.COM (1989) SOUNDCLOUD.COM (1866) P.TWIPPLE.JP (1757) SHINDANMAKER.COM (1640) LATIMES.COM (1605) M.VK.COM (1598) TRIBEZ-GAME.COM (1587) ADF.LY (1558) BBC.CO.UK (1461) NICOVIDEO.JP (1424) REUTERS.COM (1385) ITEM.RAKUTEN.CO.JP (1353) TWITCOM.COM.BR (1316) WASHINGTONPOST.COM (1219) JUSTUNFOLLOW.COM (1209) TWITCASTING.TV (1112) REKKACOPY.COM (1096) URANAITTER.COM (1062) PATH.COM (1000) ABCNEWS.GO.COM (977) TMI.ME (970) PAPER.LI (966) GETGLUE.COM (885) PLAY.GOOGLE.COM (884) NEWS.DETIK.COM (860) USTREAM.TV (822) MEDIA.TUMBLR.COM (819) BLOG.LIVEDOOR.JP (805) OW.LY (783) HUFFINGTONPOST.COM (774) MASHABLE.COM (774) LIVE.NICOVIDEO.JP (770) ETSY.COM (750) FLLWRS.COM (728) VIA.ME (722) GAME-INSIGHT.COM (722) M.YOUTUBE.COM (716) GUARDIAN.CO.UK (688) GUNGHO.JP (639) REVERBNATION.COM (632) NEWS.YANDEX.RU (617) G1.GLOBO.COM (611) M.FACEBOOK.COM (591) NOTFOLLOW.ME (591) EDITION.CNN.COM (581) STARDOLL.COM (572) SHORTWEB.US (553) MATOME.NAVER.JP (535) PHOTOZOU.JP (531) PBS.TWIMG.COM (526) NUBEE.COM (478) SPORT.DETIK.COM (473) EBAY.COM (463) Q.GS (452) TUITUTIL.NET (413) CDN.KEEK.COM (412) ESPN.GO.COM (410) HEADLINES.YAHOO.CO.JP (406) NETWORKEDBLOGS.COM (402) NEWS.YAHOO.COM (385) VIMEO.COM (384) I.IMGUR.COM (380) FLICKR.COM (376) KEEK.COM (365) HELIUM.COM (358) OPEN.SPOTIFY.COM (353) MIRRORSOFALBION.COM (350) ALLKPOP.COM (349) REGIONAL.KOMPAS.COM (348) BOLA.NET (344) BLOG.NAVER.COM (322) TECHCRUNCH.COM (318) FORBES.COM (315) DAILYMAIL.CO.UK (310) CUTT.US (309) DATPIFF.COM (300)

Figure 7: Most mentioned URLs by domain, news organisations in red

d) Access Points:

Previously, the increase of daily tweets has been linked to the growing importance of mobile devices (Farber), and relatedly, the sample shows a proliferation of access points. They follow a long-tail distribution: while there are 18,248 unique sources (including tweet buttons), 85.7% of all tweets are sent by the 15 dominant applications. Figure 8 shows that the Web is still the most common access point, closely followed by the iPhone. About 51.7% of all tweets were sent from four mobile platforms (iPhone, Android, Blackberry, and Twitter's mobile Web page), confirming the importance of mobile devices. This finding also highlights the variety and complexity of the contexts that Twitter practices are embedded in.

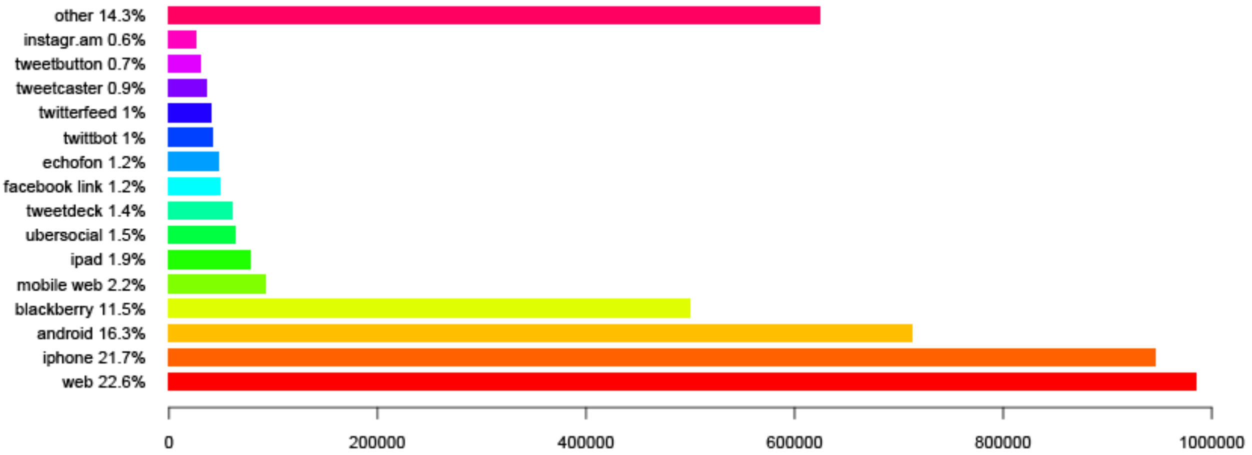


Figure 8: Twitter access points

Conclusion

Engaging with the one percent Twitter sample allows us to draw three conclusions for social media mining. First, thinking of sampling as the making of "knowable objects of knowledge" (Uprichard 2), it entails bringing data points into different relations with each other. Just as Mackenzie contends in relation to databases that it is not the individual data points that matter but the relations that can be created between them (Mackenzie), sampling involves such bringing into relation of medium-specific objects and activities. Small data collection techniques based on queries, hashtags, users or markers, however, do not relate to the whole population, but are defined by internal and comparative relations, whilst random samples are based on the relation between the sample and the full dataset.

Second, thinking sampling as assembly, as relation-making between parts, wholes and the medium thus allows research to adjust its focus on either issue or medium dynamics. Small sample research, we suggested, comes with an investment into specific use scenarios and the subsequent validity of how the collection criteria themselves are grounded in medium specificity. The properties of a "relevant" collection strategy can be found in the extent to which use practices align with and can be utilised to create the collection. Conversely, a mismatch between medium-specific use practices and sample purposes may result in skewed findings. We thus suggest that sampling should not only attend to the internal relations between data points within collections, but also to the relation between the collection and a baseline.

Third, in the absence of access to a full sample, we propose that the random sample provided through the Streaming API can serve as baseline for case approaches in principle. The experimental study discussed in our paper enabled the establishment of a starting point for future long-term data collection from which such baselines can be developed. It would allow to ground *a priori* assumptions intrinsic to small data collection design in medium-specificity and user practices, determining the relative importance of hashtags, URLs, @user mentions. Although requiring more detailed specification, such accounts of internal composition, co-occurrence or proximity of hashtags and keywords may provide foundations to situate case-samples, to adjust and specify queries or to approach hashtags as parts of wider issue ecologies. To facilitate this process logistically, we have made our scripts freely available.

We thus suggest that sampling should not only attend to the internal or comparative relations, but, if possible, to the entire population – captured in the baseline – so that medium-specificity is reflected both in specific sampling techniques and the relative relevance of practices within the platform itself.

Acknowledgements

This project has been initiated in a Digital Methods Winter School project called "One Percent of Twitter" and we would like to thank our project members Esther Weltevrede, Julian Ausserhofer, Liliana Bounegru, Guilio Fagolini, Nicholas Makhortykh, and Lonneke van der Velden. Further gratitude goes to Erik Borra for his useful feedback and work on the DMI-TCAT. Finally, we would like to thank our reviewers for their constructive comments.

References

boyd, danah, and Kate Crawford. "Critical Questions for Big Data." *Information, Communication & Society* 15.5 (2012): 662–679.

———, Scott Golder, and Gilad Lotan. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." *2010 43rd Hawaii International Conference on System Sciences*. IEEE, (2010). 1–10.

Bruns, Axel, and Stefan Stieglitz. "Quantitative Approaches to Comparing Communication Patterns on Twitter." *Journal of Technology in Human Services* 30.3-4 (2012): 160–185.

Bryman, Alan. *Social Research Methods*. Oxford University Press, (2012).

Burgess, Jean, and Axel Bruns. "Twitter Archives and the Challenges of 'Big Social Data' for Media and Communication Research." *M/C Journal* 15.5 (2012). 21 Apr. 2013 <http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561>.

Callon, Michel, *et al*. "From Translations to Problematic Networks: An Introduction to Co-word Analysis." *Social Science Information* 22.2 (1983): 191–235.

Cha, Meeyoung, *et al*. "Measuring User Influence in Twitter: The Million Follower Fallacy." *ICWSM '10: Proceedings of the International AAAI Conference on Weblogs* *and Social Media*. (2010).

Farber, Dan. "Twitter Hits 400 Million Tweets per Day, Mostly Mobile." *cnet*. (2012). 25 Feb. 2013 ‹http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/›.

Garcia-Gavilanes, Ruth, Daniele Quercia, and Alejandro Jaimes. "Cultural Dimensions in Twitter: Time, Individualism and Power." (2006). 25 Feb. 2013 ‹http://www.ruthygarcia.com/papers/cikm2011.pdf›.

Gilbert, Nigel. *Researching Social Life*. Sage, 2008.

Gillespie, Tarleton. "The Politics of 'Platforms'." *New Media & Society* 12.3 (2010): 347–364.

González-Bailón, Sandra, Ning Wang, and Alejandro Rivero. "Assessing the Bias in Communication Networks Sampled from Twitter." 2012. 3 Mar. 2013 ‹http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2185134›.

Gorard, Stephan. *Quantitative Methods in Social Science*. London: Continuum, 2003.

Hayles, N. Katherine. *My Mother Was a Computer: Digital Subjects and Literary Texts*. Chicago: University of Chicago Press, 2005.

Hong, Lichan, Gregorio Convertino, and Ed H Chi. "Language Matters in Twitter : A Large Scale Study Characterizing the Top Languages in Twitter Characterizing Differences Across Languages Including URLs and Hashtags." Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011): 518–521.

Huang, Jeff, Katherine M. Thornton, and Efthimis N. Efthimiadis. "Conversational Tagging in Twitter." *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia – HT '10* (2010): 173.

Krishnamurthy, Balachander, Phillipa Gill, and Martin Arlitt. "A Few Chirps about Twitter." *Proceedings of the First Workshop on Online Social Networks – WOSP '08*. New York: ACM Press, 2008. 19.

Krishnaiah, P R, and C.R. Rao. *Handbook of Statistics*. Amsterdam: Elsevier Science Publishers, 1987.

Langlois, Ganaele, *et al*. "Mapping Commercial Web 2 . 0 Worlds: Towards a New Critical Ontogenesis." Fibreculture 14 (2009): 1–14.

Mackenzie, Adrian. "More Parts than Elements: How Databases Multiply." *Environment and Planning D: Society and Space* 30.2 (2012): 335 – 350.

Marres, Noortje, and Esther Weltevrede. "Scraping the Social? Issues in Real-time Social Research." *Journal of Cultural Economy* (2012): 1–52.

Marwick, Alice, and danah boyd. "To See and Be Seen: Celebrity Practice on Twitter." *Convergence: The International Journal of Research into New Media Technologies* 17.2 (2011): 139–158.

Montfort, Nick, and Ian Bogost. *Racing the Beam: The Atari Video Computer System*.

MIT Press, 2009.

Noy, Chaim. "Sampling Knowledge: The Hermeneutics of Snowball Sampling in Qualitative Research." *International Journal of Social Research Methodology* 11.4 (2008): 327–344.

Papagelis, Manos, Gautam Das, and Nick Koudas. "Sampling Online Social Networks." *IEEE Transactions on Knowledge and Data Engineering* 25.3 (2013): 662–676.

Paßmann, Johannes, Thomas Boeschoten, and Mirko Tobias Schäfer. "The Gift of the Gab. Retweet Cartels and Gift Economies on Twitter." *Twitter and Society*. Eds. Katrin Weller *et al*. New York: Peter Lang, 2013.

Poblete, Barbara, *et al*. "Do All Birds Tweet the Same? Characterizing Twitter around the World Categories and Subject Descriptors." *20th ACM Conference on Information and Knowledge Management, CIKM 2011, ACM, Glasgow, United Kingdom*. 2011. 1025–1030.

Rieder, Bernhard. "The Refraction Chamber: Twitter as Sphere and Network." *First Monday* 11 (5 Nov. 2012).

Rogers, Richard. *The End of the Virtual – Digital Methods.* Amsterdam: Amsterdam University Press, 2009.

Savage, Mike, and Roger Burrows. "The Coming Crisis of Empirical Sociology." *Sociology* 41.5 (2007): 885–899.

Stalder, Felix. "Between Democracy and Spectacle: The Front-End and Back-End of the Social Web." *The Social Media Reader*. Ed. Michael Mandiberg. New York: New York University Press, 2012. 242–256.

Stieglitz, Stefan, and Nina Krüger. "Analysis of Sentiments in Corporate Twitter Communication – A Case Study on an Issue of Toyota." *ACIS 2011 Proceedings*. (2011). Paper 29.

Tumasjan, A., *et al*. "Election Forecasts with Twitter: How 140 Characters Reflect the Political Landscape." *Social Science Computer Review* 29.4 (2010): 402–418.

Tukey, John Wilder. Exploratory Data Analysis. New York: Addison-Wesley, 1977.

Uprichard, Emma. "Sampling: Bridging Probability and Non-Probability Designs." *International Journal of Social Research Methodology* 16.1 (2011): 1–11.

Journal of Technology in Human Services
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/wths20

The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source

Fabio Giglietto [a] , Luca Rossi [a] & Davide Bennato [b]
[a] University of Urbino Carlo Bo, Urbino, Italy
[b] University of Catania, Catania, Italy

Published online: 06 Dec 2012.

To cite this article: Fabio Giglietto , Luca Rossi & Davide Bennato (2012): The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source, Journal of Technology in Human Services, 30:3-4, 145-159

To link to this article: http://dx.doi.org/10.1080/15228835.2012.743797

Journal of Technology in Human Services, 30:145–159, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1522-8835 print/1522-8991 online
DOI: 10.1080/15228835.2012.743797

The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source

FABIO GIGLIETTO and LUCA ROSSI
University of Urbino Carlo Bo, Urbino, Italy

DAVIDE BENNATO
University of Catania, Catania, Italy

A growing amount of content is published worldwide every day by millions of social media users. Most of this content is public, permanent, and searchable. At the same time, the number of studies proposing different techniques and methodologies to exploit this content as data for researchers in different disciplines is also growing. This article presents an up-to-date literature review that frames available studies using Facebook, Twitter, and YouTube as data sources, in the perspective of traditional approaches for social scientists: ethnographical, statistical, and computational. The aim is to offer an overview of strengths and weaknesses of different approaches in the context of the possibilities offered by the different platforms.

KEYWORDS computational social science, ethnography, literature review, methodology, social media

INTRODUCTION

With the expression "social media," we describe a varied category of Internet services inspired by Web 2.0 (O'Reilly, 2007) principles and enabling the users of the site to create and share digital contents (Kaplan & Haenlein,

Received May 15, 2012; revised September 15, 2012; accepted October 16, 2012.

Address correspondence to Fabio Giglietto, Department of Communication Studies, University of Urbino Carlo Bo, Via Saffi, 15–61029, Urbino (PU), Italy. E-mail: fabio.giglietto@uniurb.it

2010). Social network sites based on visible, interconnected, and navigable profiles (boyd & Ellison, 2008) belong to this category, as well as the sites focused on content sharing. Despite the large variety of possibilities—from the short 140 characters of Twitter to the video sharing of YouTube—it is clear that the focus of a social media is on both the users' activities and the users' relationships and social networks.

The social media success that we have been observing for a decade has undoubtedly boosted the online diffusion of already existing creative practices (Burgess, 2006) and has formed the background for a growing interest of academic research on these platforms and on the phenomena that they are able to host.

This new digital scenario challenges, in a very radical way, the standard research practices within the social sciences, bringing an unprecedented rate of innovation (the Internet changes at an extraordinarily fast pace) and an exceptional data availability (Karpf, 2012). To meet these challenges, there is a new wave of studies coming from very different backgrounds—from computer sciences to behavioral sciences.

While social media are undoubtedly a broad research field, it can be surely claimed that the contemporary scenario is composed of a small group of very big actors—in terms of users and daily usage—and a large number of minor services that are often addressed to specific communities. This article presents a structured, up-to-date literature review of the studies focusing on three of the largest and most famous social media sites: Facebook, YouTube, and Twitter.

The reasons we deliberately focused on these three platforms is mainly related to their huge popularity among both users and researchers. YouTube, Facebook, and Twitter have, as a matter of fact, been studied from many different perspectives by many different researchers coming from various disciplines. The result of this success as new field of research is a wide spectrum of literature—albeit often confusing—coming from different scientific backgrounds that often ignore their reciprocal existence. The aim of this article is therefore to attempt a first categorization of a large part of the existing literature according to a methodological perspective rooted in the sociological tradition.

METHODOLOGICAL FRAMEWORK

Using social media as a data source is a relatively new phenomenon. The deeply interdisciplinary nature of these studies makes it difficult to retrieve a complete and up-to-date literature of papers employing this approach. At the same time, we felt the need to frame collected papers within a more solid and well-known analytical framework. In fact, even if social media data comprise a new phenomenon, social data—with their own large heterogeneity—have been used by sociologists for a long time. Therefore, we rooted the classification of the collected papers within the traditional distinction of sociological methodological approaches that have been summarized by Ricolfi (1997) in three major groups or research methods: ethnographical approaches, statistical approaches, and computational approaches. These categories follow the traditional distinction between quantitative/statistical methods and qualitative/ethnographic methods with the addition of the new computational methods that offer specific characteristics in terms of quantity and the nature of the data. The computational approach is different from the statistical one because data are not organized in a matrix of variables and cases. Data are instead organized in a structure that recalls more a relational database than a spreadsheet. This is the reason why computational approaches do not necessarily need the use of statistics, even if univariate or bivariate data representations are useful to visualize some results.

It is interesting to point out that even if this triple distinction dates back to the sociological tradition, it has recently been used—in a simpler form—to summarize the sociological research at large according to a scale defined by the depth of the analysis (high for ethnography and low for statistical approaches) and by the replicability of the scientific observations through time (high for statistical approaches and low for ethnography; Aharony, 2011). Therefore, due to the flexibility of this schema, we opted to use it as a common framework to describe the most relevant studies for every platform.

PLATFORMS, USERS' EXPERIENCE, AND DATA ACCESS POLICIES

Since the three platforms under examination expose data in a very different way, thus offering different user experiences and possibilities to researchers, it is necessary, before digging into the literature review, to provide an introduction on the characteristics that differentiate YouTube, Twitter, and Facebook under this perspective. This is of utmost importance because platforms matter both on the side of the social practices that are able to host and on the side of the research opportunities that offer to the researchers. Speaking of social media research *sui generis* with no connection to real-world platform makes, nowadays, very little sense.

YouTube is the most important video-sharing platform with 800 million users monthly, 4 billion videos viewed daily, and 60 hours of video uploaded every minute (YouTube, 2012). These numbers make YouTube the third website in the world based on traffic (Alexa, 2012). The user experience consists mainly in viewing videos. The videos can be found using the internal platform search engine, subscribing to a specific channel, or following links shared in other social networks (Facebook, Twitter). Users' interaction can range from video production to passive video viewing to video commenting or sharing (Burgess & Green, 2009). Therefore, the final user experience is different if the user is a content producer or a passive viewer (Cormode, Krishnamurthy, & Willinger, 2010). However, on a general level, it is possible to distinguish three different forms of interaction: audience interactions, social interactions, and platform interactions. Audience interactions can be measured by using metrics such as exposures—how many times a video or a channel is viewed. Different metrics are used to measure social interactions: number and type of comments posted by registered users, *likes* received by the video, or channel subscriptions. Platform interactions are measured by different kinds of information that are possible to enter when a video is uploaded (metadata): title, date, ID, tags, uploading account, description, category, copyright license, and so on. These pieces of information are used in different ways. One way they are used is in selecting which videos to analyze (e.g., the number of visualizations for choosing the most viewed videos). Another way is as contextual information for building social behavior patterns (e.g., strategies used in tagging). For these reasons, starting from the same data, every study that is using YouTube as its main data source can plan a brand new research strategy combining different metrics.

While YouTube was pushing the web into the online video era, Twitter introduced a text-only service that faced, since its early days, a worldwide success: microblogging. Born in 2006, Twitter reached the number of 340 million Tweets per day in 2012 (Twitter, 2012) and is now ranked as the eighth most visited website worldwide (Alexa, 2012). This result, in terms of number of users, rapidly produced a large volume of scientific research. This indisputable success of the Twitter platform among the research community, as well as its diffusion among third-party developers, can largely be explained by its data availability and structure. These data have always been freely available, public by default, mainly textual, and easily understandable. Additionally, free and public Application Programming Interface (API)-based access to the data has been available to developers and to skilled researchers since the launch of the platform, and, more recently, commercial services started selling specific portion or subset of Twitter data. Information available through Twitter API is of a very simple nature: It can be about the tweets or it can be about the users. Information about the tweets are the textual content of the tweet itself, time and location of its production, and the relational nature of the messages (whether there are replies to other messages or retweets of a previously produced message). Besides message-related information, user specific information is available: user name, user self-declared[1] location, the list of the users followed by the user, and the list of the users following the user. Despite its apparent simplicity, these sets of information can be combined in order to provide useful data on many aspects of Twitter usage—from posting topics and strategies to the establishment and the evolution of Twitter communities.

This data availability produced a large amount of research that can be sorted using the methodological schema we are adopting in this paper: ethnographic, statistical and computational. It is interesting to point out that these research approaches were not concurrent but followed a clear pattern representing the various scientific backgrounds currently involved in Twitter research. This led to an order of research approaches that is most likely different from those we are describing for the other platforms. While in many research fields related to online social activities, ethnographic approaches have predated statistical and, especially, computational approaches; in the "Twitter field," computational approaches appeared first, as can be observed by the venues or by the authors of the very first academic research on the topic (Huberman, Romero, & Wu, 2008; Java, Song, Finin, & Tseng, 2007).

With more than 900 million monthly active users worldwide and more than 500 million daily active users at the end of March 2012 (Facebook, 2012), Facebook is the most popular social medium in the world. One of the reasons for this worldwide success is the sense of protection, mainly developed during the early stages when the platform was open to selected colleges only (Joinson, 2008), experienced by users sharing their content with their bound community of "Friends" (boyd & Ellison, 2008). For this reason, Facebook managers devoted a growing amount of attention to developing a set of privacy settings and, more recently, to making these settings more usable (boyd & Hargittai, 2010). Despite that, recent research on Facebook privacy settings discovered that there is still a large gap between users' expected level of privacy and actual levels of access to their contents (Liu et al., 2011). Discussing the social implications or effectiveness of the Facebook strategy is beyond the scope of this article. However, the complexity of Facebook privacy settings deeply affects the extent and the type of data actually accessible to researchers. While most of the information is private by default in personal profiles, on pages—special profiles intended for organizations, public figures, and brands—information is publicly available. More recently, this simple distinction was blurred by new platform developments that allow the user to choose, on a post-to-post basis, the intended audience (some or all Friends or Public).

ETHNOGRAPHICAL APPROACHES

On a wide perspective, the ethnographical approach focuses on social meanings inferred by the researcher from the content intended as a unit of analysis. As an instance, YouTube videos usually define particular communities (e.g., the v-loggers in Griffith & Papacharissi, 2010) producing specific contents for the members of a community.

Studies following the ethnographical approach share common recognizable features. The group of units analyzed is usually small, with employment of different qualitative techniques, and usually a research design based on multiple methods or triangulation techniques.

There are different researchers—though not a great number of them—using ethnographic approaches, showing the potential of media ethnography when applied to YouTube (Lange, 2007; Rotman, Golbeck, & Preece, 2009). One of them is focused on a single video from the account Geriatric1927 whose title is "Teenagers and Drugs" (Harley & Fitzpatrick, 2009). By using a specific interpretive framework (multimodal interactional analysis: Norris, 2004) and different research techniques (conversation analysis audio transcription, transcription of nonlinguistic aspects of dialogues), applied also to the video responses it generates, the authors suggest that YouTube is mainly a social broadcast medium.

More recently than the attempts to describe the topological aspects of Twitter network, many researchers started to focus on the communicative practices of the platform, aiming at describing what kinds of social interactions were made possible by the exchange of simple 140-character-long text messages. These studies have been based both on quantitative-computational approaches (Huberman, Romero, & Wu, 2008) and on qualitative and often ethnographic approaches (Marwick & boyd, 2010). They have also provided precious insights, not only about the distribution of contacts and tweets but also on the tweeting and retweeting strategies of users, as well as their perception of their "invisible" audiences (boyd, 2008). Twitter research, even when it was mainly designed as an ethnographic investigation, has often been based on a computer-assisted data collection supported also through many freely available online tools that allowed one to retrieve and download Twitter messages.

The attempts of employing ethnographical approaches to data retrieved from Facebook tend to focus on small samples of "Friends" or public contents. The already-mentioned distinction between profiles and pages is important to understanding why this approach is not common among researchers. On the one hand, studies based on "Friends" profiles (boyd, 2008) tend to be biased and to produce results difficult to generalize. On the other hand, dealing with public contents (pages and groups) quickly increases the amount of data to analyze therefore discourages approaches based on in-depth observations. Nevertheless, the range of possibilities opened up by the analysis of content shared by users and organizations is as varied as the kinds of studies enabled. For instance, it is possible to analyze posted links pointing to online news articles in order to understand how a grass-roots agenda of topics can develop among a community of friends (Baresch, Knight, Harp & Yaschur, 2011).

STATISTICAL APPROACHES

Within the statistical approach, the reference model is the variable by cases data matrix, in which a single content (or a user) is the unit of analysis, and it is part of a sample extracted from the population of contents (or users).

Speaking about YouTube, the video is considered a trace of a social behavior, a way for accessing meanings of a community. For this reason, the video is not important per se, but in relation to the information it provides on the community it belongs to. Features of these studies are the use of sampling techniques and the use of content analysis (coded by humans or sometimes automatically by the computer). There are numbers of researchers using this approach, for example, the study by Bal, Campbell, Payne, and Pitt (2010) on mapping conversations around political spoof videos. In this study, researchers used a text analysis software (Leximancer) applied to the comments posted on three different political spoof videos, defining topics and words used by the commenting viewers in order to measure the sentiment of the audience.

Many researchers on Twitter used what can be defined as a *computational supported statistical approach* to sample the users' messages by acquiring—through the use of the public API—a Twitter-provided sample of the messages published on the public timeline (Honeycutt & Herring, 2009). While researchers moved into the Twitter world, the wide diffusion of Twitter as a social platform led to the emergence of unexpected social phenomena that found on this social network site a perfect sociotechnical environment able to host them. In just a few years, Twitter became a digital space where public issues could be discussed, critical information could be shared during natural disasters, and TV shows could be commented on by and with their fans. This user-led evolution of the platform produced a wave of studies focused on these phenomena with a closer perspective (Bruns, 2011; Rossi, Magnani, & Iadarola, 2011; Wohn & Na, 2011). These studies constitute the starting point for a more comprehensive understanding of the Twitter dynamics and of the wide range of social interactions that emerged from the platform, providing as well some interesting categorizations of these phenomena and of their Twitter-based characteristics. Thanks to the growing number of ad hoc studies and case studies, we now know that *crisis events* have communication patterns quite different from *media events* and that different actors are involved (Bruns, 2011). The study of Twitter usage during natural disasters has been particularly interesting and dates back to the early work of Earle (2010), which describes how the Twitter-based report of an earthquake experience provides a powerful tool to supplement instrument-based techniques in a quake's location and magnitude evaluation. Another important part of Twitter-based research—usually carried out with mixed *computational supported ethnographic or statistical methods*—is related to its political use. Following a series of political elections in 2009 and 2011, studies focused on describing how the politics were changed by the conversations taking place on the microblogging service (Bruns & Burgess, 2011) or whether Twitter itself could be used as a viable predictive tool to forecast the electoral outcomes (Jungherr, Jurgens, & Schoen, 2011; Lassen & Brown, 2010; Tumasjan, Sprenger, Sandner, & Welpe, 2010). These new approaches took a step forward toward the study of Twitter as a widely adopted social media and part of a larger media ecology. This is well beyond the studies aiming at describing Twitter as a network, and it places Twitter as a normal part of the contemporary media scene. When Lassen and Brown (2010) tried to use Twitter to predict electoral results, the basic underlying assumption was that Twitter was diffused enough within the politically active part of the society that a large-scale analysis of Twitter political communication could have represented a sample of the whole society. This should make clear the shift from a research approach focused on the whole Twitter network to an ad hoc social phenomenon-driven approach.

As opposed to Twitter, a stream analysis of public contents shared on Facebook is almost worthless since the great majority of users tend to share contents with their "Friends" only.

Most of the distinctions between profiles and pages also apply when dealing with statistical approaches to contents (posts, photos, videos, links, and activities) shared on Facebook by users or organizations. Retrieving and analyzing pages and group contents enables studies focused on specific organizations such as nonprofits (Waters, Burnett, Lamm, & Lucas, 2009) or political groups (Woolley, Limperos & Oliver, 2010). These studies often employ a statistical approach based on content analysis and are a reasonable follow-up of studies based on the analysis of pages and groups meta-information.

Profile meta-information is all the data available in the profiles' "about" and "favorites" sections. These sections contain information such as the user's list of "Friends," birthday, relationship status, family relationships, work, and education, as well as liked pages, music, books, movies, and so on. While in personal profiles these data are shared, by default, to "Friends" only, on pages and graph-enabled websites (e.g., webpages exposing the "like" button), this information (e.g., the number of users who liked a page or page description) is public. This crucial dissimilarity made Facebook pages and graph-enabled websites a viable target for studies on news dissemination (Lifshits & Clara, 2010) and brand engagement performances or popularity (Caren & Gaby, 2011; Lovari & Giglietto, 2012). Accessing most of the meta-information on personal profiles requires one to be a "Friend" of the subject. Once accepted as a "Friend," it is possible to study disclosure strategies (Kolek & Saunders, 2008) and user preferences and ego networks (Hogan, 2008).

COMPUTATIONAL APPROACHES

The computational approach is typical of the computer sciences, but when it is applied to social media, it provides interesting information for social scientists (Manovich, 2008). Within this approach, any software object expresses something about the users of the platform or the platform itself, and also helps to understand some properties that otherwise would not be directly observable. Studies based on computational approaches share common features: big or enormous data collection, a web services-based approach (e.g., Tubekit, Tubemogul: Shah, 2010), an Application Programming Interface (API) manipulation approach, the attempt to model the results (e.g., according to a power law model, or by the analysis of the graph structure). The studies based on a computational approach could be further classified: Those employing a web services approach often belong to the community of social scientists, due to the simpler use of these platforms in data collection and analysis, whereas those studies using the API manipulation approach belong to the computer scientist community because of the skills of computer programming that are needed for interacting with the platform through the API. The API manipulation is important because in this way it is possible to collect a great amount of data and to build a database with much information and metadata about the focus of the research. Although a growing number of social scientists are now using this approach thanks to the growing success of the so-called digital methods (Rogers, 2009), so far computer scientists have provided the major contributions from a pure computational perspective. An example of this approach is the study of Wallsten (2010) on the "Yes we can" viral video produced by Will.I.Am of the hip hop group The Black Eyed Peas. This study argues that bloggers and other Obama campaign supporters played a crucial role in convincing people to watch the video and attract the interest of the media mainstream. This research used Tubemogul web service to collect the data of total exposures and to create a viral model of the video. Another study on the popularity of YouTube videos (Chatzopoulou, Sheng, & Faloutsos, 2010) used YouTube's API to build a crawler for collecting different data in a database of 37 million videos to create a model of popularity of most viewed videos. An interesting result of this research is the "magic number" 400: A video receives one comment, one rating, and is added to someone's favorite list once for every 400 times it is viewed.

As previously stated, the disposal of public API makes Twitter data available to a large and growing number of researchers with some basic programming skills. In fact, many of the first studies on the topic are characterized by a computer science background and focused on the analysis of the network structure and on its topological characteristics studying the Twitter network as a whole. Within this perspective in their opening work, Java, Song, Finin, and Tseng (2007)—starting from a sample of 76,177 users—described Twitter's network structure, geographical distribution, and interaction between users. These studies were made possible by the relatively small size of the Twitter network at that time and have subsequently become less frequent, mainly due to both the growing size of the Twitter network—now 140 million daily users (Twitter, 2012)—and the consequent limits to data acquisition. Although we can still find recent studies regarding the global structure of the Twitter network (Wu, Hofman, Watts, & Mason, 2010), it has become apparent that these studies should be integrated with specific analyses of local phenomena that are not always visible at a *whole network* level of observation but still constitute the essence of the Twitter communication experience from the point of view of the users.

For the very same reason, the studies on Facebook whole-network structures are rare and always based on a data set provided by Facebook itself. Analyzing a large networks structure is often computationally challenging, and the results of these studies provide an abstract overview of users' behavior both in bonded communities (Lewis, Kaufman, Gonzales, Christakis, & Tastes, 2008) or entire countries (Traud & Mucha, 2011). Also, concerning stream analysis of public contents, the most interesting studies are made in partnership with Facebook itself (see the recent agreement between the Politico website and Facebook, which focused on the sentiment analysis of posts mentioning the candidates of the 2012 U.S. Republican primary elections). Although stream-based studies could be carried on with traditional manual-coded content analysis, the amount of data collected suggests instead the use of computational techniques.

Gaining access to meta-information for non-"Friends" requires the development of ad hoc Facebook applications such as the one described by Rauber and Almeida in their essay on users' privacy awareness (2011). An explicit informed consent of the subject is enforced by the platform and required by the standard norm of research ethics.

DISCUSSION AND CONCLUSIONS

Due to the complexity and heterogeneity of the described approaches, it is often challenging to classify a study within one of the categories of the methodological frames we have adopted. A classical distinction between ethnographic and statistical approaches, even if complemented by computational methods, is hardly able to describe all the ways in which social media data can be used to understand online users' behaviors.

The digital nature of the data, along with the amount of information available, makes the computational approach the most suitable way to collect it. Nevertheless, once data have been gathered, they can be analyzed with either a quantitative or a qualitative approach, depending on the research questions and strategies. Content analysis based on manual coding of "big data" is particularly hard because the amount of data makes the process extremely time-consuming. At the same time, sampling this data is also challenging because in most of the cases the distributions are extremely skewed (e.g., few extremely active users and a long tail of far less productive users). Under this perspective, although not yet perfect, the advances in the field of automatic semantic analysis appear to be promising, especially when used as a screening technique aimed to support manual coding.

Within this scenario, mixed methods approaches are often the most promising but the least frequently used. Although the amount of work required by these approaches may be an issue, a more pressing issue is the necessary collaboration among scholars coming from different backgrounds required by mixed methods approaches.

On the one hand, social scientists embracing social media as a data source for their studies need to gain a general understanding of the platforms from both the technical and the cultural point of view. This basic knowledge about the platform could be acquired by being part of the communities hosted by the platforms. This very first step is required not only for ethnographical approaches, where knowledge about the setting is somehow mandatory, but also for statistical and computational ones.

On the other hand, a researcher's background still matters when it comes to data interpretation. Even when research areas are similar, the perspective of analysis can be very different. When computer scientists look for online communities in social media, they usually adopt a graph mining approach to detect communities, which is, de facto, a direct evolution of graph clustering techniques (Leskovec et al., 2008) but is this a proper formalization of the sociological concept of community? Alternatively, many of the recent studies on online communities (Baym, 2007) coming from social scientist scholars still lack the level of formalization necessary to make the concepts suitable to be used in computational approaches to "big data" (boyd & Crawford, 2012; Wellman et al., 2002).

The literature review proposed in this article clearly points out a need to develop studies carefully designed to take advantage of a mixed methods approach including ethnographic, statistical and computational methods. However, the methodological skills required often exceed the traditional curriculum of social scientists. There is therefore a strong need for collaboration among scientists coming from different backgrounds in order to support studies that combine broad perspectives with in-depth and effective interpretations.

NOTE

[1] The location field can be used both to communicate the user's real geographical coordinates using global positioning system (GPS)-enabled devices and to state more generic information (like the name of the country) or even to convey sarcastic or political messages, as noted by some recent studies (Takhteyev, Gruzd, & Wellman, 2011).

REFERENCES

Aharony, N. (2011). *Social fMRI: Measuring, understanding, and designing social mechanisms in the real world*. Cambridge, MA: Massachusetts Institute of Technology.

Alexa. (2012). *YouTube site info*. Retrieved from http://www.alexa.com/siteinfo/youtube.com

Bal, A. S., Campbell, C. L., Payne, N. J., & Pitt, L. (2010). Political ad portraits: a visual analysis of viewer reaction to online political spoof advertisements. *Journal of Public Affairs*, *10*(4), 313–328. Wiley Online Library.

Baresch, B., Knight, L., Harp, D., & Yaschur, C. (2011). Friends who choose your news: An analysis of content links on Facebook. *ISOJ: The Official Research Journal of International Symposium on Online Journalism*, *1*. Retrieved from http://online.journalism.utexas.edu/2011/papers/Baresch2011.pdf

Baym, N. K. (2007). The new shape of online community: The example of Swedish independent music fandom. *First Monday*, *12*(8). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1978/1853

boyd, d. (2008). *Taken out of context: American teen sociality in networked publics*. University of California, Berkeley. Retrieved from http://www.danah.org/papers/TakenOutofContext.pdf

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society*, *15*(5), 662–679.

boyd, d., & Ellison, N. B. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. doi:10.1111/j.1083-6101.2007.00393.x

boyd, d., & Hargittai, E. (2010). Facebook privacy settings: Who cares? *First Monday*, *15*(8). Retrieved from http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3086

Bruns, A. (2011). How long is a tweet? Mapping dynamic conversation networks on Twitter Using Gawk and Gephi. *Information, Communication & Society*, *15*(9), 1–29. doi:10.1080/1369118X.2011.635214

Bruns, A., & Burgess, J. (2011). #Ausvotes: How twitter covered the 2010 Australian federal election. *Communication, Politics & Culture*, *44*(2), 37–56.

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, *17*(4-2). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3937/3193

Burgess, J. (2006). Hearing ordinary voices: Cultural studies, vernacular creativity and digital storytelling. *Continuum: Journal of Media & Cultural Studies*, *20*(2), 201–214.

Burgess, J., & Green, J. (2009). *YouTube: Online video and participatory culture*. Cambridge, England: Polity Press.

Caren, N., & Gaby, S. (2011). Occupy online: Facebook and the spread of Occupy Wall Street. *SSRN Electronic Journal*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1943168. doi:10.2139/ssrn.1943168

Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010, March). A first step towards understanding popularity in YouTube. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010* (pp. 1–6). IEEE.

Cormode, G., Krishnamurthy, B., & Willinger, W. (2010). A manifesto for modeling and measurement in social media. *First Monday*, *15*(9-6). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3072

Earle, P. (2010). Earthquake Twitter. *Nature Geoscience*, *3*(4), 221–222. doi:10.1038/ngeo832

Facebook. (2012). *Key facts*. Retrieved from http://newsroom.fb.com/content/default.aspx?NewsAreaId=22

Griffith, M., & Papacharissi, Z. (2009). Looking for you: An analysis of video blogs. *First Monday*, *15*(1). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2769/2430

Harley, D., & Fitzpatrick, G. (2009). Creating a conversational context through video blogging: A case study of Geriatric1927. *Computers in Human Behavior*, *25*(3), 679–689.

Hogan, B. (2008). Analyzing social networks via the Internet. In *The Sage handbook of online research methods* (pp. 141–160). London: Sage.

Honeycutt, C., & Herring, S. C. (2009). *Beyond microblogging: Conversation and collaboration via Twitter* (Vol. 0, pp. 1–10). Los Alamitos, CA, USA: IEEE Computer Society. doi:http://doi.ieeecomputersociety.org/10.1109/HICSS.2009.602

Huberman, B. A., Romero, D. M., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, *14*(1). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we Twitter: Understanding microblogging. *Network*, *1*, 56–65. doi:10.1145/1348549.1348556

Joinson, A. N. (2008). Looking at, looking up or keeping up with people? *Proceeding of the twenty-sixth annual CHI conference on human factors in computing systems—CHI '08* (p. 1027). New York, NY: ACM Press. doi:10.1145/1357054.1357213.

Jungherr, A., Jurgens, P., & Schoen, H. (2011). Why the Pirate Party won the German election of 2009, or The trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M., Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment. *Social Science Computer Review*, *1*(6). doi:10.1177/0894439311404119

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. Elsevier.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society*, *15*(5), 639–661.

Kolek, E. A., & Saunders, D. (2008). Online disclosure: An empirical examination of undergraduate Facebook profiles. *Journal of Student Affairs Research and Practice*, *45*(1), 1–25. doi:10.2202/1949-6605.1905

Lange, P. G. (2007). Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication*, *13*, 361–380. doi:10.1111/j.1083-6101.2007.00400.x

Lassen, D. S., & Brown, A. R. (2010). Twitter: The electoral connection? *Social Science Computer Review*, *29*(4), 419–436. doi:10.1177/0894439310382749

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008, April). Statistical properties of community structure in large social and information networks. *Proceeding of the 17th International Conference on World Wide Web* (pp. 695–704). New York, NY: ACM.

Lewis, K., Kaufman, J., Gonzales, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*(4), 330–342. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378873308000385

Lifshits, Y. (2010). *Ediscope: Social analytics for online news*. Yahoo—Yahoo Labs. Retrieved from http://www.research.yahoo.net/files/YL-2010-008.pdf

Liu, Y., Gummadi, K. P., Krishnamurthy, B., & Mislove, A. (2011, November). Analyzing Facebook privacy settings: User expectations vs. reality. *Proceedings of the 2011*

ACM SIGCOMM Conference on Internet Measurement Conference (pp. 61–70). New York, NY: ACM.

Lovari, A., & Giglietto, F. (2012). Social media and Italian universities: An empirical study on the adoption and use of Facebook, Twitter and Youtube. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/so13/papers.cfm?abstract_id=1978393. doi:10.2139/ssrn.1978393

Manovich, L. (2008). *Software takes command*. Unpublished. Retrieved from http://lab.softwarestudies.com/2008/11/softbook.html

Marwick, A. E., & boyd, d. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. doi:10.1177/1461444810365313

Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. London, UK: Psychology Press.

O'Reilly, T. (2007). *What is Web 2.0. Design patterns and business models for the next generation of software*. Retrieved from http://oreilly.com/lpt/a/6228

Rauber, G., & Almeida, V. A. F. (2011). Privacy albeit late. *Networks*, *13*, 26. Retrieved from http://precog.iiitd.edu.in/Publications_files/GR_VA_PK_SMW_2011.pdf

Ricolfi, L. (2001). *La ricerca qualitativa*. Rome, Italy: Carocci.

Rogers, R. (2009). *The end of the virtual: Digital methods*. Amsterdam, The Netherlands: Amsterdam University Press.

Rossi, L., Magnani, M., & Iadarola, B. (2011). #rescatemineros: Global media events in the microblogging age. *Selected papers of Internet research*, 0(12.0). Retrieved from http://spir.aoir.org/index.php/spir/article/view/30

Rotman, D., Golbeck, J., & Preece, J. (2009). The community is where the rapport is—On sense and structure in the youtube community. *Proceedings of the Fourth International Conference on Communities and Technologies* (pp. 41–50). New York, NY: ACM.

Shah, C. (2010). Supporting research data collection from YouTube with TubeKit. *Journal of Information Technology & Politics*, *7*(2–3), 226–240.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2011). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81.

Traud, A., & Mucha, P. (2011). Social structure of Facebook networks. *Physica A*, *391*, 4165–4180. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437111009186

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*(4), 402–418. doi:10.1177/0894439310386557

Twitter, Inc. (2012). Twitter turns six. *Twitter Blog*. Retrieved from http://blog.twitter.com/2012/03/twitter-turns-six.html

Wallsten, K. (2010). "Yes we can": How online viewership, blog discussion, campaign statements, and mainstream media coverage produced a viral video phenomenon. *Journal of Information Technology & Politics*, *7*(2–3), 163–181.

Waters, R. D., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review*, *35*(2), 102–106. doi:10.1016/j.pubrev.2009.01.006.

Wellman, B., Boase, J., & Chen, W. (2012). The networked nature of community: Online and offline. *IT & Society*, *1*(1), 151–165.

Wohn, D. Y., & Na, E. K. (2011). Tweeting about TV: Sharing television viewing experiences via social media message streams. *First Monday*, *3*(16). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3368/2779

Woolley, J. K., Limperos, A. M., & Oliver, M. B. (2010). The 2008 Presidential election, 2.0: A content analysis of user-generated political Facebook groups. *Mass Communication and Society*, *13*(5), 631–652. doi:10.1080/15205436.2010.516864.

Wu, S., Hofman, J. M., Watts, D. J., & Mason, W. A. (2010, March). Who says what to whom on Twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 705–714). New York, NY: ACM.

YouTube. (2012). *Statistic: Traffic*. Retrieved from http://www.youtube.com/t/press_statistics

ABOUT THE AUTHORS

Fabio Giglietto, PhD, is assistant professor at the Department of Communication and Human Studies of the University of Urbino “Carlo Bo”. His main research interests are theory of information, communication and social systems with specific focus on the relationship between social systems and new technologies. He is currently working on developing metrics to understand the relationship between social media practices and ‘real world’ phenomena. Since 2010 he is a board member of RC51 on Sociocybernetics, a research committee of the International Sociological Association and member of editorial board of the *Journal of Sociocybernetics*. Full and up to date list of publications and CV is available at http://www.mendeley.com/profiles/fabio-giglietto.

Luca Rossi is a fellowship researcher at the Department of Communication Studies of the University of Urbino “Carlo Bo”. His research interests are focused on Internet studies particularly in the fields of Internet culture, games studies and Social Networks studies. He is currently working on Social Network Analysis methods applied to online propagation of Information and cultural phenomena. Among his most recent publications: “Toward a Bridge between Sociocybernetics and Internet Studies,” *Journal of Sociocybernetics*, 2009, “Media & Generation: How User Generated Content Reshape Generational Identity in the Mass Media System,” *Sociologia della Comunicazione*, 2010, and Information Propagation in Social Network Sites in *Advances in Social Networks Analysis and Mining*, *2010 International Conference on IEEE.*

Davide Bennato is assistant professor of Sociology of digital media at the University of Catania and his research interest are focused on technological cultures, digital media consumption and digital media socialization. He is author of *Sociologia dei Media Digitali* (Sociology of Digital Media [2011]) and co-author of *Dizionario di informatica, dell’ICT e dei media digitali* (2012). He writes about the social impact of the internet in his blog Tecnoetica (http://www.tecnoetica.it/).

The Hermeneutics of Screwing Around; or What You Do with a Million Books

Stephen Ramsay

April 17, 2010

According to the World Wide Web, the phrase, "So many books, so little time" originates with Frank Zappa. I don't believe it, myself. If I had had to guess, I would have said maybe Erasmus or Trithemius. But even if I'm right, I'm probably wrong. This is one of civilization's oldest laments—one that (in spirit, at least) predates the book itself. There has never been a time when philosophers—lovers of wisdom broadly understood—have not exhibited profound regret over the impedance mismatch between time and truth. For surely, there are more books, more ideas, more experiences, more relationships worth having than there are hours in a day (or days in a lifetime).

What everyone wants—what everyone from Sargon to Zappa has wanted—is some coherent, authoritative path through what is known. That's the idea behind Dr. Elliot's Five Foot Shelf, Adler's *Great Books of the Western World,* Modern Library's *100 Best Books,* and all other similar attempts to condense knowledge into some ordered list of things the educated should know. It's also the idea behind every syllabus, every curriculum, and most of the non-fiction books that have ever been written. The world is vast. Art is long. What else can we do but *survey* the field, *introduce* a topic, plant a *seed* (with, what else, a *seminar).* Amazon.com has a feature that allows users to create reading guides focused on a particular topic. They call it, appropriately, "Listmania."

While the anxiety of not knowing the path is constant, moments of cultural modernity provide especially fertile ground for the creation of epitomes, summae, canons, and bibles (as well as new schools, new curricula, and new ways of organizing knowledge). It is, after all, at the end of history that one undertakes summation of "the best that has been thought and said in the world" (190). The aforementioned "great books" lists all belong to the early decades of the twentieth century, when U.S. cultural anxiety—especially concerning its relationship to Europe—could be leavened with a bold act of cultural confidence. Thomas Jefferson had said something similar at a time closer to the founding of the country, when he noted that "All that is necessary for a student is access to a library, and directions in what order the books are to be read." But the same phenomenon—the same play of anxiety and confidence—was at work in the writing of the Torah, the *Summa*, Will Durant's *Story of Civilization,* and all efforts of similar grandeur. All three of those works were written during moments, not just of rapid cultural change, but during periods of anxiety about change. "Hear, O Israel, the statutes and judgments which I speak in your ears this day, that ye may learn them, and keep, and do them" (Deutronomy 5:1); "[W]e purpose in this book to treat of whatever belongs to the Christian religion, in such a way as may tend to the instruction of beginners" (1); "I wish to tell as much as I can, in as little space as I can, of the contributions that genius and labor have made to the cultural heritage of mankind" (?) This essay will not aim quite so high.

Even in the very early days of the Web, one felt the soul-crushing lack of order. One of the first pages I ever visited was "David and Jerry's Guide to the World Wide Web," which endeavored to, what else, *guide* you through what seemed an already impossibly vast expanse of information (you may have heard of that particular compendium; it's now called Yahoo!). Google might seem something else entirely, but it shares the basic premise of those quaint guides of yore, and of all guides to knowledge. The point is not to return the over three million pages that relate in some way to Frank Zappa. The point is to say, "Relax. Here is where you start. Look at this. Then look at that."

We might say that all such systems rely on an act of faith, but it's not so much trust in the search engine (or the book, or the professor) as it is willingness to suspend disbelief about the yellow wood after having taken a particular road. Literary historian Franco Moretti states the situation starkly:

> [W]e've just started rediscovering what Margaret Cohen calls the "great unread." "I work on West European narrative, etc...." Not really, I work on its canonical fraction, which is not even one per cent of published literature. And again, some people

> have read more, but the point is that there are thirty thousand nineteenth-century British novels out there, forty, fifty, sixty thousand—no one really knows, no one has read them, no one ever will. And then there are French novels, Chinese, Argentinian, American . . . (55)

Debates about "canonicity" have been raging in my field for as long as the field has been around. Who's in? Who's out? How do we decide? Moretti reminds us of the dispiriting fact that this problem has no practical solution. It's not just that someone or something will be left off; it's that our most inclusive, most enlightened choices will fail against even the most generous requirements for statistical significance. The syllabus represents the merest fraction of the professor's knowledge, and the professor's knowledge is embarrassingly slight. It's not that the emperor has no clothes (that would be fine); it's that no one knows what the emperor looks like.

Greg Crane, who held a series of symposia on the general question, "What Do You Do With A Million Books?" a few years ago, rightly identifies it as an ancient calculus:

> The Greek historian Herodotus has the Athenian sage Solon estimate the lifetime of a human being at c. 26,250 days (Herodotus, *The Histories,* 1.32). If we could read a book on each of those days, it would take almost forty lifetimes to work through every volume in a single million book library. The continuous tradition of written European literature that began with the Iliad and Odyssey in the eighth century BCE is itself little more than a million days old. While libraries that contain more than one million items are not unusual, print libraries never possessed a million books of use to any one reader.

Way too many books, *way* too little time.

But again, the real anxiety is not that the Library of Congress contains over 500 human lifetimes worth of reading material (I'm using the highly generous Solon-Crane metric, which assumes you read a book every day from the day you're born until the day you die). The problem is that that much information *probably* exceeds our ability create reliable guides to it. It's one thing to worry that your canon isn't sufficiently inclusive, or broad, or representative. It's another thing when your canon has no better chance of being these things than a random selection. When we get up into the fourteen-million-book range, books that are known by more than two living people are already “popular.” A book like *Hamlet* has overcome enormous mathematical odds that ruthlessly favor obscurity; the fact that millions of people have read it might become a compelling argument for why you should read it too. But in the end, arguments from the standpoint of popularity satisfy neither the canoniclast nor the historian. The dark fear is that no one can really say what is “representative,” because no one has any basis for making such a claim.

Several solutions have been proposed, including proud ownership of our ignorance and dilletantism. A few years ago, Pierre Bayard famously—and with only the barest sheen of satire—exposed our condition by writing a book entitled, “How To Talk About Books You Haven't Read?” In it, intellectual facility is presented as a kind of trick. “For knowing how to speak with finesse about something with which we are unacquainted has value far beyond the realm of books” (184). It is a lesson thoroughly absorbed by anyone who stands on the right side of a Ph.D. oral exam. But amazingly, even Bayard sees this as a means toward *guiding* people through knowledge.

> [Students] see culture as a huge wall, as a terrifying specter of “knowledge.” But we intellectuals, who are avid readers, know there are many ways of reading a book. You can skim it, you can start and not finish it, you can look at the index. You learn to live with a book. [...] I want to help people organize their own paths through culture. (“Read It?”)

At some level, there is no difference at all between Pierre Bayard and, say, Mortimer Adler. Both believe in culture. Both believe that one can find an ordered path through culture. Bayard just thinks there are faster ways to do it than starting with Volume 1 of *Great Books of the Western World.* Indeed, Adler himself almost seems to agree; books two and three of *Great Books* present what he calls a “Synopticon.” What could such a thing be but the *Cliff's Notes* to the main ideas of Western civilization?

There also isn't much of a difference between Bayard on the one hand and Crane and Moretti on the other. All three would like us to dispense with the silly notion that we can read everything, so that we can get on with the task of organizing our own paths through culture. It is true that the latter—as well as Digital Humanists generally—propose that we use computers, but I would like to argue that that difference is not as crucial as it seems.

There have always been two ways to deal with a library. The first is the one we're most used to thinking about. I am doing research on the influence of French composer Edgard Varèse on the early work of Frank Zappa. I go to the library and conduct an investigation, that might include the card catalog, a bibliography or two, the good people at the reference desk, or any one of a dozen different methods and tools. This is search. I know what I'm looking for, and I have various strategies for locating it. I can't read everyting on this subject. I can't even locate everything on this subject. But I have faith in the idea that I can walk out of the library (this afternoon, or after ten years of focused research, depending on my situation) being able to speak intelligently and convincingly on this topic.

The second way goes like this: I walk into the library and wander around in a state of insoucient boredom. I like music, so I head over to the music section. I pick up a book on American rock music and start flipping through it (because it's purple and big). There's an interesting bit on Frank Zappa, and it mentions that Zappa was way into this guy named Edgard Varèse. I have no idea who that is, so I start looking around for some Varèse. One look at the cover of his biography—Varèse with that mad-scientist look and the crazy hair—and I'm already a fan. And so off I go. I check out some records and discover Varèse.

This is called browsing, and it's a completely different activity. Here, I don't know what I'm looking for, really. I just have a bundle of "interests" and proclivities. I'm not really trying to find "a path through culture." I'm really just screwing around. This is more or less how Zappa discovered Varèse. He had read an article in *LOOK* magazine in which the owner of the *Sam Goody* record chain was bragging about his ability to sell obscure records like *The Complete Works of Edgard Varèse, Vol. 1* (Occhiogrosso 31). The article described Varèse's music as, "a weird jumble of drums and other unpleasant sounds." The rest is history (of the sort that you can search for, if you're so inclined).

We think of the computer as a device that has revolutionized search—"information retrieval," to use the formal term—and that is of course true. Until recently, no one was able to search the content of all the books in the library. There was no way to ask, "Which of these books contains the phrase 'Frank Zappa?' " The fact that we can now do that changes everything, but it doesn't change the nature of the thing. When we ask that question—or any question, for that matter—we are still searching. We are still asking a question and availing ourselves of various technologies in the pursuit of the answer.

Browsing, though, is a different matter. Because once you have programmatic access to the content of the library, screwing around suddenly becomes a far more illuminating and useful activity. That is, after all, why we called the navigational framework one used to poke around the World Wide Web a "browser." From the very start, the Web outstripped our ability to say what is actually there. Dave and Jerry couldn't do it then and Google can't do it even now. "Can I help you?" "No, I'm just browsing." Translation: "I just got here! How can you help me find what I'm looking for when (a) I don't know what's here and (b) I don't what I'm looking for?" The sales clerk, of course, doesn't need a translation. He or she understands perfectly that you're just screwing around.

And that is absolutely not what the people who are thinking about the brave new world of large-scale digital corpora (Google Books, or the Web itself) want to talk about. Consider Martin Mueller's notion of "not reading"—an idea he puts forth during a consideration of the power of the digital surrogate:

> A book sits in a network of transactions that involve a reader, his interlocutors, and a "collective library" of things one knows or is supposed to know. Felicitous reading—I adapt the term from John Austin's definition of felicitous speech acts—is the art of locating with sufficient precision the place a given book occupies in that network at a given moment. Your skill as a reader, then, is measured by the speed and accuracy with which you can do that. Ideally you should do it in "no time at all." Once you have oriented a book in the right place of its network, you can stop reading. In fact, you should stop reading. (Mueller 9–10).

Perhaps this isn't "search," classically understood, but it's about as far from screwing around as the average game theory symposium is from poker night. You go to the archive to set things right—to increase the likelihood that your network of associations corresponds to the actual one (or, as seems more likely, the culturally dominant one). That technology could assist you in this august task—the task of a lifetime for most of us—should not obscure the fundamental conservatism of this vision. The vast digital library is there to help you answer the question with which you began.

Greg Crane imagines a library in which the books talk to each other—each one embedded in a swirl of data mining and machine learning algorithms. What do we do with a million books? His answer is boldly visionary: "[E]xtract from the stored record of humanity useful information in an actionable format for any given human being of any culture at any time and in any place." He notes that this "will not emerge quickly," but one might legitimately question whether, strictly speaking, such a thing is logically possible for the class of problems traditionally held within the province of screwing around. What "useful information" was Zappa looking for (in, of all places, *LOOK)?.* He didn't really know and couldn't say.

Zappa would have loved the idea of "actionable formats," however. As it turns out, it took him over a year to find a copy of a Varèse record, and when he finally did, he didn't have the money to buy it. He ended up having to convince the saleman to part with it at a discount. Lucky for us, the salesman's "network of transactions" was flawed.

How would Zappa's adventure have played out today? *LOOK Online* mentions Varèse, and the "actionable format" is (at best) a click away, and at worst, over at Pirate Bay. And it's better than that. If you like Varèse, you might also like Messiaen's *Quartet for the End of Time,* which Messiaen actually wrote in a prison camp during the Second World War, the fifth movement of which (the piece, not the war) is based on an earlier piece which uses six Ondes Martinot, which is not only one of the first electronic instruments, but possibly the most beautiful sound you have ever hearde. And I don't believe this. There's a guy in Seattle who is trying to *build* an Ondes, and he's already rigged a ring controller to a Q125 Signal Processor. And he's got video.

This is browsing. And it's not like being in a library at all.

Is it possible to imagine this kind of highly serendipitous journey replacing the ordered mannerism of conventional search? It's important here to note that the choice is not between Google and Stumble—between surfing and asking Jeeves. It's not a matter of replacing one with the other, as any librarian will tell you. It is rather to ask whether we are ready to accept surfing and stumbling—screwing around, broadly understood—as a research methodology. For to do so would be to countenance the irrefragable complexities of what "no one really knows." Could we imagine a world in which "Here is an ordered list of the books you should read," gives way to, "Here is what I found. What did you find?" Because that is the conversation I and many other professional scholars and intellectuals are having on Twitter every single day, and it's not clear that we are worse for it.

There are concerns, of course. A humanist scholar—of whatever discipline, and however postmodern—is by definition a believer in shared culture. If everyone is screwing around, one might legitimately wonder whether we can achieve a shared experience of culture sufficient to the tasks we've traditionally set for education—especially matters such as participation in the public square. Concerns about a media landscape so ramified as to allow you to listen only to those ideas with which you already agree are not without foundation. But these questions are no sooner asked than answered by the recent history of the World Wide Web. Today, the dominant format of the Web is not the "Web page," but the protean, modded forum: Slashdot, Reddit, Digg, Boing Boing, and countless others. They are guides of a sort, but they describe themselves vaguely as containing "stuff that matters," or, "a directory of wonderful things." These sites are at once the product of screwing around and the social network that invariable results when people screw with each other.

As usual, they order these things much better in France. Years ago Roland Barthes made the provocative distinction between the "readerly text" (where one is mostly a passive consumer), and the "writerly text," where, as he put it, the reader, "before the infinite play of the world (the world as function) is traversed, intersected, stopped, plasticized by some singular system (Ideology, Genus, Criticism) which reduces the plurality of entrances, the opening of networks, the infinity of languages." Many have commented on the ways such thoughts appear to anticipate the hypertext, the mashup, and the Web. But Barthes himself doubted whether "the pleasure of the text"—the writerly text—could ever penetrate the institions in which readerly paths through culture are enshrined. He writes:

> What relation can there be between the pleasure of the text and the institutions of the text? Very slight. The theory of the text postulates bliss, but it has little institutional future: what it establishes, its precise accomplishment, its assumption, is a practice (that of the writer), not a science, a method, a research, a pedagogy; on these very principles, this theory can produce only theoreticians or practitioners, not specialists (critics, researchers, professors, students). It is not only the inevitably metalinguistic nature of all institutional research which hampers the writing of textual pleasure, it is also that we are today incapable of conceiv-

> ing a true science of becoming (which alone might assemble our pleasure without garnishing it with a moral tutelage). (60)

Somewhere in there lies a manifesto for what the world looks like when digital humanities becomes the humanities. Have we not already begun to call ourselves "a community of practice," in preference to "a science, a method, a research, a pedagogy?"

But the real message of our technology is something entirely unexpected—a writerly, anarchic text that is more useful than the readerly, institutional text. *Useful* and *practical* not in spite of its anarchic nature, but as a natural consequence of the speed and scale that inhere in all anarchic systems. This is, if you like, the basis of the Screwmeneutical Imperative. There are so many books. There is so little time. Your ethical obligation is neither to read them all nor to pretend that you have read them all, but to understand each path through the vast archive as an important moment in the world's duration—as an invitation to community, relationship, and play.

Works Cited

Aquinas, Thomas. *Summa Theologiae.* Vol. 1. Scotts Valley, CA: NovAntiqua, 2008.

Arnold, Matthew. Culture and Anarchy *and Other Writings.* 1869. Ed. Stefan Collini. Cambridge: Cambridge UP, 1993.

Barthes, Roland. *The Pleasure of the Text.* New York: Farrar-Hill, 1975.

Crane, Gregory. "What Do You Do with a Million Books?" *D-Lib Magazine* 12.3 (2006)

Durant, Will. *Our Oriental Heritage.* Story of Civilization 1. New York: Simon and Shuster, 1963.

Jefferson, Thomas. "To John Garland Jefferson" 11 June 1790. *The Works.* Vol. 6. New York: Putnam, 1905.

Moretti, Franco. "Conjectures on World Literature." *New Left Review* 1 (2000): 54-68.

Mueller, Martin. "Digital Shakespeare or Toward a Literary Informatics." *Shakespeare* 4.3 (2008): 284–301.

Occhiogrosso, Peter. *The Real Frank Zappa Book.* New York: Picador, 1990.

Riding, Alan. "Read It? No, but You Can Skim a Few Pages and Fake It" *New York Times.* 24 Feb 2007. Web.

Reading Salon #2: When All You Have Is a Twitter API, Every Problem Looks Like a Hashtag

Moderators: Michael Stevenson and Erik Borra

Bruns, Axel, and Stefan Stieglitz. 2012. “Quantitative Approaches to Comparing Communication Patterns on Twitter.” Journal of Technology in Human Services 30 (3-4): 160–185.

Gilbert, Eric, and Karrie Karahalios. 2009. “Predicting Tie Strength with Social Media.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 211–220. CHI ’09. New York, NY, USA: ACM.

Hecht, Brent, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. “Tweets from Justin Bieber’s Heart: The Dynamics of the Location Field in User Profiles.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 237–246. CHI ’11. New York, NY, USA: ACM.

Marres, Noortje. 2012. “The Redistribution of Methods: On Intervention in Digital Social Research, Broadly Conceived.” The Sociological Review 60: 139–165.

Puschmann, Cornelius, and Jean Burgess. 2013. “The Politics of Twitter Data”. SSRN Scholarly Paper ID 2206225. Rochester, NY: Social Science Research Network.

Rieder, Bernhard. 2012. “The Refraction Chamber: Twitter as Sphere and Network.” First Monday 17 (11) (November 4). Twitter data.

Journal of Technology in Human Services
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/wths20

Quantitative Approaches to Comparing Communication Patterns on Twitter

Axel Bruns [a] & Stefan Stieglitz [b]
[a] Queensland University of Technology, Brisbane, Australia
[b] University of Münster, Münster, Germany

Published online: 06 Dec 2012.

To cite this article: Axel Bruns & Stefan Stieglitz (2012): Quantitative Approaches to Comparing Communication Patterns on Twitter, Journal of Technology in Human Services, 30:3-4, 160-185

To link to this article: http://dx.doi.org/10.1080/15228835.2012.744249

Journal of Technology in Human Services, 30:160–185, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1522-8835 print/1522-8991 online
DOI: 10.1080/15228835.2012.744249

Quantitative Approaches to Comparing Communication Patterns on Twitter

AXEL BRUNS
Queensland University of Technology, Brisbane, Australia

STEFAN STIEGLITZ
University of Münster, Münster, Germany

To date, the available literature mainly discusses Twitter activity patterns in the context of individual case studies, while comparative research on a large number of communicative events and their dynamics and patterns is missing. By conducting a comparative study of more than 40 different cases (covering topics such as elections, natural disasters, corporate crises, and televised events) we identify a number of distinct types of discussion that can be observed on Twitter. Drawing on a range of communicative metrics, we show that thematic and contextual factors influence the usage of different communicative tools available to Twitter users, such as original tweets, @replies, retweets, and URLs. Based on this first analysis of the overall metrics of Twitter discussions, we also demonstrate stable patterns in the use of Twitter in the context of major topics and events.

KEYWORDS communicative patterns, media events, public communication, social media, Twitter

INTRODUCTION

Since 2006, microblogging has become an increasingly widely used tool for communication on the Internet. Twitter, as one of the first and most popular microblogging providers, has some 140 million users, with some 340 million

Received May 16, 2012; revised August 25, 2012; accepted October 16, 2012.

Address correspondence to Axel Bruns, ARC Centre of Excellence for Creative Industries and Innovation, Creative Industries Faculty, Queensland University of Technology, Z1-515, Creative Industries Precinct, Musk Ave, Kelvin Grove, Qld. 4059, Australia. E-mail: a.bruns@qut.edu.au

tweets posted each day (Twitter, 2012). In contrast to social networking sites (SNS) such as Facebook, the reach of posts on Twitter is not necessarily limited to a specific group (such as subscribed "friends" or "followers"); rather, posted messages are public by default and may also be found by visitors searching the site or tracking the Twitter stream. Each user is thus able to create public posts to initiate discussions, to participate in debates, and to follow the communication of others. To manage these communicative flows and increase the efficiency of public message exchanges, Twitter users have adapted a variety of methods to classify their contributions (tweets)—for example, as a public response (@reply) or a shared message originating from another user (retweet).

Twitter has now become a widely used communications channel across a wide range of applications, from politics, journalism and crisis communication (Bruns & Burgess, 2011a; Christensen, 2011; Larsson & Moe, 2011; Lotan, Ananny, Gaffney, & boyd, 2011; Bruns, Burgess, Crawford, & Shaw, 2012; Mendoza, Poblete, & Castillo, 2010; Palen, Starbird, Vieweg, & Hughes, 2010; Stieglitz & Dang-Xuan, in press) through its use as a backchannel for television shows, cultural and sporting events, and conferences to a wide variety of uses for everyday interpersonal communication (e.g., boyd, Golder, & Lotan, 2010; Deller, 2011; Dröge, Maghferat, Puschmann, Verbina, & Weller, 2011; Marwick & boyd, 2011; Papacharissi, 2011; Weller, Dröge, & Puschmann, 2011). Significant research into some such uses is now emerging, but largely remains in the form of topic-, context-, and event-related case studies that are able to shed substantial light on specific uses of Twitter but do not yet lead to a more comprehensive overall picture of how Twitter is used.

Individuals and organizations may use Twitter to subscribe to the update feeds of other users, as well as to publish their own short messages (to a maximum of 140 characters) about various topics (e.g., from personal and professional updates to press releases and other corporate information). To widely disseminate information on Twitter, the mechanism of retweeting has been adopted by users. By retweeting, users may not only share information but also entertain a certain audience or (by adding comments to the retweets) publicly agree or disagree with someone (boyd, Golder, & Lotan, 2010). As a result, Twitter has become an important platform for users to spread information about topics of shared interest: retweets propagate the original tweet to a new set of audiences, namely, the followers of the retweeting user. Given the growing Twitter user base, the high speed of information dissemination on Twitter, and the significant influence of Twitter as a driver of web traffic, new questions arise about the way it is used to support public information sharing and information search.

Other studies have already made some first steps to investigate how and why certain information items spread more widely than others (Stieglitz & Dang-Xuan, forthcoming [Suh, Hong, Pirolli, & Chi, 2010]). However, so far the literature mainly discusses Twitter activity patterns in the context of individual case studies, while comparative research on a large number of discussions and their dynamics and patterns is missing. By conducting a comparative study of several dozen different cases (including topics such as elections, natural disasters, corporate crises, and televised sporting and cultural events) we have identified a number of distinct types of discussion that can be observed on Twitter. Drawing on a range of communicative metrics, we show that thematic and contextual factors influence the usage of different communicative tools available to Twitter users, such as original tweets, @replies, retweets, and URLs. We also demonstrate patterns in the structure of the user community involved (e.g., number of participants, relevance of lead users). This article presents a first analysis of the overall metrics of Twitter discussions relating to different areas of content, and outlines two standard, stable types of Twitter usage in the context of major topics and events. As such, it represents a significant advance for research that investigates the usage of different communication tools in public discussions.

This article pursues this larger picture by exploring general patterns of Twitter usage, drawing on detailed usage metrics for a wide range of cases and events over the past 2 years. By collating these data points and identifying cases that exhibit similar patterns of activity, we observe a range of common, apparently well-established user practices on Twitter. We suggest that these observations point to regularities in the popular responses to specific themes and events that may also be identified well beyond the Twitter platform itself.

RELATED WORK

In recent years, a substantial amount of literature has been published in the field of Twitter communication. Therefore, we provide a short literature review of those articles that explicitly reflect metrics within Twitter communication in the field of politics, natural and human disasters, and entertainment and brand-related communication.

Politics

In a study of approximately 100,000 messages containing a reference to either a political party or a politician in the context of the 2009 German federal election, Tumasjan, Sprenger, Sandner, and Welpe (2011) show that Twitter is used extensively for the dissemination of politically relevant information and that the mere number of party mentions accurately reflects the election result, suggesting that microblogging messages on Twitter seem to validly mirror the political landscape offline and can be used to predict election results to a certain extent. Conover et al. (2011) examine two networks of political communication on Twitter with more than 250,000 tweets from the 6 weeks leading up to the 2010 U.S. congressional midterm elections. Using a combination of network clustering algorithms and manually annotated data, the authors demonstrate that the network of political retweets exhibits a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users. Surprisingly, this is not the case for the user-to-user mention network, which is dominated by a single politically heterogeneous cluster of users in which ideologically opposed individuals interact at a much higher rate compared to the network of retweets. Similarly, Yardi and boyd (2010) find that in a political context Twitter users are more likely to interact with others who share the same views as they do in terms of retweeting, but they are also actively engaged with those with whom they disagree. In addition, replies between like-minded individuals would strengthen group identity, whereas replies between different-minded individuals would reinforce in-group and out-group affiliation. In a large-scale study, Suh, Hong, Pirolli, and Chi (2010) addressed these questions and identified several factors that significantly impact on the retweetability of Twitter messages (tweets), including the presence of URLs and hashtags, as well as the number of followers and the age of the originating user's account. Beside these case-based analyses, Stieglitz and Dang-Xuan (2012) provide a general framework that presents methods for social media analytics in the political context.

Natural and Human Disasters

In recent years, a growing body of literature has emerged in the field of social media and crisis communication (Bruns, Burgess, Crawford, & Shaw, 2012; Hughes & Palen, 2009; Mendoza, Poblete, & Castillo, 2010; Palen, Starbird, Vieweg, & Hughes, 2010). Cheng, Sun, Hu, and Zeng (2011) investigated Twitter as a tool to monitor and capture emerging trends and patterns of time-critical knowledge. They extensively evaluated a diffusion-based recommendation framework and a proposed algorithm using Twitter data collected during the early outbreak of H1N1 Flu. Studies by Bruns, Burgess, Crawford, and Shaw (2012) and by Cheong and Cheong (2011) analyzed Twitter-based communication in the context of natural disasters, focusing on the Australian floods in 2011. By using social network analysis methods, they found that several different groups of actors, including affected locals, emergency services, and mainstream media organizations, played important roles in providing and sharing information about the disaster.

Other studies (Bruns, Highfield, & Burgess, in press; Lotan, Ananny, Gaffney, & boyd, 2011; Vis, 2012) examined uses of Twitter during major civil unrest in the context of the 2011 London and UK riots or the Arab Spring uprisings in a number of North African and Middle East countries. They identified a diverse range of uses of social media for information dissemination alongside and in addition to other media channels and word-of-mouth information, and also highlighted significant differences in activity between local and more distant observers of these events.

Entertainment and Brand-Related Communication

Stieglitz and Krüger (2011) investigated a 2010 brand crisis involving car manufacturer Toyota and showed that, measured by the published number of tweets, crisis discussions are characterized by peaks and quiet periods in the communication of enterprise-related issues. Further, they found that the lead users involved in the Twitter debate played an important role in the discussion, for example by publishing a significantly high amount of all tweets and generating a large amount of retweets.

Park, Cha, Kim, and Jeong (2009) investigated the Domino's Pizza crisis and analyzed the diffusion of bad news through Twitter. They separately classified sentiments in tweets generated by customers and those generated by the enterprise itself, and proved that the diffusion of bad news is faster than that of other types of content, such as apologies.

In a study of users and their behaviors in the Twitter network, Krishnamurthy, Phillipa, and Arlitt (2008) identified three types of users (broadcaster, acquaintances, and miscreants) by analyzing a crawled data set that covered nearly 100,000 users. The broadcasters, also called power-tweeters, are characterized by a large number of followers as well as a large amount of self-created postings. One finding in this study was that these users update their status more often and post more tweets than users of the two other categories.

METHODOLOGY

In order to establish a sound basis for the identification of shared patterns in Twitter-based communication around specific issues, the majority of the Twitter phenomena which we observe in this article are centered around common hashtags (brief keywords included in tweets, prefixed with the hash symbol #). Hashtags are an originally user-generated mechanism for making messages related to a specific topic more easily discoverable and are now well-supported by central Twitter infrastructure as well as by specific Twitter client software; it is now possible for users (and even for nonregistered visitors to the site) to search Twitter for specific hashtags, and to follow the stream of new messages containing specific hashtags in real time. This makes hashtags a useful and an important mechanism for coordinating conversations around identified themes and events, ranging from breaking news (such as #eqnz for the 2010/11 earthquakes in Christchurch, New Zealand) through major media events (e.g., #euro2012 for the 2010 European Football Championships) to viral marketing campaigns (such as #kony2012 and #stopkony for the campaign to bring a fugitive Ugandan warlord to justice). Beyond such world events, hashtags are also used to coordinate much more low-key discussions and user communities, from providing a backchannel for conference delegates to organizing Twitter-based user meetups (such as the long-standing #phdchat, a global discussion for PhD candidates). Finally, a different use of hashtags, which we do not consider in detail here, is as markers of emphasis or emotion, as in, "My bus is running late again. #fail."

Hashtags, then, may emerge ad hoc in response to breaking news and other unforeseen events, spreading virally as more and more users with an interest in the topic see the hashtag in their Twitter feeds and begin to use it themselves (see Bruns & Burgess, 2011b). They may also be used repeatedly for recurring events (such as #ausvotes for Australian federal elections, or #eqnz for each of the four major earthquakes which affected Christchurch in 2010/2011). Alternatively, they may be promoted *praeter hoc* by relevant organizations as the appropriate hashtag to be used for an upcoming event (this is the case for backchannel hashtags for conferences or TV shows, for example). Such diverse hashtags may in turn attract widely varying groups of users. Breaking news events, especially where they are of national or global relevance, may find hundreds of thousands of Twitter users posting or retweeting hashtagged messages, while hashtags related to conferences or TV shows may involve only a much smaller number of users who happen to be attending or watching at the time. Standing hashtags for the discussion of specific continuing topics (from #phdchat to the day-to-day tracking of long-term events such as the popular revolts in #Libya, #Egypt, or #Syria), in turn, may involve only a comparatively small group of committed contributors. At the same time, they may see a temporary influx of a large number of interested users as key developments unfold and are widely covered by mainstream media outlets.

Using the Twitter Application Programming Interface (API), it is comparatively simple to capture comprehensive data sets of the vast majority of all tweets containing a specific hashtag (within limits determined by the reliability of the API and of real-time Twitter tracking tools; cf. Bruns & Liang, 2012). During 2010–2012, as part of a project collaboration between Queensland University of Technology, Brisbane, and the University of Münster, we have done so for some 40 hashtags as well as a number of non-hashtagged keywords (which we discuss later). Individual hashtag data sets for this study were captured using the open-source platform *yourTwapperkeeper* (2012), which utilizes Twitter streaming API and search API functionality to capture, in real time, any tweets containing the keywords (including hashtags) selected by the operator. For more details on Twitter research methods using *yourTwapperkeeper* and alternative technologies, see Bruns and Liang (2012). *yourTwapperkeeper* does not provide for post hoc data gathering; it is able only to capture tweets for set keywords as they are sent. Therefore, the selection of hashtag and keyword data sets used for this study was a function of the long-term research interests of the Brisbane and Münster research groups, which specialize in political, crisis, and brand communication research. In themselves, these over 40 data sets cover a diverse range of uses, therefore, but we also encourage the further extension of this initial work through the addition of hashtag and keyword metrics extracted from data sets gathered by researchers interested in other areas of communication using social media.

To better understand the diversity of uses evident in the present collection of cases, and to identify any common patterns between individual cases, we draw on a catalogue of metrics for describing the communicative patterns which may be observed for each hashtag (see Bruns & Stieglitz, in press, for a detailed introduction of these metrics). In the first place, these included:

- The number of *tweets* in the hashtag data set.
- The number of *unique users* contributing to the hashtag data set.
- The percentage of *original tweets* in the hashtag data set (i.e., tweets that are neither @replies nor retweets).
- The percentage of *genuine @replies* in the hashtag data set (i.e., @replies that are not retweets).
- The percentage of *retweets* in the hashtag data set.[1]
- The percentage of tweets in the hashtag data set that contain URLs.

Additionally, we also divided the total user base for each hashtag data set into three groups, following a standard 1/9/90 distribution (Tedjamulia, Dean, Olsen, & Albrecht, 2005):

- The top 1% of most active *lead users*.
- The next 9% of still *highly active users*.
- The remaining 90% of *least active users*.

For each of these three groups in each hashtag data set, we again calculated the metrics already outlined, taking into account only the tweets sent by that percentile group. Compared across the groups, this provides a measure for each hashtag of how dominant within the overall hashtag conversation the leading user groups were. This enabled us to examine any obvious differences in the Twitter activity patterns of the three user groups.

In the following discussion, we collate and compare these metrics for the range of hashtags which we tracked over the past two years. This enables us to identify communication patterns that are common across these diverse cases, and to develop a typology of hashtagged Twitter usage. First, however, we provide an overview of the hashtag data sets that were used in this analysis, and outline their relevant features.

Hashtag Data Sets

This study draws on a wide variety of data sets, whose key features we outline in Table 1. While the scope of this article does not permit a detailed discussion of the themes and contents of each data set, we note that these cases encompass a wide variety of topical hashtag uses. These range from political themes through natural disasters to entertainment and sports; from

TABLE 1 Overview of hashtag and keyword data sets used for the comparative analysis

Hashtag/ keyword	Description	Theme	Timeframe	Notes on time frame	Total unique users	Total tweets
#0zapftis	Scandal around trojan horse virus developed by German intelligence service	Politics	11–31 Oct. 2011	Three weeks following scandal	6716	26158
#aflgf	Australian Football League 2011 grand final	Sports	1–2 Oct. 2011	Matchday and following day	3793	6135
#angryboys	*Angry Boys*: popular weekly TV sitcom on Australian free-to-air television	Television	12 May to 31 July 2011	Full season	30121	63333
#auspol	Australian politics (general discussion)	Politics	8 Feb. to 8 Dec. 2011	Eight months	26290	854019
#ausvotes	Australian federal election 2010	Politics	20–22 Aug. 2010	Three days around election day	36286	415511
#chch	February 2011 earthquake in Christchurch, New Zealand	Natural disaster	22–28 Feb. 2011	First week after earthquake (initial alternative to #eqnz)	9688	24400
#earthquake	March 2011 earthquake and tsunami in Japan	Natural disaster	11–24 Mar. 2011	First 2 weeks after earthquake	183794	358737
#egypt	Arab Spring protests	Political unrest	26 Feb. to 26 Nov. 2011	Nine month period since first major protests	281978	6277782
#eqnz	February 2011 earthquake in Christchurch, New Zealand	Natural disaster	22 Feb. to 7 Mar. 2011	First 2 weeks after earthquake	37635	156940
#eurovision	Eurovision Song Contest 2011	Television	9–15 May 2011	Semifinals and and finals broadcasts on 10/12/14 May	137745	520543
#ge11	Irish general election	Politics	26 Feb. 2011	Election day	6151	28468
#gobacksbs	*Go Back to Where You Came From*: weekly reality TV show with political connotations on Australian free-to-air television	Television	21 June to 4 July 2011	Second half of season	8691	29009
#irene	Hurricane Irene along the East Coast of the United States	Natural disaster	27 Aug. to 17 Sep. 2011	Three weeks following first impact	37891	64315
#kony2012	Viral campaign to arrest warlord Joseph Kony	Politics	8–21 Mar. 2012	First 2 weeks of campaign	80874	101425

(*Continued*)

TABLE 1 Continued

Hashtag/ keyword	Description	Theme	Timeframe	Notes on time frame	Total unique users	Total tweets
#libya	Arab Spring protests	Political unrest	26 Feb. to 26 Nov. 2011	Nine-month period since first major protests	363489	3825272
#londonriots	Violent riots in London and the United Kingdom	Crisis	8–21 Aug. 2011	First 2 weeks after riots	127631	212213
#masterchef	*Masterchef*: popular weekly reality TV show on Australian free-to-air television	Television	1 May to 8 Aug. 2011	Whole season	54117	210773
#mkr	*My Kitchen Rules*: popular weekly reality TV show on Australian free-to-air television	Television	13 Feb. to 31 Mar. 2012	Final 27 episodes	12671	63866
#mw3	*Modern Warfare 3*: popular computer game	Net culture	1–30 Nov. 2011	One month around official launch	207858	413922
#norway	Right-wing terrorist attacks in Oslo and Utøya	Crisis	24 July to 9 Aug. 2011	First 2 weeks after attacks	38224	63244
#nrlgf	Australian National Rugby League 2011 grand final	Sports	1–2 Oct. 2011	Buildup and match day	2049	4182
#occupy	Global Occupy protests	Political protests	19 Dec. 2011 to 19 Apr. 2012	Four months of protests	121952	560560
#occupywallstreet	Occupy protests in New York	Political protests	27 Sep. to 27 Nov. 2011	Three months at height of protests	234514	885174
#oscars	Academy Awards 2011	Entertainment	27 Feb. 2011	Event day	236103	639251
#qanda	*Q&A*: popular weekly political talk show on Australian free-to-air television	Politics	21 Feb. to 21 Nov 2011	Whole season	246231	47131
#qldfloods	Major flooding in southeast Queensland	Natural disaster	10–16 Jan. 2011	First week of floods	15553	35658
#qldvotes	2012 Queensland state election	Politics	23–25 Mar. 2012	Three days around election day	5788	17456
#riotcleanup	Cleanup after violent riots in London and the United Kingdom	Crisis	8–21 Aug. 2011	First two weeks after riots	38511	53381
#royalwedding	Wedding between Prince William and Kate Middleton	Entertainment	29 Apr. 2011	Wedding day	492566	926527

#spill	Party room revolt against Australian Prime Minister Kevin Rudd	Politics	23–24 June 2010	First rumors and confirmation of party room vote	11309	46937
#stopkony	Viral campaign to arrest warlord Joseph Kony	Politics	8–21 Mar. 2012	First 2 weeks of campaign	117050	140958
#syria	Arab Spring protests	Political unrest	26 Mar. to 26 Nov. 2011	Eight months since first major protests	229030	5230025
#tdf	Tour de France 2011	Sports	4 July to 26 July 2011	Whole tour (except first two days)	94830	427467
#tsunami	March 2011 earthquake and tsunami in Japan	Natural disaster	11 Mar. to 11 Apr. 2011	First month after earthquake	529913	948640
#ukriots	Violent riots in London and the United Kingdom	Crisis	8–21 Aug. 2011	First 2 weeks after riots	61766	126664
#wikileaks	Political controversy	Politics	26 Feb. to 26 Nov. 2011 + 1–7 Sep. 2011	Nine months + period around Julian Assange arrest in the United Kingdom	119853 16930	422635 35451
bin laden	Death of Osama bin Laden	Politics	2 May to 2 June 2011	First month following bin Laden killing in Abbottabad	1868127	3987919
masterchef	*Masterchef*: popular weekly reality TV show on Australian free-to-air television	Television	1 May to 8 Aug. 2011	Whole season	238689	609714
qantas	Global grounding of flights by Qantas management in response to industrial action	Brand crisis	26 Oct. to 8 Nov. 2011	Two weeks around grounding crisis	42144	98636
steve jobs	Death of Apple founder Steve Jobs	Net culture	7 Oct. to 7 Nov. 2011	One month after death	403321	562411
tsunami	March 2011 earthquake and tsunami in Japan	Natural disaster	11 Mar. to 11 Apr. 2011	First month after earthquake	1936553	4246019

breaking news events through foreseeable, regularly occurring activities to channels for continuous thematic discussion; from local issues to global events; from events that unfolded over the span of a few hours to themes that were discussed for close to a year (and remain active beyond the time span covered in our analysis); and from activities that involve only a relatively small subset of the global Twitter user base, measuring in the thousands, to events that attracted the participation of close to 2 million unique users or generated more than 6 million tweets (see Table 1).

Real-time data collection for these data sets generally commenced as the hashtags related to specific themes and events became prominent on Twitter, especially in the case of acute crisis events. This required researchers to react speedily as news of these crises broke (as in the case of natural disasters such as the Christchurch earthquake or the Japanese tsunami), to rapidly determine the most prominent hashtags (#eqnz, #tsunami), and to add those hashtags to the existing *yourTwapperkeeper* installations for tracking. In other cases (such as #royalwedding, #eurovision, or #ausvotes), hashtags were foreseeable prior to the event and could be added to the tracker in advance. In each case, however, we have further determined appropriate start and end points for the data timeframes to be considered in the present article, in order to focus on the key period of activity for each hashtag or keyword. For natural disaster events, this usually means limiting the analysis to the first days or weeks after the initial disaster event, and for election-related discussion, to the days around election day itself. The specific time frames chosen for each hashtag or keyword are outlined in Table 1. It is necessary to draw on this disparate collection of data sets for our analysis in order to detect any patterns of Twitter use that persist even in spite of such marked differences between individual cases.

In addition to the hashtag data sets (which contain only those tweets about a topic that were explicitly hashtagged—e.g., #tsunami for the March 2011 tsunami), we also include five *keyword* data sets. These were gathered by capturing all tweets that contained only the specific keyword (e.g., "tsunami"), regardless of whether or not the # symbol was prefixed to that term. We discuss these data sets in more detail in the following. Our aim in including them in the following analysis is to examine whether there are any indications that beyond the use of dedicated hashtags, topical communication patterns on Twitter may follow similar principles as we outline them for deliberately hashtagged exchanges.

USER ACTIVITY METRICS

Given the divergence in the number of tweets and unique users for each data set, it is first useful to compare the relative prominence of the leading user groups across these cases. Figure 1 presents the relative number of

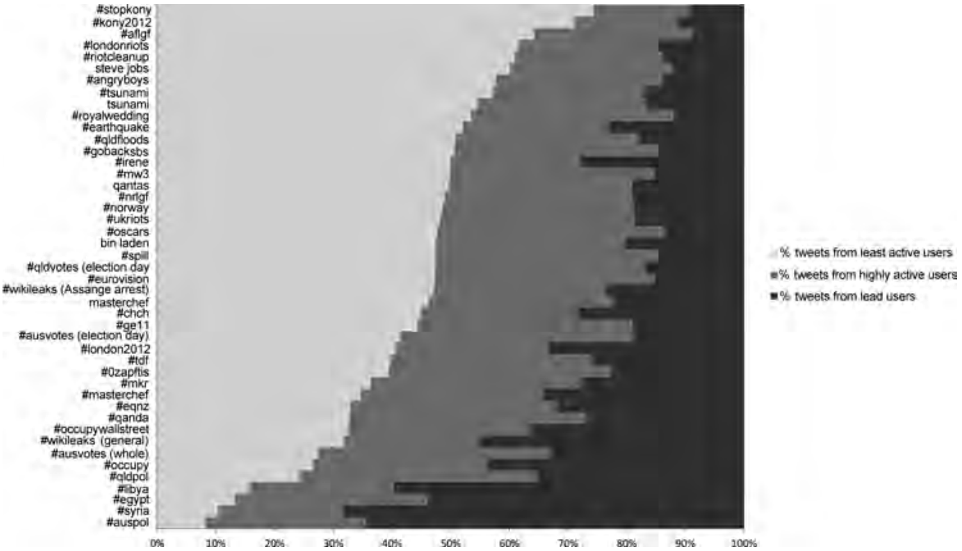


FIGURE 1 Relative contributions from the three user groups.

tweets contributed by each of the three user groups we have outlined already: lead users (top 1% most active users), highly active users (next 9% of active users), and least active users (the remaining 90% of users: the "long tail" of the user base).

Clear distinctions between the cases examined here emerge from this analysis. Roughly half of the cases are comparatively dominated by the two most active user groups, who (in combination) contribute 50% or more of the total number of tweets. For a smaller number of cases, that percentage grows to well above 70%; here, the "long tail" of least active users remains largely silent, while any meaningful exchanges take place mainly within a dedicated in-group of highly active participants.

It is notable in this context that the hashtags that feature the most active groups of leading users are generally also those that cover the longest time frames. In our comparison, the 10 data sets that see the fewest tweets from the least active user group are #syria, #egypt, and #libya (each of which attracted hundreds of thousands of participants and was active throughout 2011); #auspol, a standing hashtag for the discussion of Australian federal politics with a small but highly active contributor community, and #ausvotes (for the discussion of the 2010 Australian federal election); #occupy, #occupywallstreet, and #wikileaks (which serve as key distribution tools for information about global political protest and counterculture movements, over the long term, and attract hundreds of thousands of participants); and #qanda (the hashtag promoted by the Australian Broadcasting Corporation for its weekly political talk show *Q&A*).

By contrast, those hashtags that feature the most active “long tails” of contributors in our analysis are also those that unfold over comparatively short time frames. They include scheduled media events such as #royalwedding (the 29 April 2011 wedding between Prince William and Kate Middleton) or #aflgf (the 1 Oct. 2011 Australian Football League grand final); breaking news events such as the March 2011 Japanese tsunami (in both its hashtag and keyword variants); the August 2011 London riots and the subsequent #riotcleanup initiative organized by affected communities; the death of Apple leader Steve Jobs (which we captured in a keyword data set centered on mentions of “Steve Jobs”); and short-lived viral marketing campaigns such as the initiative to bring Ugandan warlord Joseph Kony to justice (under the hashtags #kony2012 and #stopkony).

From these observations, we suggest that the relative prominence of leading user groups in a hashtag conversation is related to the overall longevity of the hashtag itself (the amount of time during which it was significantly active, and during which we gathered tweets for it). In a comparatively new hashtag, more striated community structures have not yet had a chance to crystallize, while in a long-lived hashtag it is perhaps logical that committed long-term contributors will emerge as lead users as more casual participants come and go. Figure 2, which plots the longevity of hashtags against the relative contribution made by the 90% least active users, supports this finding. Although the possibility is intriguing, it should also be noted that our evidence does not permit us to establish any causal relations between these factors. However, from our data alone it is impossible to determine (1) whether hashtags persist for the longer term *because* a strong group of lead users keeps them going, or (2) whether these lead users inevitably emerge if a hashtag continues to remain active for a long enough time.

TWEET TYPE METRICS

The hashtag data sets examined here also differ widely in their underlying communicative practices, as Figure 3 shows. Here, we examine the relative presence of the three key tweet types we have outlined already (original tweets, genuine @replies, and retweets), as well as the occurrence of URLs in any such tweets. It is again obvious that there are distinct differences in communicative patterns between hashtags: Most obviously, a small number of cases consist overwhelmingly (at 65% or above) of original tweets that neither mention nor reply to other users. These cases (and indeed, all hashtags with more than 50% original tweets) are also marked by the relative absence of URLs in tweets; the vast majority of this group of hashtags contain URLs in less than 20% of all tweets, while the average percentage of tweets with URLs for the remainder of our hashtags is close to 50%.

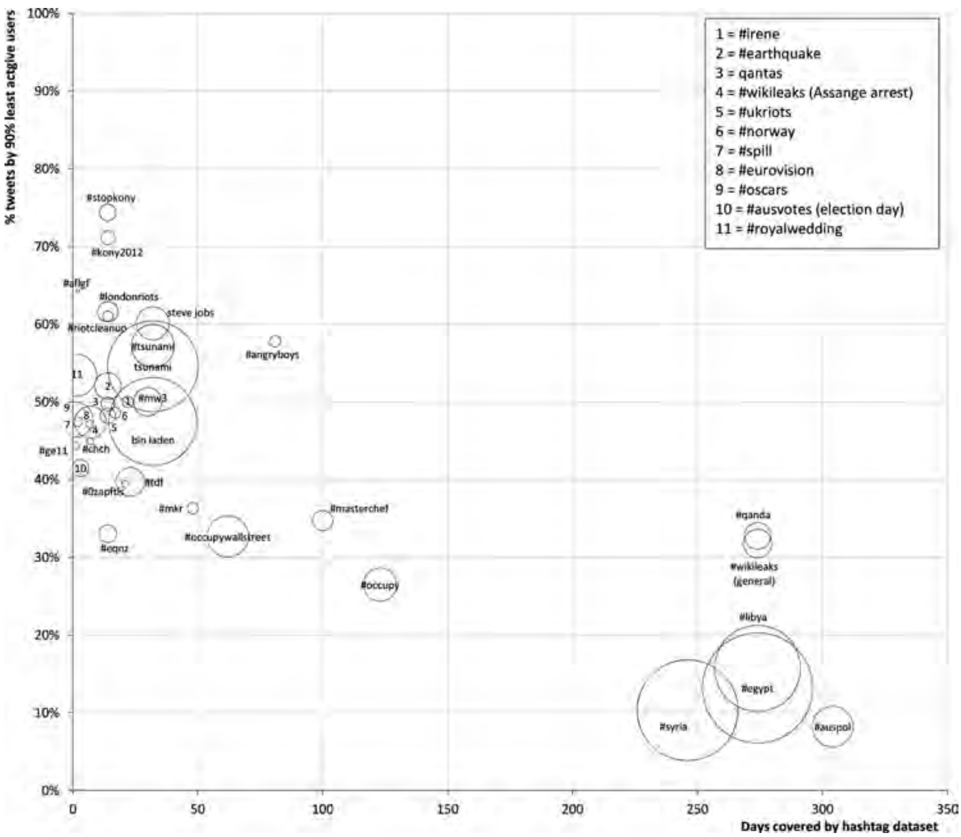


FIGURE 2 Hashtag longevity compared to percentage of tweets contributed by 90% least active users (size of data points indicates total number of tweets for each hashtag/keyword case).

It is notable in this context that hashtags that exhibit a large percentage of original tweets share a number of key contextual characteristics. For the most part, these hashtags relate to major media events, ranging from internationally televised entertainment broadcasts (#eurovision, #royalwedding, #oscars) through important sporting events (#tdf for the Tour de France, #aflgf for the Australian Football League Grand Final, #nrlgf for the Australian National Rugby League Grand Final), to popular daily or weekly television shows (the Australian reality TV programs *Masterchef* and *My Kitchen Rules*—#mkr—the sitcom *Angry Boys*, or the political talk show *Q&A*). Other Australian political events—such as #spill for the 2010 partyroom challenge against Prime Minister Kevin Rudd, or the #ausvotes discussion around election day 2010—also fit this model, as they were (or for #spill, rapidly became) major media events in their own right.

On the other hand, hashtags that saw a substantial amount of retweeting, and comparatively few original tweets, largely fall into a category that

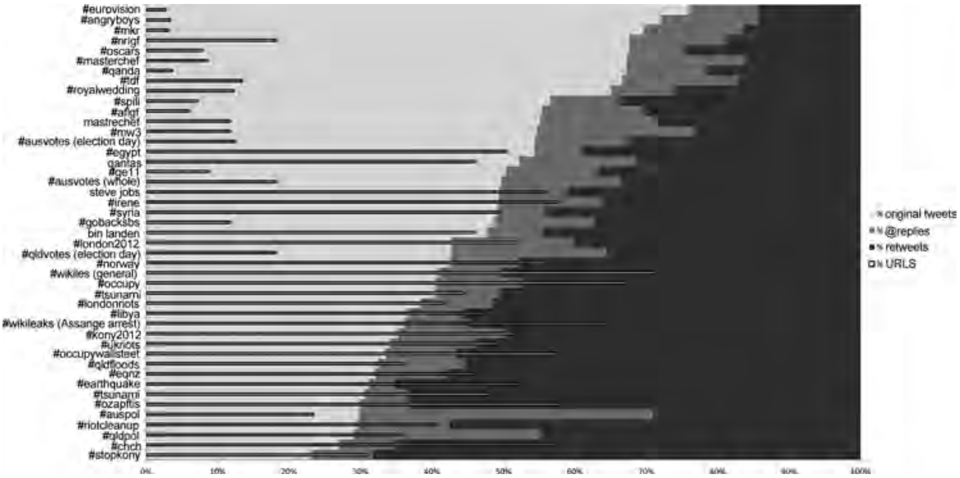


FIGURE 3 Relative percentages of different tweets types across all hashtags. (Note that the percentage of URLs is shown on a separate scale, as URLs can occur across all three tweet types.)

may be best described as "breaking news" or "rapid information dissemination"; they include, most obviously, many hashtags related to natural disasters, from #eqnz (and the alternative #chch, short for Christchurch) through #earthquake and #tsunami (both relating specifically to the March 2011 event in Japan), to #qldfloods, as well as to civil unrest (from #libya through #occupywallstreet to the London #riotcleanup). Additional examples for this category are #stopkony (an orchestrated viral marketing campaign that to some extent behaved like a crisis event) and #0zapftis (a scandal around a Trojan horse virus developed by German intelligence services for covert investigation purposes). Generally, such retweet-heavy hashtags also contain a substantial number of tweets with URLs: On average, half of all tweets in hashtags with more than 50% retweets contain URLs.

Between these two key metrics, patterns in genuine @replies for each case move somewhat more randomly, and this category generally accounts only for a relatively small percentage of tweets in each data set (with @replies constituting more than 41% of all tweets, #auspol is the one major exception to this rule; this may be related to the very well-established, dominant group of lead users in this case). In this context, a limitation of our hashtag-based Twitter research approach must be noted: As users respond to hashtagged tweets, they frequently do not again include the hashtag in their @replies, and such non-hashtagged replies are therefore not included in our data sets. Those users who do include a hashtag in their @replies, by contrast, usually do so explicitly to make their responses visible to a wider audience again; hashtagged @reply conversations are in essence performed in front of a larger public, in other words, but constitute a special case. For this reason, the following discussion largely focuses on the complementary metrics of original tweets and retweets only, as well as on the presence of URLs.

TOWARD A TYPOLOGY OF HASHTAGS

These observations enable the development of a tentative typology of hashtag events, based on the activity metrics which are observable in each case. For the data sets we have examined here, Figure 4 plots the percentage of URLs in tweets against the percentage of retweets in the overall data set, and indicates the combined contribution of lead and highly active users through the size of each data point.

On this graph, two distinct clusters of hashtag cases emerge. At the center of Figure 4 is a cluster that mainly contains hashtags relating to crises and other breaking news events. These range from natural disasters to political protests and civil unrest. This group of hashtags is characterized by both substantial posting of links to further information, and significant retweeting

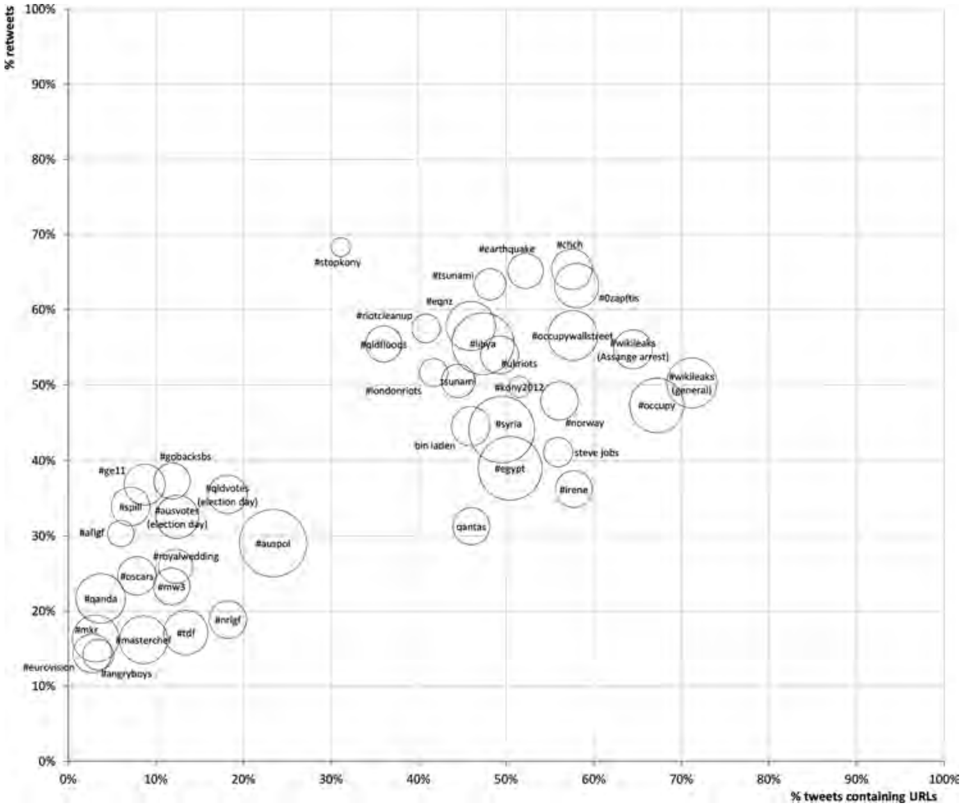


FIGURE 4 Percentage of URLs in tweets versus percentage of retweets among all tweets (size of data points shows combined contribution of lead users and highly active users).

activity. We suggest, therefore, that it is largely centered around a shared practice of *gatewatching* (Bruns, 2005): the collaborative identification, sharing, and passing along of situationally relevant information, here especially in the context of "acute events" (Burgess & Crawford, 2011).

Indeed, it is notable for this cluster of hashtags that most natural disasters, except for the #irene hashtag, for the 2011 Hurricane Irene, are positioned toward the top of the cluster (indicating an especially high percentage of retweets). This may indicate a widespread desire of users to help in sharing key emergency information, and a limited interest in posting comments or other statements about to the unfolding event, while political crises attract a comparatively higher number of such comments in the form of original tweets and @replies. By contrast, on the far right of the cluster we find a number of hashtags that are related to political protest movements (#wikileaks, #occupy, and #occupywallstreet[2]). Their positioning indicates a substantial percentage of URLs being shared through the hashtag, pointing perhaps to contributors' perception of these themes as countercultural issues that are under- or misreported by mainstream media and require constant support through online social networks. In this context, it is notable that #occupywallstreet, which deals more centrally with the struggle between protesters and law enforcement in New York, and the subset of #wikileaks activity around the arrest of Julian Assange share more similarities with the overall crisis cluster than the longer term #wikileaks and #occupy hashtags.

In addition to these hashtags, we have also included metrics for keyword data sets covering mentions of Steve Jobs following his death, for Osama bin Laden after his death in a raid on his Abbottabad compound, and for the Australian airline Qantas during a global grounding of all flights by management in response to an industrial dispute. Of these, the activity metrics for Jobs and bin Laden show strong similarities to other crisis events, pointing to similar gatewatching and news-sharing activities in the context of these breaking news events. The Qantas event behaves somewhat differently, due to a comparatively lower percentage of retweets (and thus a larger number of tweets making original comments). This is in keeping with the significant political implications of the event, beyond the international air transport crisis it caused. Further, we also include the keyword data set for "tsunami" in addition to the hashtag #tsunami, and find a comparatively smaller percentage of retweets for the keyword case. This indicates, not unexpectedly, that hashtagged tweets are more likely to be found and retweeted than non-hashtagged messages, but also points to the likelihood that overall Twitter activity patterns around crisis events, beyond their hashtagged core, are broadly similar to those for the crisis hashtag itself.

Finally, we find the #kony2012 hashtag at the center of the crisis cluster, while its cousin #stopkony is present as an outlier in the overall graph, with a very substantial percentage of retweets but comparatively few URLs. Further analysis must determine the reasons for the latter hashtag's divergent behavior, but it appears sensible that #kony2012, a campaign designed to be disseminated virally and with reference to further information on the campaign Website, and videos on YouTube, would show similar tendencies to the crisis hashtags. In essence, we may understand #kony2012 (and similar viral campaigns) as a deliberately “manufactured” crisis.

A second distinct cluster of hashtags is located in the bottom left quadrant of the graph. The hashtags assembled here are characterized by a very low percentage of URLs in each data set, as well as a comparatively low percentage of retweets. Put differently, these hashtags are mainly used to post original tweets and a limited amount of @replies, with few references to additional information outside of Twitter. The hashtags assembled in the acute events cluster are largely concerned with information sharing, whereas the hashtags in this second cluster are focused on original commentary.

Further, the majority of these hashtags are clearly related to mainstream media events, as we have already seen in the discussion of Figure 3. We therefore interpret these hashtags as cases in which Twitter functions as a backchannel for live events (especially as these events are broadcast by national and international television). These hashtags, in other words, support a shared experience of “audiencing” (e.g., Fiske, 1992): of talking back at the television (or at the live event), along with thousands of other viewers. This sense of temporary, imagined community persists even if—as our data show—actual direct interaction between users through hashtagged @replies and retweets remains relatively rare; it may be sufficient to observe the stream of hashtagged comments, even without engaging with and replying to them. Such a sense of community is further enhanced, of course, if—as is increasingly common practice—television shows include selected tweets from the hashtag stream in an on-screen ticker.

In this cluster, too, further subdivisions can be observed. Interaction through retweets is lowest for sporting and other entertainment events, while political themes attract a somewhat larger percentage of retweets—the #qldvotes, #ausvotes, and #ge11 (for the 2011 Irish general election) hashtags on their respective election days, as well as #spill (for the 2010 Australian political leadership crisis), are located toward the top of this cluster. *Go Back to Where You Came From* (#gobacksbs), while in principle a reality TV show, must similarly be included here, as it thematized the highly controversial theme of asylum seeker policy in Australia. By contrast, it is notable that the hashtag for the overtly political talk show *Q&A* does not show a retweet rate that is comparable to other political backchannel cases, for reasons that remain as yet unclear. Finally, while not immediately connected to any one mainstream media channel or show, the well-established #auspol hashtag, hosting a continuous discussion of Australian political events, appears to operate much like the other hashtags within this cluster. It may be understood, therefore, as an aggregate backchannel to mainstream political news reporting in the country, rather than as collective effort to engage in gatewatching or other citizen journalism activities.

In discussing both these clusters, it is important to note that they emerge from our analysis even in spite of the widely divergent time frames for these individual hashtag cases (ranging from hours and days to close to a year) and the varying sizes of the hashtags' user bases (from less than 10,000 to more than 2 million participants). This points to the fact that these patterns of activity reflect standard uses of Twitter, which participants engage in as the theme and purpose of the hashtag demand it. It appears that the backchannel to a minor television series does not operate much differently from that for a global media event, and the response to a natural disaster does not change substantially as a greater number of people are affected.

Similarly, while (as shown earlier) the dominance of leading user groups appears to be related to the longevity of a hashtag, the activity patterns that we have observed here do not depend on the activities of that leadership group alone, as Figure 5 demonstrates. It explores the presence of any correlations between the combined contributions made by the top 10% of most active users, and the percentage of URLs in the total data set, and finds no significant connections between these metrics. A corresponding graph comparing the contributions of leading users and the percentage of retweets would similarly yield no correlations.

This is an important observation, as it shows activity patterns in a hashtag, as measured by the percentage of retweets or URLs, to be independent of the internal makeup of the hashtag community, as measured through the 1/9/90 distinction between user groups. In Figure 5, the group of backchannel hashtags at the bottom of the graph remains clearly separate from the group of acute event hashtags at the center. The same is true for the percentage of retweets in each hashtag. Leading and peripheral users may be different in many respects, but their understanding of acute events and shared audiencing experiences appears to be similar nonetheless.

CONCLUSION

What emerges from this wide-ranging comparison of participation patterns across a diverse collection of hashtag data sets is that Twitter activities, especially around defined themes and events, are far from random, but instead appear to be governed by a number of standard practices. Of these, the practices of gatewatching and audiencing are most obviously visible in our analysis, and relate clearly to the underlying themes of the hashtags we have examined. A standard response to the emergence of breaking news and other acute events is the tendency to find, share, and reshare relevant information, resulting in a high rate of URLs and retweets. By contrast, for live,

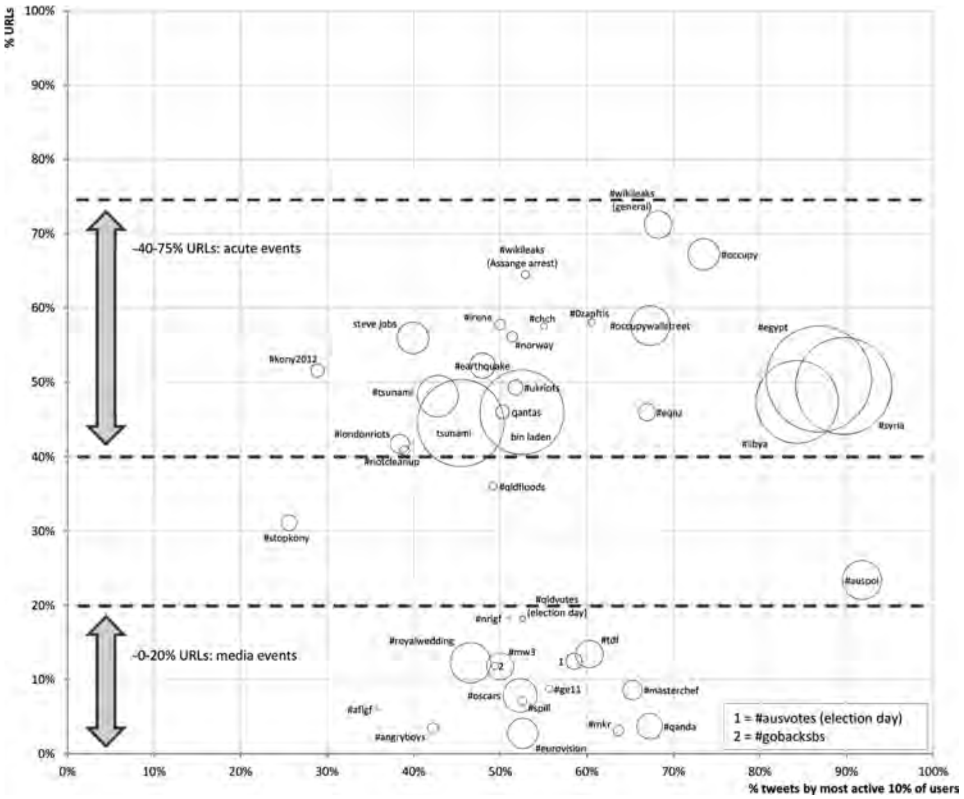


FIGURE 5 Percentage of tweets contributed by lead and highly active users vs. percentage of URLs in tweets (size of data points shows total number of tweets per hashtag).

mainstream media events Twitter acts as a backchannel, containing mainly original commentary that does not engage with the tweets of others or provide a substantial number of links to further information.

For any research dealing with Twitter data, it must be noted that due to the vagaries of working with the Twitter API itself, as well as because of unavoidable disruptions caused by regular maintenance to the university servers on which *yourTwapperkeeper* was run, the data sets thus created do not constitute an entirely comprehensive corpus of all tweets that included the specific keywords. Indeed, it is true that unless the Twitter API can be trusted to deliver all matching tweets without disruption, no study of Twitter that uses these processes can possibly achieve 100% accuracy. Further, as the API is the only access point to large-scale Twitter data which is available to researchers outside of Twitter itself, there is no opportunity to independently verify the quality of the data set. This is a necessary and unavoidable limitation that does not invalidate the findings of studies such as ours, however. Any sufficiently complex system of communication will suffer from a certain level of message loss.

Furthermore, it has to be considered that the types of topical hashtags addressed here are not the only ones that may be observed on Twitter. Further research is required to establish similar metrics for a wider range of Twitter events and to compare them with the metrics we have presented here. For example, it may well be possible that a greater number of counterculture and protest politics hashtags may exhibit similar patterns to what we have already observed for #wikileaks and #occupy, forming their own distinct cluster of cases. The intracluster distinctions we have noted for both the acute events and the backchannel cluster may also turn out to be more pronounced as more examples are added. Further, it must also be remembered that the uses of Twitter continue to evolve, especially also as a consequence of each major new event. While in combination our data sets cover a period of some 2 years, it remains to be seen whether future events will continue to show similar patterns of activity.

At the same time, if these patterns are indeed consistent across a larger number of cases and for the longer term, then our findings may also open up possibilities to operationalize them in the detection of new Twitter events. It may be possible, for example, to distinguish new acute events from other hashtags by calculating their activity metrics. This could be of use for media monitoring and emergency operations, as it would point to the emergence of crisis events purely on the basis of activity metrics, even if relevant keywords have not yet been identified. Further, if—as our examination of a handful of keyword archives appears to suggest—non-hashtagged keywords behave largely similarly to their related hashtags, this may support the identification of acute events (and their distinction from other trending topics) before users have even agreed on a standard hashtag to adopt.

Finally, our research points to the potential of understanding patterns of Twitter activity at large scale, beyond (but building on) the study of individual communicative events. Such "big data" research (boyd & Crawford 2012), drawing on comprehensive access to user activity data through platform APIs, remains in its infancy but is set to generate significant new opportunities for researchers in the humanities and allied disciplines. Current work on Twitter, such as the research presented here, will be able to be usefully combined and compared with studies of other (social) media platforms in order to develop a more comprehensive and detailed picture of information and communication flows in society. In turn, this may provide the basis for a more sophisticated understanding both of the place of social media in society, and of potential points of leverage for relevant institutions (e.g., governments, media, or emergency services) as they seek to engage with and influence such information flows.

But while this article has focused almost exclusively on the examination of large-scale, quantitative patterns in Twitter data sets of considerable size, this should not be misunderstood to privilege such quantitative research over other approaches. Rather, we close by noting the substantial opportunities for qualitative and combined quantitative/qualitative research that also exist in this field (e.g., Krüger, Stieglitz, & Potthoff, 2012). To begin with, the quantitative approaches to understanding communicative patterns on Twitter that we have introduced here provide an opportunity to pinpoint specific areas for further detailed, qualitative investigation (for a more detailed discussion, also see Bruns & Stieglitz, in press). A focus on the communicative activities of representatives of the different groups of lead, highly active, and least active users that we have introduced in this article enables an examination of a variety of distinct tweeting styles within the same hashtag exchange. For example, representatives of each group could also be studied in much greater detail through in-depth ethnographic work, or engaged through survey or interview techniques, to better understand these diverse approaches to using Twitter in specific communicative contexts.

What must be noted in this context is that the process of generating overall headline metrics for each of these data sets does not destroy the data sets themselves, which remain available for much closer, tweet-by-tweet analysis. So, for example, for data sets that follow the overall gatewatching pattern of collaborative sourcing and sharing information that appears to be common to crisis events, a possible avenue for further research is the qualitative (or mixed-methods) study of how these patterns emerge in each case and of whether these processes of emergence generally follow similar steps. Potential questions to be addressed here include how groups of lead users crystallize from the early participant base; how they come to structure their activities as the acute event unfolds; and how common principles and shared understandings of how to engage with the event are established in each case. Such questions (and similar questions that apply to the noncrisis events among the data sets we have examined here) may be addressed, inter alia, through a close, qualitative reading of the relevant tweets in the data set, through ethnographic studies of user communities, or through other methods drawn from media, cultural and communication studies, anthropology, or the social sciences. The metrics that we have outlined here, and their utilization for a bird's-eye comparison of large-scale communicative events on Twitter, are intended to serve as a useful, necessary starting point for such further research endeavors.

NOTES

1. A number of equivalent user conventions for marking messages as retweets now exist, and are included in this figure; in addition to the most common variant RT @*user* [message], we tested for MT @ *user* [message] (for "manual tweet"), [message] (via @*user*), and "@*user* [message]" (username and original message enclosed in quotation marks), which are also common, with or without added comments from the retweeting user. A second retweeting mechanism was introduced by Twitter itself, by providing a "retweet button" next to each tweet displayed on its Website or in Twitter clients; this mechanism passes along the original tweet verbatim and in full length, without inserting the "RT @*user*" into the message. Many Twitter clients—including the version of the Twitter Website optimized for mobile devices—now

offer a choice between the two formats (cf. Bruns, 2012). Because, contrary to "manual" retweets, such "button" retweets do not result in a new message but simply add to the metadata of the original tweet, they are not captured by our Twitter tracking solution and are therefore excluded from the data sets it captures. This is an unavoidable gap in the data sets, resulting in a systematic underestimation of retweeting activity. It is possible, however, to extrapolate overall retweeting activities from the patterns of "manual" retweeting.

2. We consider such protest movements, which largely remain focused on leading nations in the West, to be distinct from in the civil unrest in Libya, Egypt, or Syria, whose hashtags are located at the center of the cluster.

REFERENCES

boyd, d., & Crawford, K. (2012). *Critical questions for big data. Information, Communication and Society, 15*(5), 662–679.

boyd, d., Golder, S., & Lotan, G. (2010, January). *Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter*. Presented at HICSS-43, Kauai, Hawai'i. Retrieved from http://www.danah.org/papers/TweetTweetRetweet.pdf

Bruns, A. (2005). *Gatewatching: Collaborative online news production*. New York, NY: Peter Lang.

Bruns, A. (2012). Ad hoc innovation by users of social networks: The case of Twitter. *ZSI Discussion Paper* 16. Retrieved from https://www.zsi.at/object/publication/2186

Bruns, A., & Burgess, J. (2011a). #ausvotes: How Twitter covered the 2010 Australian federal election. *Communication, Politics & Culture, 44*(2), 37–56.

Bruns, A., & Burgess, J. (2011b, August). *The use of Twitter Hashtags in the formation of* ad hoc *publics*. Paper presented at the 6th European Consortium for Political Research General Conference, University of Iceland, Reykjavík. Retrieved from http://eprints.qut.edu.au/46515

Bruns, A., & Liang, E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday, 17*(4). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3937/3193

Bruns, A., & Stieglitz, S. (in press). Metrics for understanding communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society*. New York, NY: Peter Lang.

Bruns, A., Burgess, J., Crawford, K., & Shaw, F. (2012). *#qldfloods and @QPSMedia: Crisis communication on Twitter in the 2011 South East Queensland floods*. Brisbane: ARC Centre of Excellence for Creative Industries and Innovation. Retrieved from http://cci.edu.au/floodsreport.pdf

Bruns, A., Highfield, T., & Burgess, J. (in press). The Arab Spring and its social media audiences: English and Arabic Twitter users and their networks. *American Behavioural Scientist*.

Burgess, J., & Crawford, K. (2011, October). *Social media and the theory of the acute event*. Paper presented at Internet Research 12.0—Performance and Participation, Seattle, WA.

Cheng, J., Sun, A., Hu, D., & Zeng, D. (2011). An information diffusion-based recommendation framework for micro blogging. *Journal of the Association of Information Systems, 12*(7), 463–486.

Cheong, F., & Cheong, C. (2011). Social media data mining: A social network analysis of tweets during the 2010-2011 Australian floods. *Proceedings of the Pacific Asia Conference on Information Systems*, Paper 46. Retrieved from http://aisel.aisnet.org/pacis2011/46

Christensen, C. (2011). Twitter revolutions? Addressing social media and dissent. *Communication Review, 14(3)*, 155–157.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonalves, B., Flammini, A., & Menczer, F. (2011). Political polarization on Twitter. *Proceedings of the 5th International Conference on Weblogs and Social Media*. AAAI Press.

Deller, R. (2011). Twittering on: Audience research and participation using Twitter. *Participations*, *8*(1). Retrieved from http://www.participations.org/Volume%208/Issue%201/deller.htm

Dröge, E., Maghferat, P., Puschmann, C., Verbina, J., & Weller, K. (2011). Konferenz-Tweets. Ein Ansatz zur Analyse der Twitter-Kommunikation bei wissenschaftlichen Konferenzen. *Proceedings of the 12th International Symposium for Information Science*. Verlag Werner Hülsbusch. Retrieved from http://ynada.com/pubs/isi2010.pdf

Fiske, J. (1992). Audiencing: A cultural studies approach to watching television. *Poetics*, *21*(4), 345–359.

Hughes, A., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, *6*(3–4), 248–260.

Krishnamurthy, B., Phillipa, G., & Arlitt, M. (2008). A few chirps about Twitter. *Proceedings of the first workshop on online social networks (WOSN ´08)*. ACM publications (pp. 19–24). doi:10.1145/1397735.1397741

Krüger, N., Stieglitz, S. & Potthoff, T. (2012). Brand communication on Twitter—A case study on Adidas. *Proceedings of the 16th Pacific Asia Conference on Information Systems*, paper 161. Retrieved from http://aisel.aisnet.org/pacis2012/161

Larsson, A. O., & Moe, H. (2011). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747.

Lotan, G., Ananny, M., Gaffney, D., & boyd, d. (2011). The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, *5*, 1375–1405.

Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT? *1st Workshop on Social Media Analytics (SOMA '10)*. Washington, DC, July 25.

Palen, L., Starbird, K., Vieweg, S., & Hughes, A. (2010). Twitter-based information distribution during the 2009 Red River Valley flood threat. *Bulletin of the American Society for Information Science and Technology*, *36*(5), 13–17.

Papacharissi, Z. (2011). Conclusion: A networked self. In Z. Papacharissi (Ed.), *A networked self: Identity, community and culture on social network sites* (pp. 304–318). New York, NY: Routledge.

Park, J., Cha, M., Kim, H., & Jeong, J. (2011). Managing bad news in social media: A case study on Domino's Pizza crisis. *Proceedings of the ICWSM 2012*. AAAI

Press. Retrieved from www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4672

Stieglitz, S., & Dang-Xuan, L. (2012). Social media and political communication—A social media analytics framework. *Social Network Analysis and Mining*. doi 10.1007/s13278-012-0079-3

Stieglitz, S., & Dang-Xuan, L. (in press). Emotions and information diffusion in social media—An investigation of sentiment of microblogs and sharing behavior. *Journal of Management Information Systems.*

Stieglitz, S., & Krüger, N. (2011). Analysis of sentiments in corporate Twitter communication—A case study on an issue of Toyota. *Proceedings of the 22nd Australasian Conference on Information Systems*, paper 29. Retrieved from http://aisel.aisnet.org/acis2011/29

Suh, B., Hong, L., Pirolli, P., & Chi, E. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. *Proceedings of the IEEE International Conference on Social Computing*, IEEE Computer Society (pp. 177–184). Retrieved from http://www.parc.com/content/attachments/want-to-be-retweeted.pdf

Tedjamulia, S. J. J., Dean, D. L., Olsen, D. R., & Albrecht, C. C. (2005). Motivating content contributions to online communities: Toward a more comprehensive theory. *Proceedings of the 38th Annual Hawaii International Conference on System Science*, paper 193b. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1385630&tag=1

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, L. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*(4), 402–418.

Twitter. (2012, March). *Twitter turns six*. Retrieved from http://blog.twitter.com/2012/03/twitter-turns-six.html

Vis, F. (2012, 24 Jan.). Reading the riots on Twitter: Who tweeted the riots? *Researching Social Media*. Retrieved from http://researchingsocialmedia.org/2012/01/24/reading-the-riots-on-twitter-who-tweeted-the-riots

Weller, K., Dröge, E., & Puschmann, C. (2011). *Citation analysis in Twitter: Approaches for defining and measuring information flows within tweets during scientific conferences*. #MSM2011—1st Workshop on Making Sense of Microposts, Heraklion, Greece, May 30. Retrieved from http://files.ynada.com/papers/msm2011.pdf

Yardi, S., & boyd, d. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology and Society*, *20*, 1–8.

YourTwapperkeeper. (2012). Retrieved from https://github.com/jobrieniii/yourTwapperKeeper

ABOUT THE AUTHORS

Dr. Axel Bruns is an Associate Professor in the Creative Industries Faculty at Queensland University of Technology in Brisbane, Australia, and a Chief Investigator in the ARC Centre of Excellence for Creative Industries and Innovation (http://cci.edu.au/). He is the author of *Blogs, Wikipedia, Second Life and Beyond: From Production to Produsage* (2008) and *Gatewatching:*

Collaborative Online News Production (2005), and a co-editor of *A Companion to New Media Dynamics* (2012, with John Hartley and Jean Burgess) and *Uses of Blogs* (2006, with Joanne Jacobs). Bruns is an expert on the impact of user-led content creation, or produsage, and his current work focuses especially on the study of user participation in social media spaces such as *Twitter*, especially in the context of acute events. His research blog is at http://snurb.info/, and he tweets at @snurb_dot_info. See http://mappingonlinepublics.net/ for more details on his current social media research.

Stefan Stieglitz is an Assistant Professor in communication and collaboration management at the Department of Information Systems at the University of Münster in Germany. He studied business economics and received his doctorate degree at University of Potsdam in Germany. He is founder and academic director of the Competence Center Smarter Work at the European Research Center for Information Systems (ERCIS). Of particular interest in his work is to investigate the usage of social media in enterprises and political context. His research focuses on economic, social, and technological aspects of collaboration software. He published several articles in reputable international journals such as *Journal of Management Information Systems*, *MIS Quarterly Executive,* and *International Journal of Social Research Methodology.*

Predicting Tie Strength With Social Media

Eric Gilbert and Karrie Karahalios
University of Illinois at Urbana-Champaign
[egilber2, kkarahal]@cs.uiuc.edu

ABSTRACT

Social media treats all users the same: trusted friend or total stranger, with little or nothing in between. In reality, relationships fall everywhere along this spectrum, a topic social science has investigated for decades under the theme of *tie strength*. Our work bridges this gap between theory and practice. In this paper, we present a predictive model that maps social media data to tie strength. The model builds on a dataset of over 2,000 social media ties and performs quite well, distinguishing between *strong* and *weak* ties with over 85% accuracy. We complement these quantitative findings with interviews that unpack the relationships we could not predict. The paper concludes by illustrating how modeling tie strength can improve social media design elements, including privacy controls, message routing, friend introductions and information prioritization.

Author Keywords

Social media, social networks, relationship modeling, ties, sns, tie strength

ACM Classification Keywords

H5.3. Group and Organization Interfaces; Asynchronous interaction; Web-based interaction.

INTRODUCTION

Relationships make social media *social*. Yet, different relationships play different roles. Consider the recent practice of substituting social media friends for traditional job references. As one hiring manager remarked, by using social media "you've opened up your rolodex for the whole world to see" [38]. To the dismay of applicants, employers sometimes cold call social media friends expecting a job reference "only to find that you were just drinking buddies." Although clearly not the norm, the story illustrates a basic fact: not all relationships are created equal.

For decades, social science has made much the same case, documenting how different types of relationships impact individuals and organizations [16]. In this line of research, relationships are measured in the currency of *tie strength* [17]. Loose acquaintances, known as *weak ties*, can help a friend generate creative ideas [4] or find a job [18]. They also expedite the transfer of knowledge across workgroups [20]. Trusted friends and family, called *strong ties*, can affect emotional health [36] and often join together to lead organizations through times of crisis [24]. Despite many compelling findings along this line of research, social media does not incorporate tie strength or its lessons. Instead, all users are the same: friend or stranger, with little or nothing in between. Most empirical work examining large-scale social phenomena follows suit. A link between actors either exists or not, with the relationship having few properties of its own [1, 2, 27].

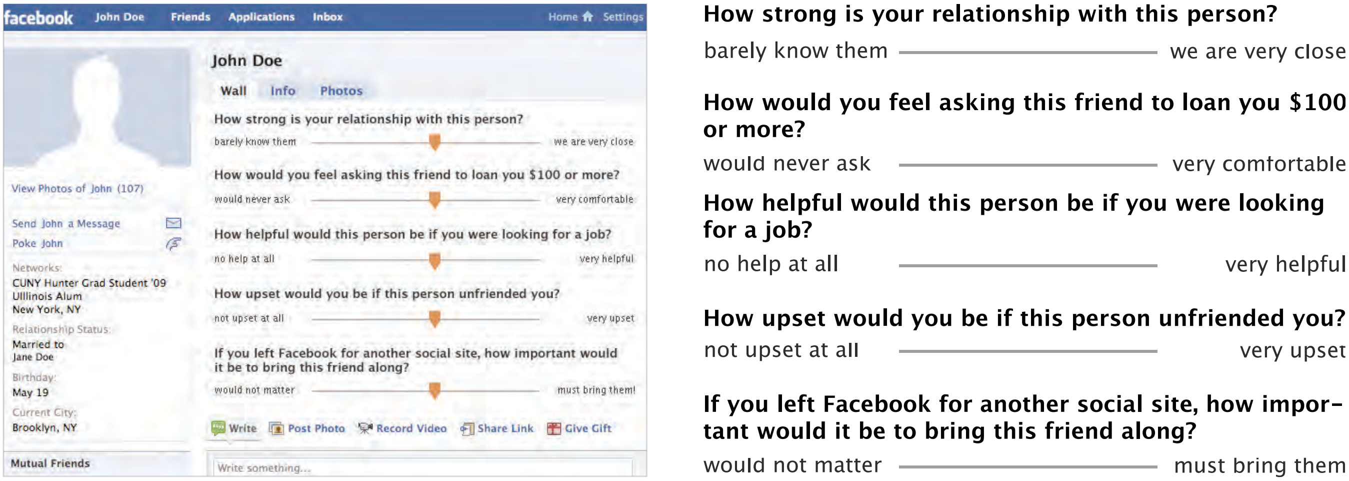
This paper aims to bridge the gap, merging the theory behind tie strength with the data behind social media. We address one central question. With theory as a guide, can social media data predict tie strength? This is more than a methodological or theoretical point; a model of tie strength has the potential to significantly impact social media users. Consider automatically allowing the friends of strong ties to access your profile. Or, as one participant cleverly suggested, remaking Facebook's Newsfeed to get rid of "people from high school I don't give a crap about." The model we present builds on a dataset of over 2,000 Facebook friendships, each assessed for tie strength and described by more than 70 numeric indicators. It performs with surprising accuracy, modeling tie strength to 10-point resolution and correctly classifying friends as *strong* or *weak* ties more than 85% of the time.

We begin by reviewing the principles behind tie strength, and then discuss its proposed dimensions. Using theory to guide the selection of predictive variables, we next present the construction of our tie strength model. It performs well, but not perfectly. To understand our model's limitations, we also present the results of follow-up interviews about the friendships we had the most difficulty predicting. The paper concludes by applying our findings toward implications for theory and practice.

TIE STRENGTH

Mark Granovetter introduced the concept of *tie strength* in his landmark 1973 paper "The Strength of Weak Ties" [17]. In this section we review *tie strength* and the substantial line of research into its characteristics. We then discuss four researchers' proposals for the dimensions of tie strength, laying a foundation for our treatment of it as a predictable quantity. The section concludes by introducing the research questions that guide the rest of this paper.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.



John Doe

How strong is your relationship with this person?
barely know them — we are very close

How would you feel asking this friend to loan you $100 or more?
would never ask — very comfortable

How helpful would this person be if you were looking for a job?
no help at all — very helpful

How upset would you be if this person unfriended you?
not upset at all — very upset

If you left Facebook for another social site, how important would it be to bring this friend along?
would not matter — must bring them!

Networks:
CUNY Hunter Grad Student '09
UIllinois Alum
New York, NY

Relationship Status:
Married to
Jane Doe

Birthday:
May 19

Current City:
Brooklyn, NY

How strong is your relationship with this person?

barely know them — we are very close

How would you feel asking this friend to loan you $100 or more?

would never ask — very comfortable

How helpful would this person be if you were looking for a job?

no help at all — very helpful

How upset would you be if this person unfriended you?

not upset at all — very upset

If you left Facebook for another social site, how important would it be to bring this friend along?

would not matter — must bring them

Figure 1. The questions used to assess tie strength, embedded into a friend's profile as participants experienced them. An automated script guided participants through a random subset of their Facebook friends. As participants answered each question by dragging a slider, the script collected data describing the friendship. The questions reflect a diversity of views on tie strength.

Definition and Impact

> The strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie. [17]

While Granovetter left the precise definition of tie strength to future work, he did characterize two types of ties, *strong* and *weak.* Strong ties are the people you really trust, people whose social circles tightly overlap with your own. Often, they are also the people most like you. The young, the highly educated and the metropolitan tend to have diverse networks of strong ties [31]. Weak ties, conversely, are merely acquaintances. Weak ties often provide access to novel information, information not circulating in the closely knit network of strong ties.

Many researchers have adopted tie strength as an analytic framework for studying individuals and organizations [16]. (Google Scholar, for instance, claims that over 7,000 papers cite "The Strength of Weak Ties" [15].) The social support offered by strong ties can actually improve mental health [36]. Banks that find the right mix of weak and strong ties to other firms tend to get better financial deals [39]. It has also been shown that weak ties, as opposed to strong ones, benefit job-seekers [18]. However, socioeconomic class reverses this effect: job-seekers from lower socioeconomic backgrounds often rely heavily on strong ties [16].

Strong ties between employees from different organizational subunits can help an organization withstand a time of crisis [24]. Yet, strongly tied coworkers are also the ones likely to create crises by pushing for institutional change [23]. Employees who weakly tie themselves beyond organizational boundaries tend to receive better performance reviews and generate more creative ideas [4]. Weak ties also act as a conduit for useful information in computer-mediated communication [8]. However, weak ties often rely on a few commonly available media [22], whereas strong ties diversify, communicating through many channels [21].

The Dimensions of Tie Strength

> At what point is a tie to be considered weak? This is not simply a question for the methodologically curious … the theory makes a curvilinear prediction. How do we know where we are on this theoretical curve? Do all four indicators count equally toward tie strength? [23]

Granovetter proposed four tie strength dimensions: *amount of time, intimacy, intensity* and *reciprocal services.* Subsequent research has expanded the list. Ronald Burt proposed that structural factors shape tie strength, factors like network topology and informal social circles [5]. Wellman and Wortley argue that providing emotional support, such as offering advice on family problems, indicates a stronger tie [40]. Nan Lin, et al., show that social distance, embodied by factors such as socioeconomic status, education level, political affiliation, race and gender, influences tie strength [29].

In theory, tie strength has at least seven dimensions and many manifestations. In practice, relatively simple proxies have substituted for it: communication reciprocity [11], possessing at least one mutual friend [37], recency of communication [28] and interaction frequency [13, 17]. In a 1984 study, Peter Marsden used survey data from three metropolitan areas to precisely unpack the predictors of tie strength [33]. While quite useful, Marsden pointed out a key limitation of his work: the survey asked participants to recall only their three closest friends along with less than ten characteristics of the friendship.

The present research can be seen as updating Marsden's work for the era of social media. Our work differs primarily in setting and scale. By leveraging social media, participants no longer have to recall; we can take advantage of long friend lists and rich interaction histories. In this way, our work also overcomes the problem of retrospective informant accuracy [3, 30, 32]. In addition, a tie strength model built from social media has the potential to feed back into social media, in ways that benefit its users.

Research Questions

The work above leads us to introduce two research questions that guide the remainder of this paper:

R1: The existing literature suggests seven dimensions of tie strength: *Intensity, Intimacy, Duration, Reciprocal Services, Structural, Emotional Support* and *Social Distance.* As manifested in social media, can these dimensions predict tie strength? In what combination?

R2: What are the limitations of a tie strength model based solely on social media?

METHOD

To answer our research questions, we recruited 35 participants to rate the strength of their Facebook friendships. Our goal was to collect data about the friendships that could act, in some combination, as a predictor for tie strength. Working in our lab, we used the Firefox extension Greasemonkey [19] to guide participants through a randomly selected subset of their Facebook friends. (Randomly sampling participants' friends guards against those with large networks dominating the results.) The Greasemonkey script injected five tie strength questions into each friend's profile after the page loaded in the browser. Figure 1 shows how a profile appeared to a participant. Participants answered the questions for as many friends as possible during one 30-minute session. On average, participants rated 62.4 friends ($\sigma = 16.2$), resulting in a dataset of 2,184 rated Facebook friendships.

Social media experiments often employ completely automated data collection. We worked in the lab for two important reasons. First, we captured all data at the client side, after a page loaded at the user's request. This allowed us to stay within Facebook's Terms of Service. More importantly, however, we asked participants to give us sensitive information: their relationship strengths plus personal Facebook data. We collected data in the lab to guard our participants' privacy and to increase the accuracy of their responses.

Predictive Variables

While participants responded to the tie strength questions, our script automatically collected data about the participant, the friend and their interaction history. The tie strength literature reviewed in the previous section pointed to seven major dimensions of predictive variables. With these dimensions as a guide, we identified 74 Facebook variables as potential predictors of tie strength. Table 1 presents 32 of these variables along with their distributions. In choosing these predictive variables, we tried to take advantage of Facebook's breadth while simultaneously selecting variables that could carry over to other social media. Below, we clarify some variables listed in Table 1 and present those not included in the table. All predictive variables make an appearance either in the text or in Table 1.

Intensity Variables

Each Facebook user has a Wall, a public communication channel often only accessible to a user's friends. *Wall words exchanged* refers to the total number of words traded between the participant and the friend via Wall posting. *Inbox messages exchanged* counts the number of appearances by a friend in a participant's Facebook Inbox, a private commu-

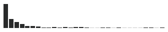
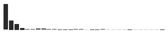
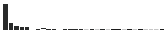
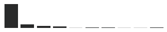
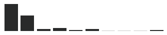
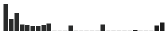
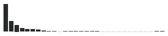
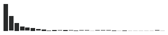
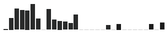
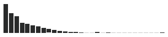
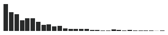
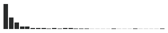
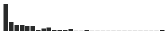
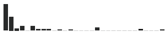
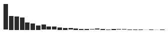
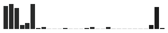
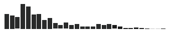
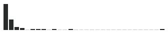
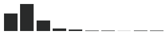
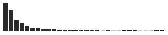
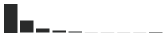
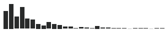
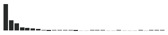
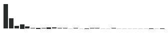
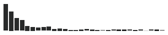
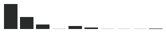
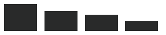
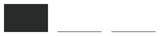
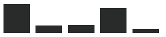
Predictive Intensity Variables	**Distribution**	**Max**
Wall words exchanged		9549
Participant-initiated wall posts		55
Friend-initiated wall posts		47
Inbox messages exchanged		9
Inbox thread depth		31
Participant's status updates		80
Friend's status updates		200
Friend's photo comments		1352
Intimacy Variables		
Participant's number of friends		729
Friend's number of friends		2050
Days since last communication		1115
Wall intimacy words		148
Inbox intimacy words		137
Appearances together in photo		73
Participant's appearances in photo		897
Distance between hometowns (mi)		8182
Friend's relationship status	6% engaged 32% married 30% single 30% in relationship	
Duration Variable		
Days since first communication		1328
Reciprocal Services Variables		
Links exchanged by wall post		688
Applications in common		18
Structural Variables		
Number of mutual friends		206
Groups in common		12
Norm. TF-IDF of *interests* and *about*		73
Emotional Support Variables		
Wall & inbox positive emotion words		197
Wall & inbox negative emotion words		51
Social Distance Variables		
Age difference (days)		5995
Number of occupations difference		8
Educational difference (degrees)		3
Overlapping words in *religion*		2
Political difference (scale)		4

Table 1. Thirty-two of over seventy variables used to predict tie strength, collected for each of the 2,184 friendships in our dataset. The distributions accompanying each variable begin at zero and end at the adjacent maximum. Most variables are not normally distributed. The *Predictive Variables* subsection expands on some of these variables and presents those not included in this table.

nication channel. *Inbox thread depth*, on the other hand, captures the number of individual Inbox messages sent between the pair. A helpful analogy for *Inbox thread depth* is the number of messages in a newsgroup thread.

Intimacy Variables

To complement our aggregate measures, we used the Linguistic Inquiry and Word Count (LIWC) dictionary to perform content analysis [34]. Our hypothesis was that friends of different tie strengths would use different types of words when communicating. LIWC matches text against lists of word stems assembled into categories. *Wall intimacy words* refers to the number of Wall words matching at least one of eleven LIWC categories: Family, Friends, Home, Sexual, Swears, Work, Leisure, Money, Body, Religion and Health. Similarly, *Inbox intimacy words* refers to the number of Inbox words matching at least one of these categories. The Home category, for example, includes words like *backyard* and *roommate*, while the Work category includes *busy*, *classes* and *commute.* In total, the intimacy variables checked for matches against 1,635 word stems. Although not presented in Table 1, we also included each LIWC intimacy category as its own predictive variable.

Days since last communication measures the recency of written communication in some Facebook channel (Wall, Inbox, photo comments) from the day we collected data.

Duration Variable

We did not have access to the date when two people became friends. Instead, *Days since first communication* is a proxy for the length of the friendship. It measures time in the same way as *Days since last communication.*

Reciprocal Services Variables

Facebook friends have relatively few opportunities to exchange informational, social or economic goods. (These practices clearly differ by social media; consider a LinkedIn user who exploits his social capital by introducing business contacts to one another.) To capture *Reciprocal Services* on Facebook, *Links exchanged by wall post* measures the number of URLs passed between friends via the Wall, a common Facebook practice. Similarly, *Applications in common* refers to the number of Facebook applications a participant and friend share. Facebook applications usually provide a tightly scoped service (e.g., displaying a virtual bookshelf on a profile) and often spread between friends by word of mouth.

Structural Variables

Facebook allows users to join groups organized around specific topics and interests. *Groups in common* refers to the number of Facebook groups to which both the participant and the friend belong. *Normalized TF-IDF of interests and about* measures the similarity between the free text *interests* and *about* profile fields. It does so by computing the dot product between the TF-IDF vectors representing the text. TF-IDF is a standard information retrieval technique [10] that respects the baseline frequencies of different words in the English language. We also measured *Number of overlapping networks,* the number of Facebook networks to which both the participant and the friend belong. Facebook networks often map to universities, companies and geographic areas.

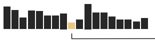
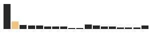
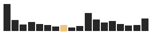
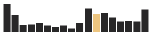
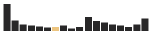
Dependent Variables	0–1 Scale	Median
How strong is your relationship?		0.411
How comfortable asking for loan?		0.076
How helpful if looking for job?		0.362
How upset if unfriended?		0.552
How important to bring friend?		0.324

Table 2. The five questions used to assess tie strength, accompanied by their distributions. The distributions present participant responses mapped onto a continuous 0–1 scale. Our model predicts these responses as a function of the variables presented in Table 1.

Emotional Support Variables

In a way similar to the content analysis variables described above, *Wall & inbox positive emotion words* is two variables referring to matches against the LIWC category Positive Emotion. The Positive Emotion category includes words like *birthday, congrats* and *sweetheart.* Similarly, *Wall & inbox negative emotion words* is two variables counting matches in the Negative Emotion category, including words like *dump*, *hate* and *useless*. We also recorded the number of gifts given between a participant and a friend. A Facebook gift is a small icon often given to a friend to show support. Gifts sometimes cost a small amount of money.

Social Distance Variables

We measured the difference in formal education between a participant and a friend in terms of academic degrees. It is computed by searching for the letters *BS*, *BA*, *MS*, *MA*, *JD*, *MD* and *PhD* in the *education* profile field. *Educational difference* measures the numeric difference between a participant and a friend along a scale: 0: None, 1: BS/BA, 2: MS/MA, 3: JD/MD/PhD.

1,261 people in our dataset completed the *politics* profile field. Of those, 79% reported their political affiliation as *very conservative*, *conservative*, *moderate*, *liberal* or *very liberal.* Assigning a scale in that order, *Political difference* measures the numeric difference between a participant and a friend. While the education and politics scales do not completely reflect the diversity of our sample, they do provide useful tools for assessing the importance of these variables for the majority of it.

Demographic and Usage Variables

Finally, in addition to the variables described above, we collected demographic and usage information on our participants and their friends: gender, number of applications installed, number of inbox messages, number of wall posts and number of photo comments.

Dependent Variables

Previous literature has proposed various manifestations of tie strength [17, 18, 21, 24]. To capture a diversity of views, we asked our participants to answer five tie strength questions. Participants moved a slider along a continuum to rate a friend. Figure 1 shows how those questions were embedded into a friend’s profile. Table 2 illustrates the responses. We chose a continuum instead of a discrete scale for three reasons. First, Mark Granovetter conjectured that tie strength may in fact be continuous [17]. The literature has not resolved the issue, let alone specified how many discrete tie strength levels exist. A continuum bypasses that problem. Second, a continuum lends itself to standard modeling techniques. Finally, applications can round a continuous model’s predictions to discrete levels as appropriate.

Participants

Our 35 participants, primarily students and staff from the University of Illinois community, came from more than 15 different academic departments. The sample consisted of 23 women (66%) and 12 men (34%) ranging between 21 and 41 years old, with a mean and median of 26. The minimum number of Facebook friends was 25; the maximum was 729 (median of 153). In terms of age and number of friends, previous empirical work suggests that our participants fall within the mainstream of Facebook users [14, 35]. All participants used Facebook regularly and had been members for at least one year.

Statistical Methods

We modeled tie strength as a linear combination of the predictive variables, plus terms for dimension interactions and network structure:

$$s_i = \alpha + \beta R_i + \gamma D_i + N(i) + \epsilon_i$$

$$N(i) = \lambda_0 \mu_M + \lambda_1 med_M + \sum_{k=2}^{4} \sum_{s \in M} \lambda_k (s - \mu_M)^k + \lambda_5 min_M + \lambda_6 max_M$$

$$M = \{s_j : j \text{ and } i \text{ are mutual friends}\}$$

More complex models were explored, but a (mostly) linear model allows us to take advantage of the full dataset and explain the results once it is built. In the equations above, s_i represents the tie strength of the i^{th} friend. R_i stands for the vector of 67 individual predictive variables. ε_i is the error term. D_i represents the pairwise interactions between the dimensions presented in Table 1. Pairwise interactions are commonly included in predictive models [12]; in this case, including all pairwise interactions would force more variables than data points into the model. Instead, we nominated variables with the fewest missing values to represent each dimension. (Not every participant or friend contributes every variable.) D_i represents all pairwise interactions between the 13 variables with a 90% or greater completion rate. Choosing 90% as a threshold ensured that every dimension was represented. To the best of our knowledge, exploring the interactions between the dimensions of tie strength is a novel approach.

N(i) encodes network structure. It captures the idea that a friendship’s tie strength not only depends on its history, but also on the tie strengths of mutual friends. In other words, it models the idea that a friend who associates with your business acquaintances is different than one who knows your mother, brother and sister. Since every friend has a potentially unique set of mutual friends, the model uses seven descriptors of the tie strength distribution over mutual friends: mean, median, variance, skew, kurtosis, minimum and maximum. These terms belong to the *Structural* dimension. However, *N(i)* introduces a dependency: every tie strength now depends on other tie strengths. How can we incorporate the tie strengths of mutual friends when it is tie strength we want to model in the first place? To solve this problem, we fit the equations above using an iterative variation of OLS regression. In each iteration, the tie strengths from the previous round are substituted to calculate *N(i)*, with all s_i initially set to zero. (Note that *N(i)* is mostly linear in the predictive variables.) Using this procedure, all s_i converged in nine iterations (0.001 average relative change threshold). This approach parallels other “neighborhood effect” models [6].

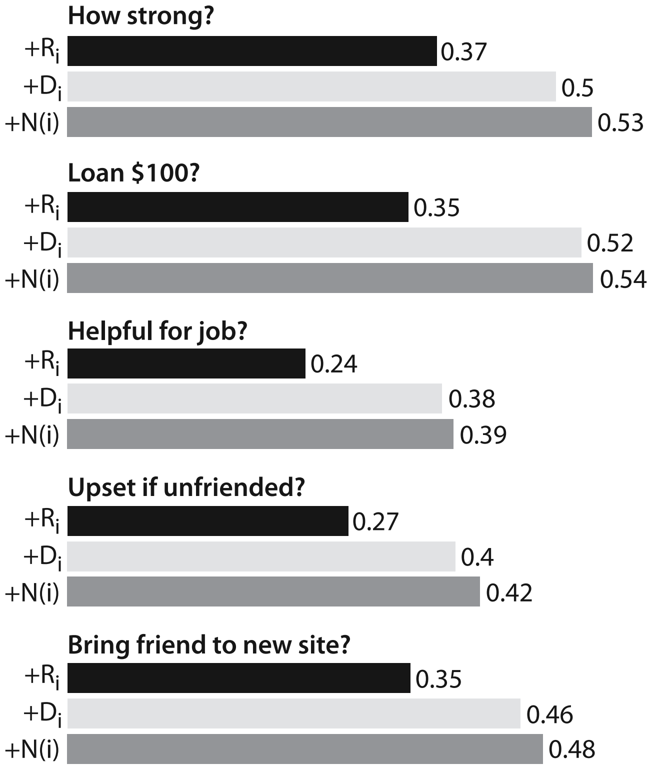


Figure 2. The model’s Adjusted R^2 values for all five dependent variables, broken down by the model’s three main terms. Modeling interactions between tie strength dimensions results in a substantial performance boost. The model performs best on *Loan $100?* and *How strong?*, the most general question.

We did not standardize, or “ipsatize” [9], the dependent variables. Because we employed network subsampling, we could not be sure participants saw the Facebook friend they would rate highest or lowest. Furthermore, not all real-life friends have Facebook accounts. It is reasonable to assume that some participants would reserve the ends of the spectra for people our experiment would never turn up. Finally, to account for the violations of normality exhibited by the distributions in Table 1, every variable is log-transformed.

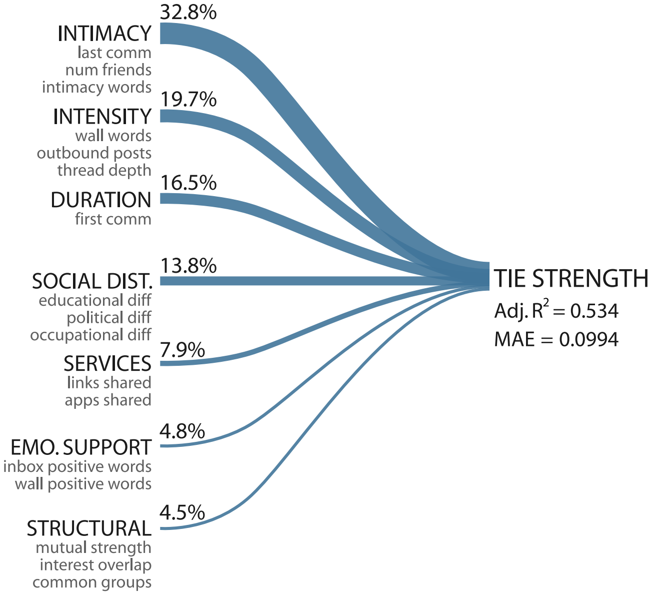


Figure 3. The predictive power of the seven tie strength dimensions, presented here as part of the *How strong?* model. A dimension's weight is computed by summing the absolute values of the coefficients belonging to it. The diagram also lists the top three predictive variables for each dimension. On average, the model predicts tie strength within one-tenth of its true value on a continuous 0–1 scale.

RESULTS

Because each participant rated more than one friend, observations within a participant were not independent. This is a common obstacle for ego-centric designs. To roughly adjust for it, all of the results presented here cut the degrees of freedom in half, a technique borrowed from the social networks literature [33].

On the first tie strength question, *How strong is your relationship with this person?*, the model fits the data very well: *Adj.* $R^2 = 0.534$, $p < 0.001$. It achieves a Mean Absolute Error of 0.0994 on a continuous 0–1 scale, where 0 is weakest and 1 is strongest. In other words, on average the model predicts tie strength within one-tenth of its true value. This error interval tightens near the ends of the continuum because predictions are capped between 0 and 1. In addition, we found strong evidence of four dimension interactions ($p < 0.001$): *Intimacy* × *Structural*, $F_{1,971} = 12.37$; *Social Distance* × *Structural*, $F_{1,971} = 34$; *Reciprocal Services* × *Reciprocal Services*, $F_{1,971} = 14.4$; *Structural* × *Structural*, $F_{1,971} = 12.41$. As we demonstrate shortly, the *Structural* dimension plays a minor role as a linear factor. However, it has an important modulating role via these interactions. One way to read this result is that individual relationships matter, but they get filtered through a friend's clique before impacting tie strength.

Figure 2 summarizes the model's performance on all five tie strength questions, broken down by the model's three main terms. Modeling dimension interactions boosts performance significantly, with smaller gains associated with modeling network structure. The model fits the second tie strength question as well as the first: *How would you feel asking this friend to loan you $100 or more?* However, it does not fit the last three questions as well. The lower performance on these questions may have resulted from participant fatigue. We considered randomizing the questions for each friend to account for ordering effects like fatigue, but we feared that randomizing would confuse and frustrate our participants, contributing to lower accuracy across the board. Therefore, we chose to prioritize the first question, the most general of the five. With the exception of *How helpful would this person be if you were looking for a job?,* all dependent variable intercorrelations were above 0.5 (Table 4).

Top 15 Predictive Variables	β	F	p-value
Days since last communication	-0.76	453	< **0.001**
Days since first communication	0.755	7.55	< **0.001**
Intimacy × Structural	0.4	12.37	< **0.001**
Wall words exchanged	0.299	11.51	< **0.001**
Mean strength of mutual friends	0.257	188.2	< **0.001**
Educational difference	-0.22	29.72	< **0.001**
Structural × Structural	0.195	12.41	< **0.001**
Reciprocal Serv. × Reciprocal Serv.	-0.19	14.4	< **0.001**
Participant-initiated wall posts	0.146	119.7	< **0.001**
Inbox thread depth	-0.14	1.09	0.29
Participant's number of friends	-0.14	30.34	< **0.001**
Inbox positive emotion words	0.135	3.64	**0.05**
Social Distance × Structural	0.13	34	< **0.001**
Participant's number of apps	-0.12	2.32	0.12
Wall intimacy words	0.111	18.15	< **0.001**

Table 3. The fifteen predictive variables with highest standardized beta coefficients. The two *Days since* variables have large coefficients because of the difference between never communicating and communicating once. The utility distribution of the predictive variables forms a power-law distribution: with only these fifteen variables, the model has over half of the information it needs to predict tie strength.

Figure 3 visualizes the predictive power of the seven tie strength dimensions as part of the *How strong?* model. The figure also includes each dimension's top three contributing variables. The weight of a dimension is calculated by summing the coefficients of the the variables belonging to it. Although not uniformly distributed, no one dimension has a monopoly on tie strength.

Table 3 presents the standardized beta coefficients of the top fifteen predictive variables. The *F* statistics signify a variable's importance in the presence of the other variables. The two *Days since* variables have such high coefficients due to friends that never communicated via Facebook. Those observations were assigned outlying values: zero in one case and twice the maximum in the other. In other words, the simple act of communicating once leads to a

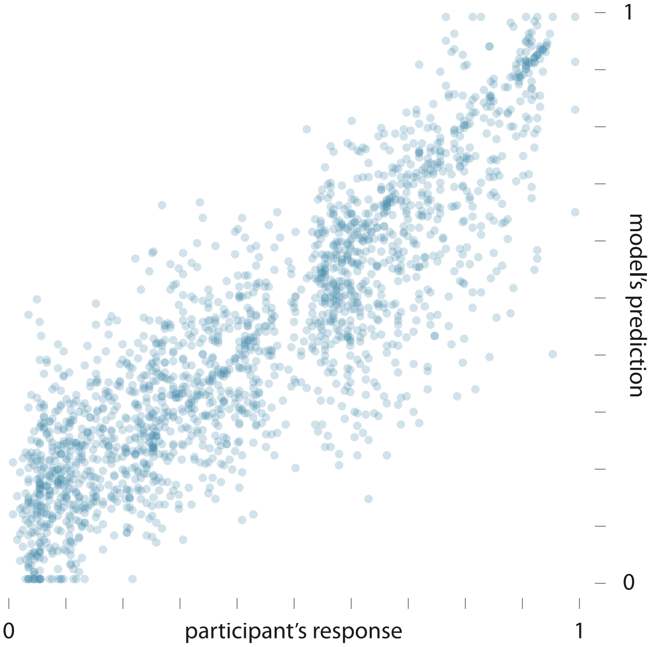


Figure 4. The model's performance across all ties in our dataset. There is a strong correlation, yet the model shows a slight bias toward underestimation, represented as the larger cloud in the bottom-right of the figure. The gap in the center results from participants' inclination to move the slider from its starting point, if only slightly.

very large movement in tie strength. *Educational difference* plays a large role in determining tie strength, but that may reflect the university community from which we sampled participants. Curiously, *Inbox thread depth* has a negative effect on tie strength; the more messages friends exchange on a single topic, the lower their tie strength. It is important to note that Table 3 orders the variables by their weights, or β coefficients, not their p-values. The p-value for *Inbox thread depth* does not express confidence in its coefficient; it expresses confidence in its utility relative to other variables. (The coefficient confidence is greater than 99.9%.) For example, *Inbox thread depth* is highly correlated with *Inbox intimacy words,* resulting in a lower F statistic.

Figure 4 compares the model's prediction to participant responses across the entire dataset. The figure illustrates a strong correlation and another view on the MAE presented above. We discuss the practical significance of the findings illustrated in Figure 4, along with the discretization of tie strength, in the next section.

Error Analysis Interviews

The model performs well, but not perfectly. To understand its limitations, we conducted ten follow-up interviews about the friendships we had the most difficulty predicting. After identifying the friends with the highest residuals, we asked participants to tell us about this particular friendship, including anything that makes it special. For instance, one participant described a "friend" he barely knew:

> I don't know why he friended me. But I'm easy on Facebook, because I feel like I'm somehow building (at least a miniscule amount of) social capital, even when I don't know the person. We went to the same high school and have a few dozen common friends. We've never interacted with each other on Facebook aside from the friending.
>
> **rating: 0; prediction: 0.44**

Correlations	Strong	Loan	Job	Un	Bring
Strong	1	0.69	0.45	0.75	0.7
Loan	0.69	1	0.4	0.55	0.55
Job	0.45	0.4	1	0.5	0.46
Unfriend	0.75	0.55	0.5	1	0.74
Bring	0.7	0.55	0.46	0.74	1

Table 4. The intercorrelations of the five dependent variables. With the exception of *Job-Strong*, *Job-Loan* and *Bring-Job*, the dependent variables are well-correlated with one another.

Notice how the participant recalls that "he friended me." Although these friends had communicated via Facebook only twice (the participant mistakenly recalled "never"), the friend's clique confused the model. The friend came from a group of relatively strong friends. As we mentioned earlier, the model filters individual relationships through cliques, leading to the high residual. Perhaps having deeper network knowledge could help, such as how the mutual friends see this friend. But this is beyond our ego-centric design.

Asymmetric Friendships

Two participants rated a friend highly because of how the friendship compared to others like it. In one case, a participant described a close bond with a professor:

> This is a professor from one of the classes I TA-ed. We have a very good relationship, because in the past we have worked out a lot of difficult class problems. The professor still remembers my name, which for some of my "friends" on Facebook may not be true. But not only that, she also knows how things are going at school, and when we meet in a hallway we usually stop for a little chat, rather then exchanging casual "Hi! Hello!" conversation.
>
> **rating: 0.85; prediction: 0.41**

Educational difference and the directionality of the wall posts pushed this prediction toward *weak tie*. Many people would not remark that a close friend "remembers my name." However, in the context of this participant's "networking" friends, the professor breaks the mold.

Participants' responses often revealed the complexity of real-life relationships, both online and offline. One participant grounded her rating not in the present, but in the hope of reigniting a friendship:

> Ah yes. This friend is an old ex. We haven't really spoken to each other in about 6 years, but we ended up friending each other on Facebook when I first joined. But he's still important to me. We were best friends for seven years before we dated. So I rated it where I did (I was actually even thinking of rating it higher) because I am optimistically hoping we'll recover some of our "best friend"-ness after a while. Hasn't happened yet, though.
>
> **rating: 0.6; prediction: 0.11**

Confounding the Medium

As might be expected, Facebook friends do not only stick to Facebook. One participant described a close friendship with a diverse digital trail:

> This friend is very special. He and I attended the same high school, we interacted a lot over 3 years and we are very very close. We trust each other. My friend are I are still interacting in ways other than Facebook such as IM, emails, phones. Unfortunately, that friend and I rarely interact through Facebook so I guess your predictor doesn't have enough information to be accurate.
>
> **rating: 0.96; prediction: 0.47**

However, even friends that stick to Facebook sometimes do so in unexpected ways:

> We were neighbors for a few years. I babysat her child multiple times. She comes over for parties. I'm pissed off at her right now, but it's still 0.8. ;) Her little son, now 3, also has an account on Facebook. We usually communicate with each other on Facebook via her son's account. This is our "1 mutual friend."
>
> **rating: 0.8; prediction: 0.28**

This playful use of Facebook clearly confused our model. With the exception of the *Social Distance* dimension, all indicators pointed to a weak tie. In fact, it is hard to imagine a system that could ever (or should ever) pick up on scenarios like this one.

DISCUSSION

Our results show that social media can predict tie strength. The *How strong?* model predicts tie strength within one-tenth of its true value on a continuous 0–1 scale, a resolution probably acceptable for most applications. In other words, discretizing our continuum onto a 10-point Likert scale, the *How strong?* model would usually miss by at most one point. The *Intimacy* dimension makes the greatest contribution to tie strength, accounting for 32.8% of the model's predictive capacity. This parallels Marsden's finding that emotional closeness best reflects tie strength [33]. However, the *Intensity* dimension also contributes substantially to the model, contrasting with Marsden's finding that *Intensity* has significant drawbacks as a predictor. One way to explain this discrepancy is that the sheer number of people available through social media strengthens *Intensity* as a predictor. In other words, when you choose to interact with someone over and over despite hundreds of people from which to choose, it significantly informs tie strength. The number of variables representing each dimension also plays a role in its overall impact. For example, *Emotional Support* might impact tie strength more if more variables represented it. (*Emotional Support* is particularly hard to quantify.) However, more variables does not always equal greater impact. As *Duration* illustrates, a single variable can account for a large part of the model's predictive capacity.

Some applications will not need 10-point resolution; the coarse categories of *strong* and *weak* may suffice. In "The Strength of Weak Ties," Granovetter himself performs his analytic work with only these approximate distinctions. One way to accomplish this is to use the model's mean, classifying all friends above it as *strong* and all below it as *weak*. Correct predictions are those where the participant's rating is correspondingly above or below the mean in the participant dataset. The *How strong?* model classifies with 87.2% accuracy using this procedure, significantly outperforming the baseline, $\chi^2(1, N = 4368) = 700.9$, $p < 0.001$. (Note that this situation does not require more sophisticated evaluation techniques, like cross-validation, because the model is highly constrained and the threshold is not learned.)

Some predictive variables surprised us. For instance, *Inbox thread depth* negatively (and strongly) affects tie strength. This finding also clashes with existing work. In [41], Whittaker, et al., report that familiarity between Usenet posters increases thread depth. One way to resolve this disparity is to note that there may be a fundamental difference between the completely private threads found on Facebook (essentially a variant of email) and Usenet's completely public ones. Common ground theory [7] would suggest that strong ties can communicate very efficiently because of their shared understanding, perhaps manifesting as shorter Inbox threads. *Educational difference* also strongly predicts tie strength, with tie strength diminishing as the difference grows. This may have resulted from the university community to which our participants belonged. On the other hand, the result may have something to do with Facebook itself, a community that spread via universities. Some variables we suspected to impact tie strength did not. *Number of overlapping networks* and *Age difference*, while intuitively good predictors, made little appreciable difference to tie strength. ($\beta = 0.027$, $F_{1,971} = 3.08$, $p = 0.079$ and $\beta = -0.0034$, $F_{1,971} = 10.50$, $p = 0.0012$, respectively.)

The error analysis interviews illustrate the inherent complexity of some relationships. They also point the way toward future research. A model may never, and perhaps should never, predict some relationships. Wanting to reconnect with an ex-boyfriend comes to mind. Relationships like these have powerful emotions and histories at play. However, it may be possible to make better predictions about relationships like the professor-student one, a strong relationship relative to others like it. Incorporating organizational hierarchy may also improve a system's ability to reason about relationships like these. Merging deeper network knowledge with data about who extended the friend request also looks promising, as evidenced by the "he friended me" interview.

Practical Implications

We foresee many opportunities to apply tie strength modeling in social media. Consider privacy controls that understand tie strength. When users make privacy choices, a system could make educated guesses about which friends fall into trusted and untrusted categories. This might also depend on media type, with more sensitive media like photos requiring higher tie strengths. The approach would not help users set privacy levels for brand new friends, ones with whom there is no interaction history. Yet, it has two main advantages over the current state of the art: it adapts with time, and it establishes smart defaults for users setting access levels for hundreds of friends.

Or, imagine a system that only wants to update friends with novel information. Broadcasting to weak ties could solve this problem. Consider a politician or company that wants to broadcast a message through the network such that it only passes through trusted friends. Because strongly tied friends often reconcile their interests [17], a politician might look for new supporters among the strong ties of an existing one. Limiting the message's audience in this way may increase the success rate relative to the effort expended.

Social media has recently started suggesting new friends to users. However, sometimes we choose not to friend someone with good reason. For instance, a strong tie of a strong tie is not necessarily a friend at all: consider the beloved cousin of a best friend. Granovetter writes, "if strong ties A–B and A–C exist, and if B and C are aware of one another, anything short of a positive tie would introduce a 'psychological strain' into the situation" [17]. A system that understands tie strength might avoid "strain" by steering clear of these delicate situations. In fact, weak ties of existing friends may make better friend candidates, as it is less likely that users have already declined to friend them. More broadly, systems that understand tie strength might apply it to make better friend introductions, although deeper study would need to uncover how best to use it in this context.

Recent work suggests that the average number of social media friends continues to grow, currently above 300 [25]. With users keeping so many friends, social media has started to consolidate friend activity into a single stream. Facebook calls this the Newsfeed. However, the multiplicative nature of the types of friends crossed with the types of updates, e.g., photos, status, new friends, comments, etc., presents a difficult design problem. A system that prioritizes via tie strength, or allows users to tune parameters that incorporate tie strength, might provide more useful, timely and enjoyable activity streams.

Theoretical Implications

There is still more variance to understand. Certainly, more predictive variables could help, such as "behind-the-scenes" data like who friended who. However, throwing more data at the problem might not solve it; perhaps social media needs novel indicators. This raises new questions for theory. When modeling tie strength exclusively from social media, do we necessarily miss important predictors? What is the upper limit of tie strength predictability?

We believe our work makes three important contributions to existing theory. First, we defined the importance of the dimensions of tie strength as manifested in social media. This is novel especially in light of the fact that these weights do not always align with prior work. Second, we showed that tie strength can be modeled as a continuous value. Third, our findings reveal how the *Structural* dimension modulates other dimensions by filtering individual relationships through cliques. Previously, it was not well-understood how or if tie strength dimensions interacted.

Finally, we see a home for our results in social network analysis. Most work to date has assumed a present link or an absent link, omitting properties of the link itself. Introducing a complete tie strength model into social network analyses, perhaps even joining a social media model with real-world data, may enable novel conclusions about whole systems [26].

Limitations

We purposely worked from theory to extend this research beyond just Facebook. The specific predictive variable coefficients may not move beyond Facebook, but the dimension weights may. That being said, this work looks only at one social media site, at one time, using data available through the browser. We look forward to work evaluating the utility of "behind-the-scenes" data and to work contrasting these findings with other social media.

CONCLUSION

In this paper, we have revealed a specific mechanism by which tie strength manifests itself in social media. Many paths open from here. Social media designers may find traction fusing a tie strength model with a range of social media design elements, including privacy controls and information prioritization. Our follow-up interviews suggest profitable lines of future work. We hope that researchers in this field will find important new theoretical questions in this work, as well as opportunities to use tie strength to make new conclusions about large-scale social phenomena.

We believe this work addresses fundamental challenges for understanding users of socio-technical systems. How do users relate to one another in these spaces? Do the data left behind tell a consistent story, a story from which we can infer something meaningful? We think this work takes a significant step toward definitively answering these questions.

ACKNOWLEDGMENTS

This work was funded in part by NSF Grant 0643502 and by Microsoft. We thank Keith Hampton, Caroline Haythornthwaite, Cliff Lampe, Scott Golder and the entire Social Spaces group for early feedback on this work. We also thank Carolyn Gilbert, who in no way helped this paper get written, but is wonderful so she gets mentioned anyway.

REFERENCES

1. Adamic, L. A. and Adar, E. 2003. Friends and neighbors on the Web. *Social Networks, 25*(3), 211–230.
2. Albert, R. and Barabási, A. L. 2002. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics, 74*(1).
3. Bernard, H. R., Killworth, P., et al. 1984. The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Review of Anthropology, 13*, 495–517.
4. Burt, R. S. 2004. Structural Holes and Good Ideas. *American Journal of Sociology, 110*(2), 349–399.
5. Burt, R. *Structural Holes: The Social Structure of Competition.* Harvard University Press, 1995.
6. Chopra, S., Thampy, T., et al. Discovering the Hidden Structure of House Prices with a Non-parametric Latent Manifold Model. *Proc. KDD,* 2007. 173–182.
7. Clark, H. H. *Arenas of Language Use.* University Of Chicago Press, 1993.

8. Constant, D., Sproull, L., et al. 1996. The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice. *Organization Science, 7*(2), 119–135.

9. Cunningham, W. H., Cunningham, I. C., et al. 1977. The Ipsative Process to Reduce Response Set Bias. *Public Opinion Quaterly, 41*(3), 379–384.

10. Frakes, W. B. and Baeza-Yates, R. *Information Retrieval: Data Structures and Algorithms.* Prentice Hall, 1992.

11. Friedkin, N. E. 1980. A Test of Structural Features of Granovetter's Strength of Weak Ties Theory. *Social Networks, 2*, 411–422.

12. Gergle, D., Kraut, R. E., et al. The Impact of Delayed Visual Feedback on Collaborative Performance. *Proc. CHI,* 2006. 1303–1312.

13. Gilbert, E., Karahalios, K., et al. The Network in the Garden: An Empirical Analysis of Social Media in Rural Life. *Proc. CHI,* 2008. 1603–1612.

14. Golder, S., Wilkinson, D. M., et al. Rhythms of Social Interaction: Messaging Within a Massive Online Network. *Proc. 3rd International Conference on Communities and Technologies (CT2007),* 2007.

15. Google Scholar. http://scholar.google.com/scholar?q=weak+ties. Accessed September 16, 2008.

16. Granovetter, M. 1983. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory, 1*, 201–233.

17. Granovetter, M. S. 1973. The Strength of Weak Ties. *The American Journal of Sociology, 78*(6), 1360–1380.

18. Granovetter, M. *Getting a Job: A Study of Contacts and Careers.* University Of Chicago Press, 1974.

19. Greasemonkey. http://www.greasespot.net. Accessed September 16, 2008.

20. Hansen, M. T. 1999. The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge Across Organization Subunits. *Administrative Science Quarterly, 44*(1), 82–111.

21. Haythornthwaite, C. 2002. Strong, Weak, and Latent Ties and the Impact of New Media.. *Information Society, 18*(5), 385–401.

22. Haythornthwaite, C. and Wellman, B. 1998. Work, Friendship, and Media Use for Information Exchange in a Networked Organization. *J. Am. Soc. Inf. Sci., 49*(12), 1101–1114.

23. Krackhardt, D. The Strength of Strong Ties: The Importance of Philos in Organizations. In N. Nohria and R. Eccles (Ed.), *Networks and Organizations: Structure, Form and Action* (216–239). Boston, MA: Harvard Business School Press.

24. Krackhardt, D. and Stern, R. N. 1988. Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly, 51*(2), 123–140.

25. Lampe, C., Ellison, N., et al. Changes in Participation and Perception of Facebook. *Proc. CSCW,* 2008. 721–730.

26. Laumann, E. O., Gagnon, J. H., et al. 1989. Monitoring the AIDS Epidemic in the United States: A Network Approach. *Science, 244*(4909), 1186–1189.

27. Liben-Nowell, D. and Kleinberg, J. The Link Prediction Problem for Social Networks. *Proc. of the 12th international conference on Information and knowledge management,* 2003. 556–559.

28. Lin, N., Dayton, P. W., et al. 1978. Analyzing the Instrumental Use of Relations in the Context of Social Structure. *Sociological Methods Research, 7*(2), 149–166.

29. Lin, N., Ensel, W. M., et al. 1981. Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment. *American Sociological Review, 46*(4), 393–405.

30. Marin, A. 1981. Are Respondents More Likely To List Alters with Certain Characteristics?: Implications for Name Generator Data. *Social Networks, 26*(4), 289–307.

31. Marsden, P. V. 1981. Core Discussion Networks of Americans. *American Sociological Review, 52*(1), 122–131.

32. Marsden, P. V. 1990. Network Data and Measurement. *Annual Review of Sociology, 16*(1), 435–463.

33. Marsden, P. V. and Campbell, K. E. 1990. Measuring Tie Strength. *Social Forces, 63*(2), 482–501.

34. Pennebaker, J. W. and Francis, M. E. *Linguistic Inquiry and Word Count.* Lawrence Erlbaum, 1999.

35. Rapleaf. RapLeaf Study of Social Network Users vs. Age. http://business.rapleaf.com/company_press_2008_06_18.html. Accessed September 16, 2008.

36. Schaefer, C., Coyne, J. C., et al. 1990. The Health-related Functions of Social Support. *Journal of Behavioral Medicine, 4*(4), 381–406.

37. Shi, X., Adamic, L. A., et al. 2007. Networks of Strong Ties. *Physica A: Statistical Mechanics and its Applications, 378*(1), 33–47.

38. Tahmincioglu, E. Facebook Friends As Job References? http://www.msnbc.msn.com/id/26223330. Accessed September 16, 2008.

39. Uzzi, B. 1999. Embeddedness in the Making of Financial Capital: How Social Relations and Networks Benefit Firms Seeking Financing. *American Sociological Review, 64*(4), 481–505.

40. Wellman, B. and Wortley, S. 1990. Different Strokes from Different Folks: Community Ties and Social Support. *The American Journal of Sociology, 96*(3), 558–588.

41. Whittaker, S., Terveen, L., et al. The Dynamics of Mass Interaction. *Proc. CSCW,* 1998. 257–264.

Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles

Brent Hecht[*], Lichan Hong[†], Bongwon Suh[†], Ed H. Chi[†]

*Northwestern University
Electrical Engineering and Computer Science
brent@u.northwestern.edu

[†]Palo Alto Research Center
Augmented Social Cognition Group
3333 Coyote Hill Road, Palo Alto, CA
{hong,suh,echi}@parc.com

ABSTRACT

Little research exists on one of the most common, oldest, and most utilized forms of online social geographic information: the "location" field found in most virtual community user profiles. We performed the first in-depth study of user behavior with regard to the location field in Twitter user profiles. We found that 34% of users did not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools. When users did input their location, they almost never specified it at a scale any more detailed than their city. In order to determine whether or not natural user behaviors have a real effect on the "locatability" of users, we performed a simple machine learning experiment to determine whether we can identify a user's location by only looking at what that user tweets. We found that a user's country and state can in fact be determined easily with decent accuracy, indicating that users implicitly reveal location information, with or without realizing it. Implications for location-based services and privacy are discussed.

Author Keywords

Location, location-based services, Twitter, privacy, geography, location prediction, social networks

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

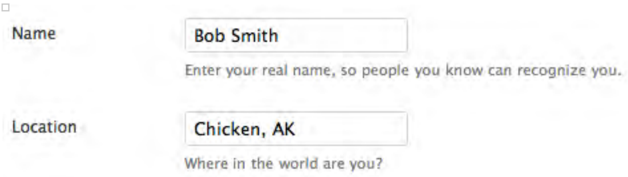
Human Factors, Design

INTRODUCTION

Interest in geographic information within the HCI community has intensified in the past few years. Academic HCI research has seen an increase in the number of papers on geographic information (e.g. [8, 18, 20, 21, 24, 26]). Industry has experienced an even greater spike in activity. Geotagged photo map interfaces have become commonplace (e.g. in Flickr and iPhoto), Google's Buzz has integrated a geographic component since its inception, and companies like Yelp have embraced the geographic nature of their user-generated content wholeheartedly.

Despite this increased interest in "geo", one of the oldest, most common forms of geographic information in the Web 2.0 world has escaped detailed study. This is the information that exists in the "location" field of user profiles on dozens of immensely popular websites. Facebook has had "Current City" and "Hometown" fields for years. Flickr allows users to enter their hometown and current location in their user profile, and the recently-launched music social network Ping by Apple has "Where I Live" as one of its profile fields.

This gap in understanding has not stopped researchers and practitioners from making ample use of the data entered into location fields. In general, it has been assumed that this data is strongly typed geographic information with little noise and good precision – an assumption that has never been validated. Backstrom et al. [1], for instance, wrote that "there is little incentive to enter false information, as leaving the field blank is an easier option". Similarly, Twitter reported that many location-based projects "are built using the simple, account-level location field folks can fill out as part of their profile". [25] This includes the "Nearby" feature of Twitter's official iPhone app, which is designed to show tweets that are close to the user's present location.



Name: Bob Smith
Enter your real name, so people you know can recognize you.

Location: Chicken, AK
Where in the world are you?

Figure 1. A screenshot from the webpage on which Twitter users enter location information. Location entries are entirely freeform, but limited to 30 characters.

In this paper, we conduct an in-depth study of user profile location data on Twitter, which provides a freeform location field without additional user interface elements that

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

encourage any form of structured input (Figure 1). The prompt is simply "Where in the world are you?" This environment allows us to observe users' natural, "organic" behavior as best as possible, thus illuminating actual user practices.

In the first part of this paper, we report the results derived from an extensive investigation of thousands of users' location entries on Twitter. We demonstrate that users' behavior with respect to the location field is richly varied, contrary to what has been assumed. We also show that the information they enter into the field is both highly diverse and noisy. Finally, our results suggest that most users organically specify their location at the city scale when they do specify their location.

For practitioners and researchers, it may be important to discover the rough location of the large percentage of users who did not disclose their true location. How can location-based services (LBS) ranging from information retrieval to targeted advertising leverage location field information given its noisy nature? Do users reveal location information through other behaviors on Twitter that can be used to effectively "fill in" the location field?

To answer both these questions, we considered users' *implicit* location sharing behavior. Since there are many forms of this implicit behavior, we decided to evaluate the most basic: the act of tweeting itself. In other words, how much information about her or his location does the average Twitter user disclose implicitly *simply by tweeting*? The second part of this paper presents a machine learning experiment that attempts to answer this question. We found that by observing only a user's tweets and leveraging simple machine learning techniques, we were reasonably able to infer a user's home country and home state. While we might never be able to predict location to GPS-level accuracy reliably using tweet content only, knowing even the country or the state of a user would be helpful in many areas such as answering search queries and targeted advertisement. In other words, users' most basic behavior on Twitter somewhat implicitly "fills out" the location field for them, better enabling LBS but also raising privacy concerns.

In summary, our contributions are fourfold:

- To the best of our knowledge, we provide the first in-depth study of user behavior in relation to one of the oldest and most common forms of online social geographic information: the location field in user profiles.
- We find that users' natural location field behavior is more varied and the information they submit is more complex than previously assumed.
- We show that the traditional tools for processing location field information are not properly equipped to handle this varied and noisy dataset.
- Using simple machine learning techniques to guess at users' locations, we demonstrate that the average user reveals location information simply by tweeting.

Following this introduction and a related work section, we describe how we collected our data from Twitter, as this is central to both of our studies. Next, we detail our characterization study and its implications. Following that, we describe the machine learning study. Finally, we close with a conclusion and discussion of future work.

Finally, before moving on, it is important to note that this work is descriptive in nature and does not focus on causal explanations for users' natural behavior. For instance, some users may decide not to enter their location for privacy reasons, while others may do so due to lack of interest or the belief that interested people already know their location. While some clues as to users' motivations can be gleaned from our first study, we leave in-depth causal analysis to future work.

RELATED WORK

Work related to this paper primarily arises from four areas: (1) research on microblogging sites like Twitter, (2) work on location disclosure behavior, (3) the location detection of users who contribute content to Web 2.0 sites, and (4) prediction of private information.

Various researchers have studied Twitter usage in depth. For instance, Honeycutt and Herring [10] examined the usage of the "@" symbol in English tweets. boyd et al. [3] studied how retweets are used to spread information. By manually coding 3,379 tweets, Naaman et al. [17] found that 20% of users posted tweets that are informational in nature, while the other 80% posted tweets about themselves or their thoughts.

With regard to the Twitter location field, Java et al. [11] found that in their dataset of 76K users, 39K of them provided information in their "location" field. They applied the Yahoo! Geocoding API[1] to the location field of these 39K users to show the geographical distribution of users across continents. Using the self-reported "utc_offset" field in user profiles, Krishnamurthy et al. [12] examined the growth of users in each continent over time. In the area of machine learning, Sakaki et al. [22] used the location field as input to their spatiotemporal event detection algorithms.

Location disclosure behavior has been investigated both in the research community and in the popular press. For instance, Barkhuus et al. [2] concluded that this behavior must be understood in its social context. In our case, this context is the entire "Twittersphere", as all data examined was in public profiles. Ludford et al. [15] identified several heuristics for how people decide which locations to share, such as "I will not share residences [or] private workplaces." In the popular press, the New York Times

[1] http://developer.yahoo.com/maps/rest/V1/geocode.html

recently featured an article [4] reporting that just 4% of U.S. residents had tried location-based services.

In the third area – location detection – the most relevant works include Lieberman and Lin [13], Popescu and Grefenstette [19], and Backstrom et al. [1]. The recentness of these papers, all published in the past two years, demonstrates that this is an active area of research. Lieberman and Lin sought to determine the location of Wikipedia users, but did so using very specific properties of the Wikipedia dataset that do not generalize to the rest of the Web 2.0 world. In addition, they did not examine the natural behavior of Wikipedia users on their "user pages", which are the Wikipedia equivalent of user profiles.

Popescu and Grefenstette [19] attempted to predict the home country of Flickr users through the analysis of their place name photo tags and latitude and longitude geotags. In contrast to both this paper and the Lieberman and Lin work, once our model has been trained, our location prediction algorithms do not depend on a user submitting any geographic information. Popescu and Grefenstette also did no qualitative examination.

Backstrom et al. [1] used the social network structure of Facebook to predict location. As noted below, our work focuses on the content submitted by users, not the social network, although both approaches could be combined in future work.

In terms of prediction of profile fields or other withheld information, our work stands out from other recent research (e.g. [1, 14]) in two ways: (1) first we examine the user practices surrounding the information that we are trying to predict, and (2) we make predictions solely from content innate to its medium and do not leverage any portion of the social graph.

DATA COLLECTION

From April 18 to May 28, 2010, we collected over 62 million tweets from the Spritzer sample feed, using the Twitter streaming API[2]. The Spritzer sample represents a random selection of all public messages. Based on a recent report that Twitter produced 65 million tweets daily as of June 2010 [23], we estimate that our dataset represents about 3-4% of public messages.

From these 62 million tweets, we further identified the tweets that were in English using a two-step combination of LingPipe's text classifier[3] and Google's Language Detection API[4]. All together, we identified 31,952,964 English tweets from our 62 million tweets, representing 51% of our dataset.

This research purposely does not consider the recent change to the Twitter API that allows location information to be embedded in each individual tweet [25]. We made this choice for two reasons. First, our focus is on the geographic information revealed in the "location" field of user profiles, a type of geographic information that is prevalent across the Web 2.0 world. Second, we found that only 0.77% of our 62 million tweets contained this embedded location information. With such a small penetration rate, we were concerned about sampling biases.

STUDY 1: UNDERSTANDING EXPLICIT USER BEHAVIOR

Study 1: Methods

Our 32 million English tweets were created by 5,282,657 unique users. Out of these users, we randomly selected 10,000 "active" users for our first study. We defined "active" as having more than five tweets in our dataset, which reduced our sampling frame to 1,136,952 users (or 22% of all users). We then extracted the contents of these 10,000 users' location fields and placed them in a coding spreadsheet. Two coders examined the 10,000 location field entries using a coding scheme described below. Coders were asked to use any information at their disposal, from their cultural knowledge and human intuition to search engines and online mapping sites. Both coders agreed initially on 89.2% of the entries, and spent one day discussing and coming to an agreement on the remaining 10.8%.

The coding scheme was designed to determine the *quality* of the geographic information entered by users as well as the *scale* of any real geographic information. In other words, we were interested in examining the 10,000 location entries for their properties along two dimensions: quality and geographic scale. We measured quality by whether or not geographic information was imaginary or whether it was so ambiguous as to refer to no specific geographic footprint (e.g. "in jail" instead of "in Folsom Prison"). In the case of location field entries with even the most rudimentary real geographic information, we examined at what scale this information specified the user's location. In other words, did users disclose their country? Their state? Their city? Their address?

Since both coders are residents of the United States, only data that was determined to be within the United States was examined for scale. This choice was made due to the highly vernacular nature of many of the entries, thus requiring a great deal of cultural knowledge for interpretation.

Study 1: Results

Information Quality

As shown in Figure 2, only 66% of users manually entered any sort of valid geographic information into the location field. This means that although the location field is usually assumed by practitioners [25] and researchers (e.g. in [11] and [22]) to be a field that is as associated with geographic information as a date field is with temporal information,

[2] http://dev.twitter.com/pages/streaming_api

[3] http://alias-i.com/lingpipe/demos/tutorial/langid/read-me/html

[4] http://code.google.com/apis/ajaxlanguage/documentation/

this is definitely not the case in our sample. The remaining one-third of users were roughly split between those that did not enter any information and those that entered either non-real locations, obviously non-geographic information, or locations that did not have specific geographic footprints.

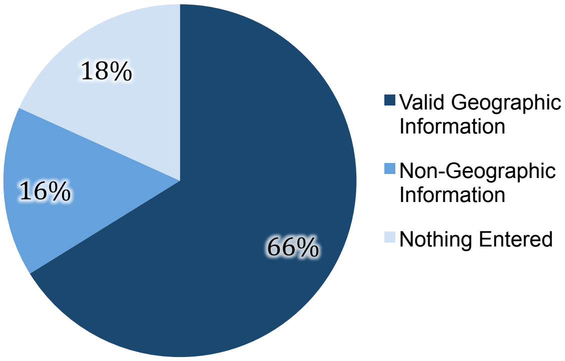


Figure 2: The distribution of manually entered location field data. Roughly one-third of users did not enter valid geographic information into the location field. 16% entered non-geographic information, while 18% entered nothing at all.

An analysis of the non-geographic information entered into the location field (the 16% in Figure 2) revealed it to be highly unpredictable in nature (see Table 1). A striking trend was the theme of Justin Bieber, who is a teenage singer. A surprising 61 users (more than 1 in 200 users) co-opted the location field to express their appreciation of the pop star. For instance, a user wrote that s/he is located in "Justin Biebers heart" (inspiring the title of this paper) and another user indicated s/he is from "Bieberacademy". Justin Bieber was not the only pop star that received plaudits from within the location field; United Kingdom "singing" duo Jedward, Britney Spears, and the Jonas Brothers were also turned into popular "locations".

Another common theme involved users co-opting the location field to express their desire to keep their location private. One user wrote "not telling you" in the location field and another populated the field with "NON YA BISNESS!!" Sexual content was also quite frequent, as were "locations" that were insulting or threatening to the reader (e.g. "looking down on u people"). Additionally, there was a prevalent trend of users entering non-Earth locations such as "OUTTA SPACE" and "Jupiter".

A relatively large number of users leveraged the location field to express their displeasure about their current location. For instance, one user wrote "preferably anywhere but here" and another entered "redneck hell".

Entering non-real geographic information into the location field was so prevalent that it even inspired some users in our sample to make jokes about the practice. For instance, one user populated the location field with "(insert clever phrase here)".

Frequency counts for these types of non-geographic information are reported in Table 1. To generate this table, non-geographic entries were coded by two human coders and the lists were merged. Categories were determined using a grounded approach, and each "location" was allowed to have zero or more categories. Because of the highly vernacular nature of this data, coders were instructed to only categorize when highly confident in their choice. As such, the numbers in Table 1 must be considered lower bounds.

Information Type	# of Users
Popular Culture Reference	195 (12.9%)
Privacy-Oriented	18 (1.2%)
Insulting or Threatening to Reader	69 (4.6%)
Non-Earth Location	75 (5.0%)
Negative Emotion Towards Current Location	48 (3.2%)
Sexual in Nature	49 (3.2%)

Table 1: A selection of the types of non-geographic information entered into the location field. Many of these categories exhibited large co-occurrence, such as an overlap between "locations" that were sexual in nature and those that were references to popular culture (particularly pop and movie stars). Percentages refer to the population of non-geographic information location field entries.

Note that, in the 66% of users who did enter real geographic information, we included all users who wrote *any* inkling of real geographic information. This includes those who merely entered their continent and, more commonly, those who entered geographic information in highly vernacular forms. For example, one user wrote that s/he is from "kcmo--call da po po". Our coders were able to determine this user meant "Kansas City, Missouri", and thus this entry was rated as valid geographic information (indicating a location at a city scale). Similarly, a user who entered "Bieberville, California" as her/his location was rated as having included geographic information at the state scale, even though the city is not real.

Information Scale

Out of the 66% of users with any valid geographic information, those that were judged to be outside of the United States were excluded from our study of scale. Users who indicated multiple locations (see below) were also filtered out. This left us with 3,149 users who were determined by both coders to have entered valid geographic information that indicated they were located in the United States.

When examining the scale of the location entered by these 3,149 users, an obvious city-oriented trend emerges (Figure 3). Left to their own devices, users by and large choose to disclose their location at exactly the city scale, no more and no less. As shown in Figure 3, approximately 64% of users specified their location down to the city scale. The next most popular scale was state-level (20%).

When users specified intrastate regions or neighborhoods, they tended to be regions or neighborhoods that engendered significant place-based identity. For example, "Orange

County" and the "San Francisco Bay Area" were common entries, as were "Harlem" and "Hollywood". Interestingly, studying the location field behavior of users located within a region could be a good way to measure the extent to which people identify with these places.

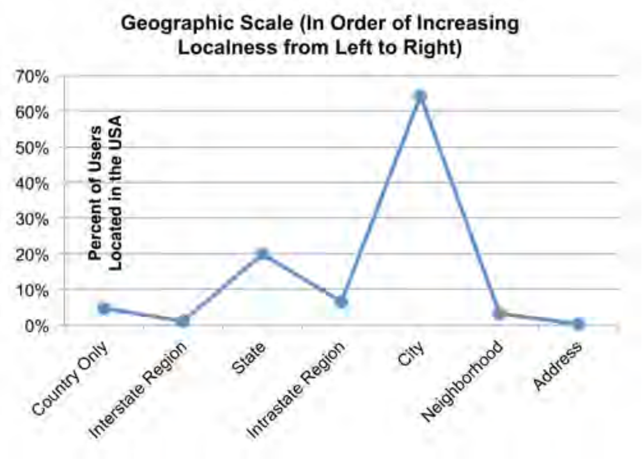


Figure 3: The scale of the geographic information entered by 3,149 users who indicated that they lived in the United States.

Multiple Locations

2.6% of the users (4% of the users who entered any valid geographic information) entered multiple locations. Most of these users entered two locations, but 16.4% of them entered three or more locations. Qualitatively, it appears many of these users either spent a great deal of time in all locations mentioned, or called one location home and another their current residence. An example of the former is the user who wrote "Columbia, SC. [atl on weekends]" (referring to Columbia, South Carolina and Atlanta, Georgia). An example of the latter is the user who entered that he is a "CALi b0Y $TuCC iN V3Ga$" (A male from California "stuck" in Las Vegas).

Automatically-entered Information

The most categorically distinct entries we encountered were the automatically populated latitude and longitude tags that were seen in many users' location fields. After much investigation, we discovered that Twitter clients such as ÜberTwitter for Blackberry smartphones entered this information. Approximately 11.5% of the 10,000 users we examined had these latitude and longitude tags in their location field. We did not include these users in Figure 2 or 3, as they did not manually enter their location data.

Study 1: Implications for Design

Failure of Traditional Geographic Information Tools

Our study on the information quality has vital implications for leveraging data in the location field on Twitter (and likely other websites). Namely, many researchers have assumed that location fields contain strongly typed geographic information, but our findings show this is demonstrably false. To determine the effect of treating Twitter's location field as strongly-typed geographic information, we took each of the location field entries that were coded as not having any valid geographic information (the 16% slice of the pie chart in Figure 2) and entered them into Yahoo! Geocoder. This is the same process used by Java et al. in [11]. A geocoder is a traditional geographic information tool that converts place names and addresses into a machine-readable spatial representation, usually latitude and longitude coordinates [7].

Of the 1,380 non-geographic location field entries, Yahoo! Geocoder determined 82.1% to have a latitude and longitude coordinate. As our coders judged none of these entries to contain any geographic information or highly ambiguous geographic information, this number should be zero (assuming no coding error). Some examples of these errors are quite dramatic. "Middle Earth" returned (34.232945, -102.410204), which is north of Lubbock, Texas. Similarly, "BieberTown" was identified as being in Missouri and "somewhere ova the rainbow", in northern Maine. Even "Wherever yo mama at" received an actual spatial footprint: in southwest Siberia.

Since Yahoo! Geocoder assumes that all input information is geographic in nature, the above results are not entirely unexpected. The findings here suggest that geocoders alone are not sufficient for the processing of data in location fields. Instead, data should be preprocessed with a geoparser, which disambiguates geographic information from non-geographic information [7]. However, geoparsers tend to require a lot of context to perform accurately. Adapting geoparsers to work with location field entries is an area of future work.

Attention to Scale in Automated Systems

Another important implication comes from the mismatch in revealed scale between the latitude and longitude generated automatically by certain Twitter clients and that revealed naturally by Twitter users. The vast majority of the machine-entered latitude and longitude coordinates had six significant digits after the decimal point, which is well beyond the precision of current geolocation technologies such as GPS. While it depends somewhat on the latitude, six significant digits results in geographic precision at well under a meter. This precision is in marked contrast with the city-level organic disclosure behavior of users. In our dataset, we found a total of only nine users (0.09% of the entire dataset) who had manually entered their location at the precision of an address, which is still less precise than a latitude and longitude coordinate expressed to six significant digits. However, this number could have been affected somewhat by the 30-character limit on the Twitter location field.

This mismatch leads us to a fairly obvious but important implication for design. Any system automatically populating a location field should do so, not with the exact latitude and longitude, but with an administrative district or vernacular region that contains the latitude and longitude coordinate. Fortunately, these administrative districts are easy to calculate with a reverse geocoding tool. Users should also be given a choice of the scale of this district or region (i.e. city, state, country), as users seem to have different preferences. This implication may apply to the "location" field on other sites as well as the location metadata associated with user-contributed content such as tweets and photos.

Other Implications

Another design implication is that users often want to have the ability to express sarcasm, humor, or elements of their personality through their location field. In many ways, this is not a surprise; people's geographic past and present have always been a part of their identity. We are particularly interested in the large number of users who expressed real geographic information in highly vernacular and personalized forms. Designers may want to invite users to choose a location via a typical map interface and then allow them to customize the place name that is displayed on their profile. This would allow users who enter their location in the form of "KC N IT GETS NO BETTA!!" (a real location field entry in our study) to both express their passion for their city and receive the benefits of having a machine-readable location, if they so desire.

Our findings also suggest that Web 2.0 system designers who wish to engender higher rates of machine-readable geographic information in users' location fields may want to force users to select from a precompiled list of places.

People who entered multiple locations motivate an additional important implication for design. This gives credence to the approach of Facebook and Flickr, which allow users to enter both a "current" location and a "hometown" location. However, the behavior of these users also suggests that this approach should be expanded. We envision a flexible system that would allow users to enter both an arbitrary number of locations and describe each of those locations (e.g. "home", "favorite place", etc.)

STUDY 2: UNDERSTANDING IMPLICIT USER BEHAVIOR THROUGH MACHINE LEARNING

In the first study, we used human judges to look closely at the explicit information included in the location field. However, in domains such as location-based services it may be important to discover the rough location of the large percentage of users who did not disclose their true location. Privacy advocates would likely also be interested in understanding whether or not this can be done. Given the results of prior research on location detection [1, 13, 19], we wanted to determine how much *implicit* location information users disclose simply by their day-to-day tweeting behavior. To do so, we used the data gathered above to conduct a set of machine learning experiments.

The goal of these experiments was to determine users' locations simply by examining the text content of their tweets. Specifically, we sought to predict a user's country and state solely from the user's tweets. We did not have enough data to work at a city level. As noted above, the contribution here is to demonstrate the implicit location sharing behavior of users in the context of their explicit behavior (with an eye towards location-based services, as well as privacy).

Study 2: Methods

In this subsection, we describe the general methodology behind our machine learning experiments, in which we use a classifier and a user's tweets to predict the country and state of that user. First, we discuss how we modeled each Twitter user for the classifier and how we shrank these models into a computationally tractable form. Next, we highlight the methodology behind the building of our training sets for the classifier and explain how we split off a subset of this data for validation purposes. Finally, we describe our classification algorithm and sampling strategies, as well as the results of our machine learning experiments.

Model Construction and Reduction

To classify user locations, we developed a Multinomial Naïve Bayes (MNB) model [16]. The model accepts input in the form of a term vector with each dimension in the vector representing a term and the value of the dimension representing the term count in a user's tweets. We also tried advanced topic models including Explicit Semantic Analysis [6]. However, a pilot study revealed that the simple term frequency (TF) MNB model greatly outperformed the more complex models. Thus, we only report the TF results.

For computational efficiency, we settled on using a fixed-length 10,000-term vector to represent each user in all cases. We tried two different methods for picking which 10,000 terms to use. The first was the standard frequency-based selection model in which we picked the 10,000 most common terms in our corpus. We called this algorithm "COUNT", for its reliance on term counting.

We also developed a more advanced algorithm designed to select terms that would discriminate between users from different locations. This simple heuristic algorithm, which we call the "CALGARI" algorithm, is based on the intuition that a classification model would perform better if the model includes terms that are more likely to be employed by users from a particular region than users from the general population. It is our assumption that these terms will help our classifier more than the words selected by the COUNT algorithm, which includes many terms that are common in all countries or states considered (e.g. "lol").

The CALGARI algorithm calculates a score for each term present in the corpus according to the following formula:

$$CALGARI(t) = \begin{cases} 0 & \text{if} \quad users(t) < MinU \\ \dfrac{\max\left(P(t \mid c = C)\right)}{P(t)} & \text{if} \quad users(t) \geq MinU \end{cases}$$

where t is the input term, *users* is a function that calculates the number of users who have used t at least once, $MinU$ is an input parameter to filter out individual idiosyncrasies and spam (set to either 2 or 5 in our experiments), and *C* is a geographic class (i.e. a state or country). The max function simply selects the maximum conditional probability of the term given each of the classes being examined. Terms are then sorted in descending order according to their scores and the top 10,000 terms are selected for the model. After picking the 10,000 terms, each user's Twitter feed was represented as a term vector using this list of 10,000 terms as dimensions, populated by the feed's term frequencies for each dimension.

A good example of the differences between CALGARI and COUNT was found in the average word vector for each algorithm for users in Canada. Among the terms with the highest weights for the CALGARI algorithm were "Canada", "Calgari", "Toronto" and "Hab". On the other hand, the top ten for COUNT included "im", "lol", "love", and "don't". Note that the CALGARI algorithm picked terms that are much more "Canadian" than those generated by the COUNT algorithm. This includes the #2 word "Calgari" (stemmed "Calgary"), which is the algorithm's namesake.

Developing Ground Truth Data

In order to build a successful classifier, we first needed to generate high-precision ground truth data. The main challenge here was to match a large group of users with their *correct* country and/or state. Through this group of users, the classifier could then learn about the tweeting patterns of each country and state population, and use these patterns to make predictions about *any* user.

Our starting point in developing the ground truth data was the 32 million English tweets created by over 5 million users. We first applied an extremely high-precision, very low-recall geocoder similar to that used in Hecht and Gergle [8]. The geocoder examines the text of the location field of each user and attempts to match it against all English Wikipedia article titles. If the location field matches (case-insensitive) a title exactly, latitude and longitude coordinates are searched for on the corresponding Wikipedia page[5]. If coordinates are found, the user is assigned that latitude and longitude as her location. If not, the user is excluded. We validated the precision of this method by testing it against the same non-geographic data that was input into the Yahoo! Geocoder in Study 1. Our Wikipedia-based geocoder correctly determined that none of the input entries was an actual location.

The Wikipedia-based geocoder and the automatically entered latitude and longitude points allowed us to identify the coordinates for 588,258 users. Next, we used spatial data available from ESRI and the United States Census to calculate the country and state (if in the United States) of the users. This process is known as reverse geocoding.

In order to avoid problems associated with having a small number of tweets for a given user, we further restricted our ground truth data to those users who had contributed ten or more tweets to our dataset. In doing so, we removed 484,449 users from consideration.

We also required that all users in our dataset have a consistent country and state throughout the sample period. A tiny minority of users manually changed their location information during the sample period. In addition, a larger minority of users had their location changed automatically by Twitter clients. This temporal consistency filter pruned an additional 4,513 users from consideration.

In the end, our ground truth data consisted of 99,296 users for whom we had valid country and state information and 10 or more tweets. As noted earlier, this ground truth data was the sampling frame for deriving our training and validation sets for all machine learning experiments.

Training and Validation Sets

In each experiment, we used a specific subset (described below) of the ground truth data as training data. Since the CALGARI algorithm and the COUNT algorithm both involve "peeking" at the ground truth data to make decisions about which dimensions to include in the term vectors, the use of independent validation sets is vital. In all experiments, we split off 33% of the training data into validation sets. These validation sets were used *only* to evaluate the final performance of each model. In other words, the system is totally unaware of the data in the validation sets until it is asked to make predictions about that data. The validation sets thus provide an accurate view of how the machine learner would perform "in the wild." We used two sampling strategies for generating training and validation sets.

Sampling Strategies

In both our country-scale and state-scale experiments, we implemented two different sampling strategies to create the training data from the ground truth data. The first, which we call "UNIFORM", generated training and validation sets that exhibited a uniform distribution across classes, or countries and states in this context. This is the sampling strategy employed by Popescu and Grefenstette [19]. The experiments based on the UNIFORM data demonstrate the ability of our machine learning methods to tease out location information in the absence of the current demographic trends on Twitter.

The second sampling strategy, which we call "RANDOM", randomly chose users for our training and validation datasets. When using "RANDOM" data, the classifier considers the information that, for example, a user is much more likely to be from the United States than from Australia given population statistics and Twitter adoption rates. In other words, prior probabilities of each class (country or state) are considered. The results from experiments on the "RANDOM" data represent the amount of location information our classifier was able to extract *given* the demographics of Twitter.

[5] Hundreds of thousands of Wikipedia articles have latitude and longitude points embedded in them by users.

Sampling Strategy	Model Selection	Accuracy	Baseline Accuracy	% of Baseline Accuracy
Country-Uniform-2500	**Calgari**	**72.71%**	**25.00%**	**291%**
Country-Uniform-2500	Count	68.44%	25.00%	274%
Country-Random-20K	**Calgari**	**88.86%**	**82.08%**	**108%**
Country-Random-20K	Count	72.78%	82.08%	89%
State-Uniform-500	**Calgari**	**30.28%**	**5.56%**	**545%**
State-Uniform-500	Count	20.15%	5.56%	363%
State-Random-20K	Calgari	24.83%	15.06%	165%
State-Random-20K	**Count**	**27.31%**	**15.06%**	**181%**

Table 2: A summary of results from the country-scale and state-scale experiments. The better performing model selection algorithm is bolded for each experiment. The CALGARI result reported is the best generated by *MinU* = 2 or *MinU* = 5.

Evaluation of the Classifier

In the end, we conducted a total of four experiments, each on a differently sampled training and validation set (Table 2). In each experiment, we tested both the CALGARI and COUNT algorithms, reporting the accuracy for both. The machine learning algorithm and training/validation set split were identical across all four experiments.

For the country-prediction experiments, we first focused on the UNIFORM sampling strategy. From our ground truth data, 2,500 users located in the United States, the United Kingdom, Canada, and Australia were randomly selected, resulting in 10,000 users total. These four countries were considered because there are less than 2,500 users in each of the other English-speaking countries represented among the 99,296 ground truth users. As noted above, 33% of these users were then randomly chosen for our validation set and removed from the training set. The remainder of the training set was passed to one of two model selection algorithms: CALGARI and COUNT. We then trained our Multinomial Naïve Bayes classifier with the models and evaluated on the validation set removed earlier.

Next, we performed the same exercise, replacing the UNIFORM with the RANDOM sampling strategy, which selected 20,000 different users from our ground truth data, all of whom lived in one of the four countries listed above.

Our state-prediction experiments were roughly the same as our country experiments, with the only major difference in the development of the UNIFORM datasets. Since the U.S. states range in population from California's 36+ million people to Wyoming's 0.5+ million people, our dataset was skewed in a similar fashion. We only had very limited data for small-population states like Wyoming. In fact, out of all our 99,296 ground truth users, we only had 31 from Wyoming. As such, we only included the 18 states with 500 or more users in our UNIFORM dataset.

Study 2: Results

Country-prediction Experiments

For the UNIFORM sampling strategy, the best performing algorithm was CALGARI, which was able to predict the country of a user correctly 72.7% of the time, simply by examining that user's tweets. Since we considered four different countries in this case, one could achieve 25% accuracy by simply randomly guessing. Therefore, we also report the accuracy of our classifier relative to the random baselines, which in the best case here was 291% (or 2.91x).

With the RANDOM sampling strategy, we needed to use a different baseline. Since 82.08% of sampled users were from the U.S., one could achieve 82.08% accuracy simply by guessing "United States" for every user. However, even with these relatively decisive prior probabilities, the CALGARI algorithm was capable of bringing the accuracy level approximately 1/3 of the way to perfection (88.9%). This represents a roughly 8.1% improvement.

State-prediction Experiments

The results of our state-prediction experiments were quite similar to those above but better. As can be seen in Table 2, the classifier's best UNIFORM performance relative to the random baseline was a great deal better than in the country experiment. The same is true for the RANDOM dataset, which included users from all 50 states (even if there were only a dozen or so users from some states).

The baselines were lower in each of these experiments because we considered more states than we did countries. The UNIFORM dataset included 18 states (or classes). The RANDOM dataset included all 50 plus the District of Columbia, with New York having the maximum representation at 15.06% of users. A baseline classifier could thus achieve 15.06% accuracy simply by selecting New York in every case.

Study 2: Discussion

Table 2 shows that in every single instance, the classifier was able to predict a user's country and/or state from the user's tweets at accuracies better than random. In most cases, the accuracy was several *times* better than random, indicating a strong location signal in tweets. As such, there is no doubt that *users implicitly include location information in their tweets*. This is true even if a user has not entered any explicit location information into the location field, or has entered a purposely misleading or humorous location (assuming that these users do not have significantly different tweeting behavior).

We did not attempt to find the optimal machine learning technique for location prediction from tweet content. As such, we believe that the accuracy of location prediction can be enhanced significantly by improving along four fronts: (1) better data collection, (2) more sophisticated machine learning techniques, (3) better modeling of implicit behaviors, especially those involving social contexts on Twitter, and (4) inclusion of more user metadata.

Study 2: Implications

An interesting implication of our work can be derived from the conditional probabilities tables of the classifier. By studying these tables, we developed a list of terms that could be used to both assist location-based services (LBS) and launch location "inference attacks" [14]. A selection of terms that have strong predictive power at the country and state scales is shown in Table 3.

Stemmed Word	Country	"Predictiveness"
"calgari"	Canada	419.42
"brisban"	Australia	137.29
"coolcanuck"	Canada	78.28
"afl"	Australia	56.24
"clegg"	UK	35.49
"cbc"	Canada	29.40
"yelp"	United States	19.08
Stemmed Word	**State**	**"Predictiveness"**
"colorado"	Colorado	90.74
"elk"	Colorado	41.18
"redsox"	Massachusetts	39.24
"biggbi"	Michigan	24.26
"gamecock"	South Carolina	16.00
"crawfish"	Louisiana	14.87
"mccain"	Arizona	10.51

Table 3: Some of the most predictive words from the (top) Country-Uniform-Calgari and (bottom) State-Uniform-Calgari experiments. Predictiveness is calculated as a probability ratio of the max. conditional probability divided by the average of the non-maximum conditional probabilities. This can be interpreted as the number of times more likely a word is to occur given that a person is from a specific region than from the average of the other regions in the dataset. In other words, an Arizonan is 10.51 times more likely to use the term "mccain" than the average person from the other states.

There appear to be four general categories of words that are particularly indicative of one's location. As has been known in the social sciences for centuries (e.g. the gravity model [5]) and seen elsewhere with user-generated content (UGC) [9,13], people tend to interact with nearby places. While in some cases this has been shown to be not entirely true [8], mentioning place names that are close to one's location is very predictive of one's location. In other words, tweeting about what you did in "Boston" narrows down your location significantly on average.

Tweeting about sports assists in location inference significantly, as can be seen in Table 3. Similarly, our classifier found that a user from Canada was six times more likely to tweet the word "hockey" than a user from any other country in our study.

A third major category of predictive terms involves current events with specific geographic footprint, emphasizing the spatio*temporal* nature of location field data. During the period of our data collection, several major events were occurring whose footprints corresponded almost exactly with the scales of our analyses. The classifier easily identified that terms like "Cameron", "Brown", and "Clegg" were highly predictive of users who were in the United Kingdom. Similarly, using terms related to the 2010 NBA playoffs was highly indicative of a user from the United States. More generally speaking, a machine learner could theoretically utilize any regionalized phenomenon. For example, a tweet about a flood at a certain time [24, 26] could be used to locate a user to a very local scale.

Finally, regional vernacular such as "hella" (California) and "xx" (U.K.) were highly predictive of certain locations. It is our hypothesis that this category of predictive words helped our term frequency models perform better than the more complex topic models. It seems that the more abstract the topic model, the more it smoothes out the differences in spelling or slang. Such syntactic features can be powerful predictors of location, however.

Given some Twitter users' inclination towards privacy, users might value the inclusion of this predictive word list into the user interface through warnings. Moreover, given some users' inclination towards location field impishness, users may enjoy the ability to easily use this type of information to fool predictive systems. In other words, through aversion or purposeful deception, users could avoid location inference attacks by leveraging these terms.

FUTURE WORK

Much future work has arisen from this study of explicit and implicit location field behavior. The most immediate is to examine the causal reasons for the organic location disclosure behavior patterns revealed by this work. This could be explored through surveys, for example.

With regard to the classifier, we are looking into including social network information into our machine learners. This would allow us to explore the combination of content-based and network-based [1] location prediction.

We also are working to extend our predictive experiments to other cultural memberships. For instance, there is nothing about our models that could not be adapted to predict gender, age group, profession, or even ethnicity.

Other directions of future work include examining per-tweet location disclosure, as well as evaluating location disclosure on social network sites such as Facebook. Of course, accessing a large and representative sample of location field data on Facebook will be a major challenge. We have also done research investigating the ability to use the surprisingly noisy yet very prevalent "time zone" field in user profiles to assist in location prediction.

CONCLUSION

In this work, we have made several contributions. We are the first to closely examine the information embedded in user profile location fields. Through this exploration, we have shown that many users opt to enter no information or non-real location information that can easily fool geographic information tools. When users do enter their real locations, they tend to be no more precise than city-scale.

We have also demonstrated that the explicit location-sharing behaviors should be examined in the context of implicit behaviors. Despite the fact that over one-third of Twitter users have chosen not to enter their location, we have shown that a simple classifier can be used to make predictions about users' locations. Moreover, these techniques only leverage the most basic activity in Twitter – the act of tweeting – and, as such, likely form something of a lower bound on location prediction ability.

Given the interest in LBS and privacy, we hope the research here will inspire investigations into other natural location-based user behaviors and their implicit equivalents.

ACKNOWLEDGEMENTS

We thank our reviewers, Bjoern Hartmann, and our colleagues at PARC and the Northwestern CollabLab for their helpful suggestions.

REFERENCES

1. Backstrom, L., Sun, E. and Marlow, C. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. *WWW '10*, Raleigh, NC.
2. Barkhuus, L., Brown, B., Bell, M., Hall, M., Sherwood, S. and Chalmers, M. From Awareness to Repartee: Sharing Location within Social Groups. *CHI '08*, Florence, Italy, 497-506.
3. boyd, d., Golder, S. and Lotan, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *HICSS '10*, Kauai, HI.
4. Miller, C. and Wortham, J. Technology Aside, Most People Still Decline to Be Located. *The New York Times.* August 30, 2010.
5. Fellman, J.D., Getis, A. and Getis, J. *Human Geography: Landscapes of Human Activities*. Mcgraw-Hill Higher Education, 2007.
6. Gabrilovich, E. and Markovitch, S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *IJCAI '07*, Hyberabad, 1606-1611.
7. Hecht, B. and Gergle, D. A Beginner's Guide to Geographic Virtual Communities Research. *Handbook of Research on Methods and Techniques for Studying Virtual Communities*, IGI, 2010.
8. Hecht, B. and Gergle, D. On The "Localness" Of User-Generated Content. *CSCW '10*, Savannah, Georgia.
9. Hecht, B. and Moxley, E. Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. *COSIT '09*, L'Aber Wrac'h, France, 88-105.
10. Honeycutt, C. and Herring, S.C. Beyond Microblogging: Conversation and Collaboration via Twitter. *HICCS '09.*
11. Java, A., Song, X., Finin, T. and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. *Joint 9th WEBKDD and 1st SNA-KDD Workshop '07*, San Jose, CA, 56-65.
12. Krishnamurthy, B., Gill, P. and Arlitt, M. A Few Chirps about Twitter. *First Workshop on Online Social Networks*, Seattle, WA, 19-24, 2008.
13. Lieberman, M.D. and Lin, J. You Are Where You Edit: Locating Wikipedia Users Through Edit Histories. *ICWSM '09,* San Jose, CA.
14. Lindamood, J., Heatherly, R., Kantarcioglu, M. and Thuraisingham, B. Inferring Private Information Using Social Network Data. *WWW '09*, Madrid, Spain.
15. Ludford, P.J., Priedhorsky, R., Reily, K. and Terveen, L.G. Capturing, Sharing, and Using Local Place Information. *CHI '07*, San Jose, CA, 1235-1244.
16. McCallum, A. and Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*, 41-48.
17. Naaman, M., Boase, J. and Lai, C.-H. Is it Really About Me? Message Content in Social Awareness Streams. *CSCW '10*, Savannah, GA.
18. Panciera, K., Priedhorsky, R., Erickson, T. and Terveen, L.G. Lurking? Cyclopaths? A Quantitative Lifecycle Analysis of User Behavior in a Geowiki. *CHI '10*, Atlanta, GA, 1917-1926.
19. Popescu, A. and Grefenstette, G. Mining User Home Location and Gender from Flickr Tags. *ICWSM '10.*
20. Priedhorsky, R. and Terveen, L.G. The Computational Geowiki: What, Why, and How. *CSCW '08*, San Diego.
21. Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D. and Srivastava, M. Biketastic: Sensing and Mapping for Better Biking. *CHI '10*, Atlanta, GA, 1817-1820.
22. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *WWW '10*, Raleigh, NC.
23. Schonfeld, E. Costolo: Twitter Now Has 190 Million Users Tweeting 65 Million Times A Day. *TechCrunch.* http://techcrunch.com/2010/06/08/twitter-190-million-users/, 2010.
24. Starbird, K., Palen, L., Hughes, A.L. and Vieweg, S. Chatter on The Red: What Hazards Thread Reveals about the Social Life of Microblogged Information. *CSCW '10*, Savannah, GA.
25. Twitter. Location, Location, Location. *twitterblog.* http://blog.twitter.com/2009/08/location-location-location.html, 2009.
26. Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. *CHI '10*, Atlanta, GA, 1079-1088.

The Sociological Review

The redistribution of methods: on intervention in digital social research, broadly conceived

Noortje Marres

Abstract: This paper contributes to debates about the implications of digital technology for social research by proposing the concept of the redistribution of methods. In the context of digitization, I argue, social research becomes *noticeably* a distributed accomplishment: online platforms, users, devices and informational practices actively contribute to the performance of digital social research. This also applies more specifically to social research methods, and this paper explores the phenomenon in relation to two specific digital methods, online network and textual analysis, arguing that sociological research stands much to gain from engaging with their distribution, both normatively and analytically speaking. I distinguish four predominant views on the redistribution of digital social methods: methods-as-usual, big methods, virtual methods and digital methods. Taking up this last notion, I propose that a redistributive understanding of social research opens up a new approach to the *re-mediation* of social methods in digital environments. I develop this argument through a discussion of two particular online research platforms: the Issue Crawler, a web-based platform for hyperlink analysis, and the Co-Word Machine, an online tool of textual analysis currently under development. Both these tools re-mediate existing social methods, and both, I argue, involve the attempt to render specific *methodology critiques* effective in the online realm, namely critiques of the authority effects implicit in citation analysis. As such, these methods offer ways for social research to intervene critically in digital social research, and more specifically, to endorse and actively pursue the re-distribution of social methods online.

Keywords: digital social research, social studies of science and technology, digital devices, online network analysis, online textual analysis, digital social methods

Introduction

As sociologists like to point out, the implications of technology for social life tend to be imagined in either highly optimistic or deeply pessimistic ways (Woolgar, 2002). Current debates about the implications of digitization for social research are no exception to this rule. The question of how digital devices,

The Sociological Review, 60:S1, pp. 139–165 (2012), DOI: 10.1111/j.1467-954X.2012.02121.x

and their proliferation across social life, transform social research is generating much interest today, and, as a consequence, the question of the 'social implications of technology' is now very often posed in relation to social research itself (Back, 2010; Savage *et al.*, 2010; boyd and Crawford, 2011). As it turns out, these discussions are no less susceptible to the polarizing effects of technology on the imagination, than, say, popular debates about the implications of cloning or robotics on society. While some propose that new technologies are opening up a golden age of social research, others argue that digitization has engendered a crisis for social research, creating a situation in which we risk to lose 'the human element' from view.

Both the optimistic and the pessimistic vision of digital social research start from a similar observation: digital technologies have enabled a broad range of new practices involving the recording, analysis and visualization of social life (Fielding *et al.*, 2008). Millions of blogs document everyday life on an ongoing basis; online platforms for social networking such as Facebook generate masses of data for social analysis; and applications of 'digital analytics' make it possible for everyone with access to these tools to analyse 'social behaviour' in real time. For the optimists, this situation implies a renaissance of social research: the new technologies and practices greatly enhance the empirical and analytic capacities of social research, and they render social research newly relevant to social life (Latour *et al.*, 2012). For the pessimists, the new digital sources of social intelligence announce not so much a rejuvenation of social research, but rather pose a serious threat to established traditions and forms of sociological research (Savage and Burrows, 2007). From this vantage point, the proliferation across social life of new technologies for recording, analysing and visualizing social life masks an underlying trend of a very different nature. These technologies are leading to the privatization of social research: they enable the displacement of social research to the corporate laboratories of big IT firms.

In this paper, I would like to unsettle this opposition between the utopian and dystopian imagination of digital technology in social research. I would like to contribute to debates about the implications of digitization for social research by exploring the concept and phenomenon of the *redistribution of research*. This notion has been put forward in the social studies of science and technology (STS) to complicate our understanding of the relations between science, technology, and society (Latour, 1988; Rheinberger, 1997; see also Whatmore, 2009). It highlights that scientific research tends to involve contributions from a *broad* range of actors: researchers, research subjects, funders, providers of research materials, infrastructure builders, interested amateurs, and so on. Scientific research, according to this notion, must be understood as a *shared accomplishment* of a diverse set of actors. This idea has clear implications for digital social research: it suggests that it may be a mistake to try and locate digital social research in a single domain, be it 'the university', or 'everyday practices like blogging', or 'the private laboratories of large IT firms'. Instead, we should examine how, in the context of digitization, the roles of social research are being distributed between a range of different actors: between researchers, research subjects, digital technologies, and so on. Moreover, the concept of redistribution directs attention to a possible implication of digitization for social research: digitization may be unsettling established divisions of labour in social research. If we use blogs in social research, does this mean that we are partly delegating the task of data collection to bloggers?

Here I would like to focus on the redistribution of a specific element in social research, namely methods. Digitization is widely said to have special implications for the role and status of social research methods in particular (Fielding *et al.*, 2008; Rogers, 2010; Adkins and Lury, 2009). Views on this matter, too, diverge: some propose that digital technology inaugurates an age of methodological innovation, as new technologies for data collection, analysis and visualization enable the further elaboration of existing methods and the development of new ones. Others are more inclined to emphasize the 'return of the same' masked by such claims to newness, proposing that the 'new' digital methods continue along the same path as the 'quantitative revolution' of the 1960s and 70s (boyd and Crawford, 2011; Uprichard *et al.*, 2008). These observations are no less pertinent than the optimistic and pessimistic diagnoses flagged above, but on the issue of method too, there seems to be potential in side-stepping the 'false choice' between an utopian and a dystopian diagnosis, and to examine instead whether and how digitization enables new ways of *distributing* methods among different agents involved in social research. Social methods, too, may be understood as a shared accomplishment, involving contributions of researchers, research subjects, technologies, and so on (Rogers, 2009). The question is how the digital inflects this circumstance.

The issue of the redistribution of methods is a slippery one, as the contributions of different agents to the enactment of methods are hard to pin down: to return to the above example, why would we call blogs agents of data collection, rather than data points in our data set? On what grounds? To prevent being paralysed by general questions like this, I will explore the redistribution of method here in a contextual and empirical way, namely by examining two online platforms for social research: Issue Crawler, a web-based application for network analysis which has been online for 10 years now, and a tool of online textual analysis that is currently under development, provisionally called The Co-Word Machine. Both of these tools adapt social research methods to the online environment, namely network and textual analysis, and more precisely, co-citation and co-word analysis.[1] And they can both be said to undertake a 'redistribution' of social research methods: these transpose onto the Web methods that have long been championed in social research and, in doing so, they come to rely on a different set of entities in the enactment of this method, such as Web crawlers and online data feeds. The translation of methods of network and textual analysis into online environments, I will emphasize, enables a form of critical intervention in digital social research: to implement these methods online is to offer a distinctive variation on more prevalent applications of methods of network and textual analysis in digital networked media. The overall aim, then, is to get a more precise sense of the space of intervention opened up by digital social methods – *of method as intervention* – online. First, however, I would like to revisit in more detail the current debate about the implications of digitization for social research.

The digitization of social life and the redistribution of social research

The ongoing debate about the implications of digital technology for social research has directed attention to three significant features of digitization. No doubt the most important one is the *proliferation of new devices, genres and formats for the documentation of social life*. The last decade has seen an explosion of digital technologies that enable people to report and comment upon social life, from photo-sharing via Flickr to the public gossip of Twitter. Such online platforms allow users to publicize their accounts of everyday life like never before, in the form of simple text or snapshots taken with mobile phones. Especially interesting about the new devices from a sociological perspective is that they enable *the routine generation of data about social life as part of social life* (Fielding *et al.*, 2008; see on this point also Marres, 2011). 'Social media' platforms, that is, embed the process of social data generation in everyday practices, whether in the form of people 'live' commenting on an event via Twitter to the smart electricity meters that record fluctuations in domestic energy use. Finally, the two previous developments cannot really be understood without considering the development of online platforms and tools for the *analysis* of digital social data.

These days, most online platforms come with 'analytics' attached: a set of tools and services facilitating the analysis of the data generated by said platforms, from blog posts to Facebook friends. In this respect, what is especially significant for social research about online platforms for 'user-generated content' is that they actively support the adaptation of these platforms for purposes of social research. An example here is Yahoo Clues, a recently launched online platform that makes data generated by the Yahoo search engine available for analysis, allowing 'you to instantly discover what's popular to a select group of searchers – by age or gender – over the past day, week or even over the past year' (see Figure 1).[2] Providing access to a searchable database of search engine queries, Yahoo Clues makes available for analysis an arguably new type of social data, in the form of millions of queries that people perform as part of everyday life. And as Yahoo Clues allows its users to break down popular queries in terms of searcher profiles (gender, age, geographic location), it enables a distinctively social form of analysis. It also provides an example of the 'relocation' of social research enabled by digitization, as it formats social analysis as a popular practice that 'anyone' might like to engage in.

Social theorists have been hard pressed to provide an integrated assessment of these various developments and their implications for social research. Some authors have sought to affirm the new popular appeal of social research, suggesting that we are today witnessing a radical expansion in the range of actors,

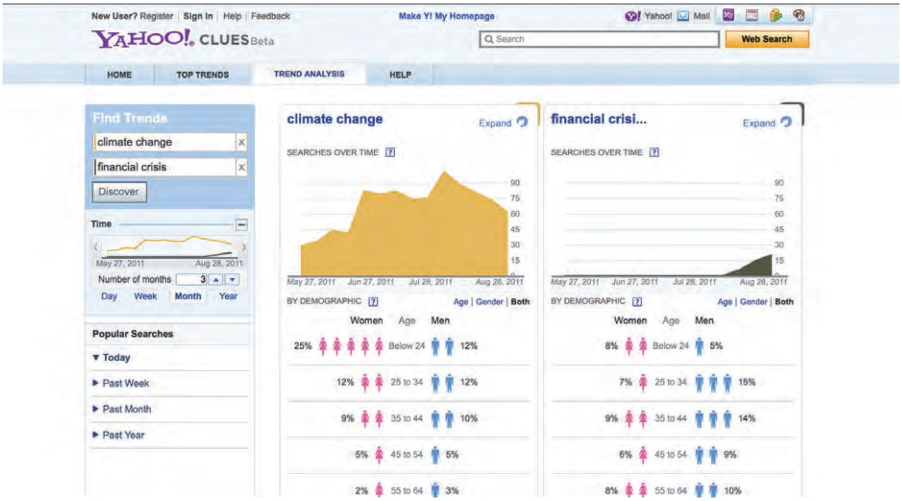


Figure 1: *Yahoo Clues: 'a new beta service that gives you a unique 'behind the scenes' look into popular trends across the millions of people who use Yahoo! to search each day' (July 2011).*

devices and settings caught up in the recording, reporting and analysis of social life. Some sociologists have been tempted to see in social media platforms a clear case of 'non-professional researchers enthusiastically engaging in the recording and reporting of social life' (my formulation). This would suggest that digitization is occasioning a revival of amateur-led social research, invoking memories of the English Mass Observation Movement, with its armies of lay people who documented scenes of everyday life in notebooks and questionnaires during the 1930s and 40s (Hubble, 2006; Savage, 2010). But others – indeed, in some cases the same authors – are more drawn to the dark side of this vision. Thus, Savage and Burrows (2007), in their influential article on 'The Coming Crisis of Empirical Sociology', prophesized that digitization signals the demise of sociology as a public form of knowledge. In their account, digitization, in spite of popular appearances, enables the *concentration* of social research capacity in a few well-resourced research centres, most notably of big IT firms. In this view, the wide popularity of online platforms for the collection, annotation and analysis of social data makes possible the displacement of social research to a few hubs of the digital economy, equipped for the central storage, processing and valuation of these data.

As has often been pointed out, the optimistic and the pessimistic diagnosis of a social phenomenon, while in some ways strictly opposed to another, may in other ways be neatly aligned (Haraway, 1991; Woolgar, 2002). As we know from the social study of consumer culture, dynamics of popularization and infrastructural concentration are by no means anti-thetical. As Celia Lury (1996, 2004) observed, popular fashion brands like Nike are marked by proliferation *and* unification, by the combination of an open-ended multiplicity of Nike-inflected social practices and a centralized orchestration of the phenomenon. To observe, then, that the spread of digital devices for the recording and analyzing social life occurs simultaneously with the concentration of control over the infrastructure that enables it is to note an all too familiar feature of post-industrial societies. It is just that, in the context of digitization, these dynamics are proving increasingly relevant to social research itself. But here I would like to argue that by concentrating on this overarching issue of the *displacement* of research capacity – to society at large, or the IT industry – we risk losing from view another, more fine-grained dynamic: the *redistribution* of social research *between* actors involved in social research. Rather than rushing to decide which sector of society will prove to be the biggest 'winner' – which will strengthen its position the most as a consequence of the digitization of social research? – we must then consider a more open-ended and complex process, namely that of the reconfiguration of the relations between the diverse set of agents caught up in social research.

The notion of the 'redistribution' of research has been put forward in STS and related fields to highlight processes of exchange *between* actors involved in social research. The notion emphasizes that the production of new knowledge and new technologies tend to involve complex interactions and transactions between a whole range of actors inside as well as outside the university, including research subjects, funding bodies, technological infrastructures, researchers, and so on. Research and innovation, then, is also a matter of the *transfer* of information, materials, and also more complex things like 'agency', between the various actors involved in research: when subjects agree to be interviewed or offer samples, when an institution allows a researcher into its archive, certain transactions occur that are critical to the production of new knowledge and/or technology. One example here is focus group research: this form of research relies on contributions from a range of actors, from research subjects, to research subject recruitment agencies and focus group moderators (Lezaun, 2007). Rather than assume that focus group research is conducted either 'in the university' or 'in the corporate sector', it therefore makes more sense to consider how this methodology enlists actors from different practices and domains, from marketing to government, activist organizations and academic research, and enables transactions among them. Indeed, social studies of focus group research have shown that the invention of the focus group in 1940s America enabled social research to take on new roles in society, among others as advisers on civic opinion (Lezaun, 2007; Grandclément and Gaglio, 2010). It also involved the development of new 'infrastructures' of social research, such as focus group research centres.

The concept of the 'redistribution of social research' has a number of implications for the debate about the consequences of digitization for social research. It suggests that some of the assumptions informing the debate about the displacement of research capacity, from the university to society, or from the public university to private industries, may be too simplistic. It suggests that the idea of the self-sufficient academy has *always* been a myth (Latour, 1988; Button, 1991; Callon *et al*., 2009 [2001]). For a long time already, academics have not been the only or even the main protagonists of research, as other actors have historically played active roles in the production of knowledge (Latour, 1988; Law, 2004). It is just that the conventional understanding of science and innovation makes it difficult to acknowledge the contributions of 'non-scientists' as meaningful contributions to research and innovation, without problematizing the status of our knowledge. Going against this conventional understanding, the concept of the redistribution of social research defines social research as a collective undertaking, involving a diverse set of actors in a variety of roles. Processes of inquiry, from this vantage point, are best understood as *inherently* distributed among a whole range of agencies, involving active contributions from research subjects, the experimental apparatus, funders of research, and so on (Latour, 1988; Rheinberger, 1997; Law, 2009).

Once we approach social research as distributed, the question of *displacement* of research capacity – away from academia; towards popular culture or industry – no longer seems the most relevant question to ask. Rather than trying to decide in what *singular* location research capacity is today most advantageously located, we should examine what digitization means for the distribution of roles in social research *between* various actors in and outside the university. Especially important about digitization, from this vantage point, is that it may well be unsettling divisions of labour in social research. Emerging practices of online social research that seek to take advantage of the new social data made available by platforms like Facebook and Twitter provide a case in point. Digital sociology student Sam Martin, for instance, turned to Twitter to analyse the racial abuse row over the prosecution of England footballer John Terry.[3] Using various applications from Google Docs to Yahoo Pipes and the Twitter API application programming interface, Martin culled messages mentioning John Terry from Twitter over a four-day period in February 2012. Using a programme called 'TagExplorer' she produced a network map of 'topconversationalists', which notably included 'Queens Park Ranger Captain and Footballer' Joey Barton, who was present at the pitch when the racial abuse incident occurred (see Figure 2).

This type of online research, which adapts social media applications to the purposes of social research, can be said to redistribute social research in various ways. Most notable, is its reliance on the social media platform Twitter itself: Twitter enables the ranking of twitter users according to the number of followers, tweets, and re-tweets, and in visualizing the corpus of messages using the measure of 'topconversationalists', Martin's small study arguably replicates some of the measures that are implicit in the medium under scrutiny. We should also note the various research tools and applications that allowed her to extract tweets from Twitter and visualize them, like Tagexplorer: these instruments, as well as the 'developer community' from which they sprang, here come to play a notable role in the organization of social research, and so did, arguably, the army of tweeters who in this study got a say on framing

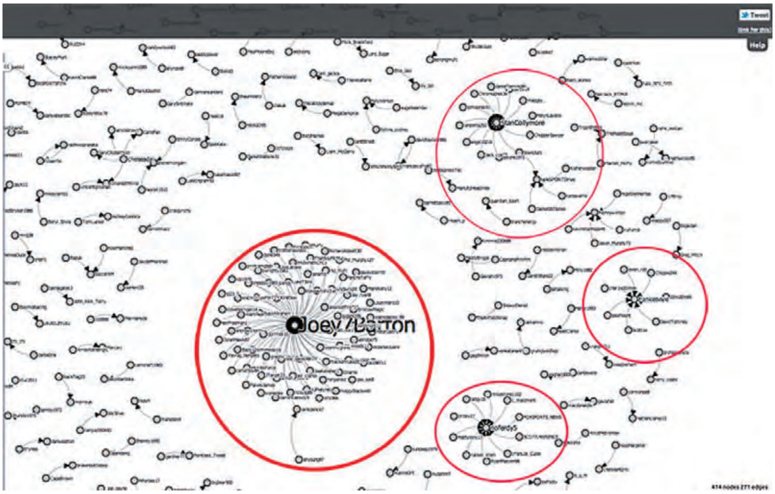


Figure 2: *Top conversationalists, the John Terry debate on Twitter, visualization using TAGSExplorer, 3–7 February 2012, by Sam Martin.*

phenomena as significant (by following some Twitter contributors rather than others).

Digital social research then enables particular redistributions of social research. Taking up digital online tools, sociologists are likely to enter into working relations with platforms, tool developers and analytic and visual devices which are operating in contexts and developed for purposes that are not necessarily those of sociology (Marres and Weltevrede, forthcoming). In examining such redistributions in digital social research, we can ask a question about the implications of digitization for social research that is at once more specific and open-ended than the question about displacement: to what extent does digitization enable *renegotiations of divisions of labour* in social research? At issue, then, is not only which institution or sector gets to define what social research is, and to occupy the ‘top spot’, but rather what relations between a range of different actors is enabled by particular, emerging digital social research practices. The notion of the redistribution of social research, furthermore, directs attention to *a much broader* set of actors and entities caught up in the process of the digitization of social research, including but not limited to: online platforms, users, databases, design agencies, algorithms, IT companies, digital culture commentators, information formats, social movements, and so on (see on this point also Madsen, 2012). The division of labour between users, devices and researchers in the conduct of social research, we then say, is being unsettled, contested and redefined in complex but quite specific ways.

The idea of the redistribution of social research can then provide some useful conceptual guidance, in examining the implications of digitization for social research. It differs from the thesis of the 'displacement' of social research, highlighted above, in at least four ways.

First, to consider the redistribution of social research is to shift attention from the *external* relations of social research to its *internal* relations. The displacement diagnosis posits a fairly strict separation between academic social science and its various outsides – industry, social life, the public. To argue that research capacity is moving away from academia to somewhere else is to accentuate the distinction between academic and other forms of social research. By contrast, a redistribution perspective highlights the contributions of actors inside and outside the university in the production of social research (Adkins and Lury, 2009; Savage *et al.*, 2010).[4] It entails a relatively loose definition of social research, to which various skills and competencies may contribute. Secondly, a redistributive understanding of social research implies a shift in perspective from *ready-made sociology* to *sociology in-the-making*. The digitization of social research, we could say, renders newly relevant a classic insight of the social studies of science and technology: our analysis of knowledge production changes radically as soon as we shift our attention from the status of social research *as a finished product*, to ongoing processes of social research (Latour, 1988).

To conjure up the spectre of the 'corporatization' or 'popularization' or 'democratization' of social research, is to build an argument that derives its normative force from a focus on outcomes. By contrast, if we focus on divisions of labour in digital social research, we explore how digitization may affect and inform the *conduct* of social research, and the normative charge of our exploration here derives from the extent to which these processes are still to a degree undecided, contested, multiple. Thirdly, and relatedly, the notion of redistribution leads us to question the distinction *between the conditions or 'context' of social research and its content.* Debates about the consequences of digitization of social research often concentrate on changes that affect the 'material base' for social research, that is, the technologies and forms of data storage on which it relies. However, of many of the features of digital social research it is actually quite hard to say whether they affect only the conditions or the substance of social research or both or neither: does Twitter research primarily signify a change of conditions in social research, as tweets can be extracted from Twitter so much faster and in quantities that are so much larger than used to be the case in popular discourse analysis (boyd and Crawford, 2011; Leavitt, 2009)? Or does the very meaning of the concept of social discourse change now that we mean by it the broadcasting of one-liners by active individuals in 'real-time' (Niederer and Van Dijck, 2010)?

Fourthly and finally, a focus on redistribution rather than displacement has implications for how we understand our own role as social researchers. That is, the practical or normative roles that we are able to envision for social research, or what we might call their 'scope of intervention', is very different depending on which of the two perspectives we adopt, displacement or redistribution. From a redistributive perspective, the principal question becomes how we may most relevantly intervene in shifting distributions of social research capacity. Here, the main point is not to paint big canvas total pictures of the unlikely future we desire for social research and the likely one that we must fend off. Rather, the question becomes where and how, given the type of redistributions of social research that are currently ongoing, we can most pertinently add a different ingredient that might change the wider mix of social research. A focus on social research *methods* appears to be especially productive in this regard.

The redistribution of social research methods: five views

Method is an important mediator of divisions of labour in social research, and this is no less the case in digital social research. The devising of new research methods, of course, has long been a strategy of choice for those attempting to establish privilege, or claim precedence or newness in science, and digital social research is no exception to this either. As in other fields, debates in social research about methodology have long been a key site and proxy for much more comprehensive controversies about the future direction of the field, with much of the 20th-century methodology contests having been dominated by the pitching of quantitative versus qualitative sociology, with the Positivismusstreit between Karl Popper and Jurgen Habermas as an illustrious example. Methods, then, offer a means to conjure up and establish particular versions of social research, and this in turn tends to involve the attempt to enforce particular divisions of labour in social research. Qualitative social research, for instance, proposes to grant much more initiative to research subjects, while much quantitative research endeavours to create a greater role for standardized tools of data collection, such as the survey, as a way to guarantee the commensurability of data.

In the area of digital social research, methods are invoked to such effects as well.[5] There have been some audacious claims about the opportunities for methodological innovation enabled by online networked media, such as the claim that changing patterns in user activity on the Web may indicate or predict real-time events, like an onslaught of the flu (Rogers, 2009; Mohebbi *et al.*, 2011). And in this context, too, qualitative and quantities methods are pitched against one another, as claims are made back and forth about the relative advantages of, for instance, digital ethnography versus large-scale online survey research (boyd and Crawford, 2011). The Internet has also been said to favour particular social methods over others, such as unobtrusive or non-interventionist methods like content analysis (Lee, 2000; Carslon and Anderson, 2007). Here I cannot do justice to these various methodology debates, but discussions about digital social research methods provide an especially useful prism through which to approach the issue of the redistribution of social research: different views on the implications of digitization for social methods imply very different understandings of what redistributions of research capacity are possible in this context, both empirically and normatively speaking. These views therefore provide a useful starting point for identifying different options in this regard. In this section, I will present some different views on the digitization of two particular methods, network and textual analysis, so as to set the stage for further discussion of the possibilities for intervention in digital social research opened up by the redistribution of methods.

It is possible to order different views on the implications of digitization for social research methods along a spectrum, which starts on one end with a minimal redistribution of research capacity and moves to a maximum redistribution on the other end. The left side of this spectrum is marked by a conservative position that is sceptical about the possibility that social methods are undergoing any significant transformation in digital environments, let alone something like a 'redistribution of methods'. This position, which might be dubbed 'methods-as-usual' can be recognized in an argument recently put forward by the eminent Chicago sociologist Andrew Abbott, who proposed that for anyone who is well versed in social research methods, the newness of the new, online media is very much overstated.[6] Abbott emphasizes that the social methodologies incorporated into digital devices like search engines, most notably network and textual analysis, are pretty standard fare, at least for trained sociologists, and has called the search engine Google 'basically a concordance machine', which matches key-words (queries) to target contexts, and which relies on 'rather routine' additional measures of network analysis, such as in-centrality, to determine the authority of sources; something which has little new to offer to sociologists who have long been familiar with such measures. This view focuses specifically on the formal metrics built into digital devices, and does not consider how these metrics are adapted to or informed by other features of digital devices, as for instance the use of 'live' data or feedback mechanisms. Indeed, it does not really consider the possibility that social research methods may be transformed by virtue of their insertion in a digital networked environment. One could accordingly say that, from this perspective, only one redistribution of research capacity has occurred, in that popular online devices now have social research methods built into them. But on the whole no real redistribution of *methods* is acknowledged: social research methods themselves are not really affected by their uptake in digital online media.

A second view differs significantly from this, and is associated with the new network science informed by mathematics, physics and computing science. This body of work is principally concerned with the opportunities that online media offer for further development of large-scale network and textual analysis, and may accordingly be called 'big methods'. It proposes that digitization has made possible new developments in the *modelling* of networks and textual worlds, and this in large part because of the very large data-sets that digital media technologies make available. The vast databases that have been built over the last decade by search engine companies, gaming industries, Internet service providers and social media platforms create opportunities to significantly expand the analytical and empirical power of network science. They enable the further development of what Duncan Watts and others (Newman *et al.*, 2007) refer to as 'the analysis of real-world network dynamics' (see also Lazer *et al.*, 2009). Contrary to methods-as-usual, this methodological programme can be said to undertake a redistribution of methods of sorts. The new network science namely favours a new set of techniques for data collection and analysis, which entail an unusual division of labour between research subjects, data collection devices, and analysts in social research. To put it somewhat crudely, the approach seeks to maximize the role of mathematical techniques, at the expense of research subjects. In their introduction to the New Network Science, Newman and Watts argue that the social data generated by digital platforms are 'more amenable to the kinds of techniques with which physicists and mathematicians are familiar', and offer a welcome substitute for survey data, and other all too 'social' types of data (Newman *et al.*, 2007).

Proposing this, the new network science reinstates a classic opposition of social research, that between subjective and objective data. Like many others, Newman *et al.* (2007) locate the opportunities that digitization offers for social research in the *type* of data that now become available for social analysis: namely transactional data, which 'record the activities and interactions of the subjects directly' and are thus routinely generated as part of social activities by digital devices, from loyalty cards to search engines (see on this point also Latour, 1998; Rogers, 2009; Savage and Burrows, 2007). Newman *et al.* (2007) give a classic positivist justification for relying on this type of data, arguing that they are much more objective and, as such, offer a welcome substitute for the 'subjective' data generated by surveys, making it possible to avoid reliance on the active contributions of erratic human subjects to data collection.[7] In their account, then, data provided by research subjects are not quite reliable data, something which in their view challenges the validity of network analysis as a whole: 'the respondent data are so contaminated by diverse interpretations of the survey instrument, along with variable recollection or even laziness, that any inferences about the corresponding social network must be regarded with scepticism' (Newman *et al.*, 2007: L-6). Paradoxically, the rise of social media like email, blogs and Facebook here makes possible the *rejection* of user-generated data for purposes of social research, and a redistribution of research capacity towards online registrational devices.

A third and fourth approach are respectively called 'virtual methods' and 'digital methods', and they can be distinguished from the former two in that they are explicitly concerned with the changing relations between social research, its devices and objects in digital online environments. These two approaches offer, however, very different accounts of these changes. The 'virtual methods' programme, developed by Christine Hine (2002, 2005) and others in the early 2000s, focused on the opportunities opened up by the transposition of qualitative social research methods into digital online environments. Its main concern was the digital transformation of *our own sociological methods*, that is, the ways in which methods like discourse analysis and ethnography were and could be transformed by their application in the new context. In focusing mostly on the fate of qualitative methods, Hine's approach to virtual methods makes the opposite manoeuvre from the new network science: it seeks to maximize the role of interpretative subjects in social research, defining the experience of this subject as one of the principal empirical objects of virtual social research. As Hine (2002) puts it: 'ethnographers of the Internet cannot hope to understand the practices of *all* users, but through their own practices they can develop an understanding of what it is to be *a* user' (2002: 54). More generally speaking, the virtual methods approach is concerned with the *digitization* of social research methods, that is, with the translation of methodologies that sociologists define as their own into online environments (Rogers, 2010). This is to recognize a significant but limited redistribution of methods: here, the role of new entities, like web users, in the performance of social method is very much acknowledged, as everyday Internet users are seen to do things online that are similar to fieldwork (taking notes, documenting practice, checking out a strange, new social world). However, such redistributions of social method are here only explored insofar as they affect actors and agencies caught up in the sociologists' research itself: researcher, research subjects, mediating infrastructures, tools used, and so on.

In adopting this strategy, virtual methods do not address the wider issue of the general uptake of social methods in digital online environments, and the consequences of this for the shape and outlook of digital social research. It is this issue that the digital methods programme formulated by Richard Rogers and others (Rogers, 2009) explicitly takes up. This approach proposes that dominant digital devices, search engines chief among them, can be adapted for purposes of social research, and accords to these devices the capacity to inform the development of new methods of social research. Because of their large, dynamic data sets, sophisticated algorithms and feedback possibilities, search engines, Rogers argues, are able to devise forms of social analysis that were not possible before, termed 'natively digital' (see also Weltevrede, n.d.). Digital methods, then, propose that social research should take advantage of the analytic and empirical capacities that are 'embedded in online media'. These can be adapted to purposes of social research, by developing online research tools that run on top of web devices, like Google. The Googlescraper, for instance, adapts Google to conduct work frequency analysis in source sets delineated by the user.[8] This methodological programme of *repurposing* entails a particular redistribution of social research methods, namely towards *devices*: in proposing to adapt existing online devices for purposes of social research, their capacities of data collection, analysis and feedback, come to be incorporated into social and cultural research. As the Digital Methods Initiative proposes to import dominant online tools for data collection, analysis and visualization into social research – or at least parts thereof – devices that constitute the *context* of digital culture come to actively inform the content of social and cultural research.

Arguably, the Digital Methods Initiative more than any other approach discussed above seeks to come to terms with the redistribution of methods in digital environments. Recently, sociologists have recognized that online environments foster a range of tools and practices that qualify as instruments of social research, acknowledging that methods lead a 'social life' online (Savage *et al.*, 2010). But the Digital Methods Initiative proposes an empirical programme that deliberately deploys this circumstance, seeking to render it analytically useful for social research. However, in its above formulation, this approach nevertheless could be said to share a blind spot with the first two approaches already discussed above. Just as with the methods-as-usual perspective and the 'big methods' of the new network science, digital methods can be seen to bracket the issue of the *re-mediation* (Bolter and Grusin, 2000) of social methods in digital online media. As mentioned, Rogers defines the methods enabled by online digital devices as 'natively digital', proposing that they have no clear correlate in the pre- or non-digital world. In making this claim, the DMI programme statement does not really consider, or even downplays, the question of how the uptake of existing social research methods in digital environments entails a *refashioning* of these methods.[9] This question, however, seems to me all too relevant if we want to explore the type of *interventions* that social research becomes capable of in the context of the redistribution of social methods online.

The notion of the 're-mediation of methods' is especially useful, I want to propose here, in that it directs attention to the ways in which prevailing digital devices have methods built into them in which we can recognize those of social research. The foundational article in which Google founders Larry Page and Sergey Brin outlined the central idea behind the new search algorithm, Pagerank, does not only cite a famous sociologist of science, Robert Merton, but it also makes an informed critique of the limitations of sociological forms of network analysis, or as the case may be, citation analysis (Page *et al.*, 1999). Below I will further discuss the particular re-mediation of citation analysis undertaken by Google. Attending to such re-mediations of social methods in the digital context, I want to propose, brings into view a particular mode of intervention for social research itself. Insofar as predominant digital devices apply existing social methods, this may render newly relevant existing sociological *critiques* of these methods. The re-mediation of social methods in the digital context, then, opens up a space of critical intervention for engaged social research. In the remainder of this piece, I will discuss the methodological strategies involved in the development of two digital research tools along these very lines: the Issue Crawler, and an online application of co-word analysis currently under development, the Co-word machine. If we consider how these devices re-mediate social methods, we get an idea of the *digital forms of methodology critique* they enable.

Issue Crawler: from co-citation to co-link analysis

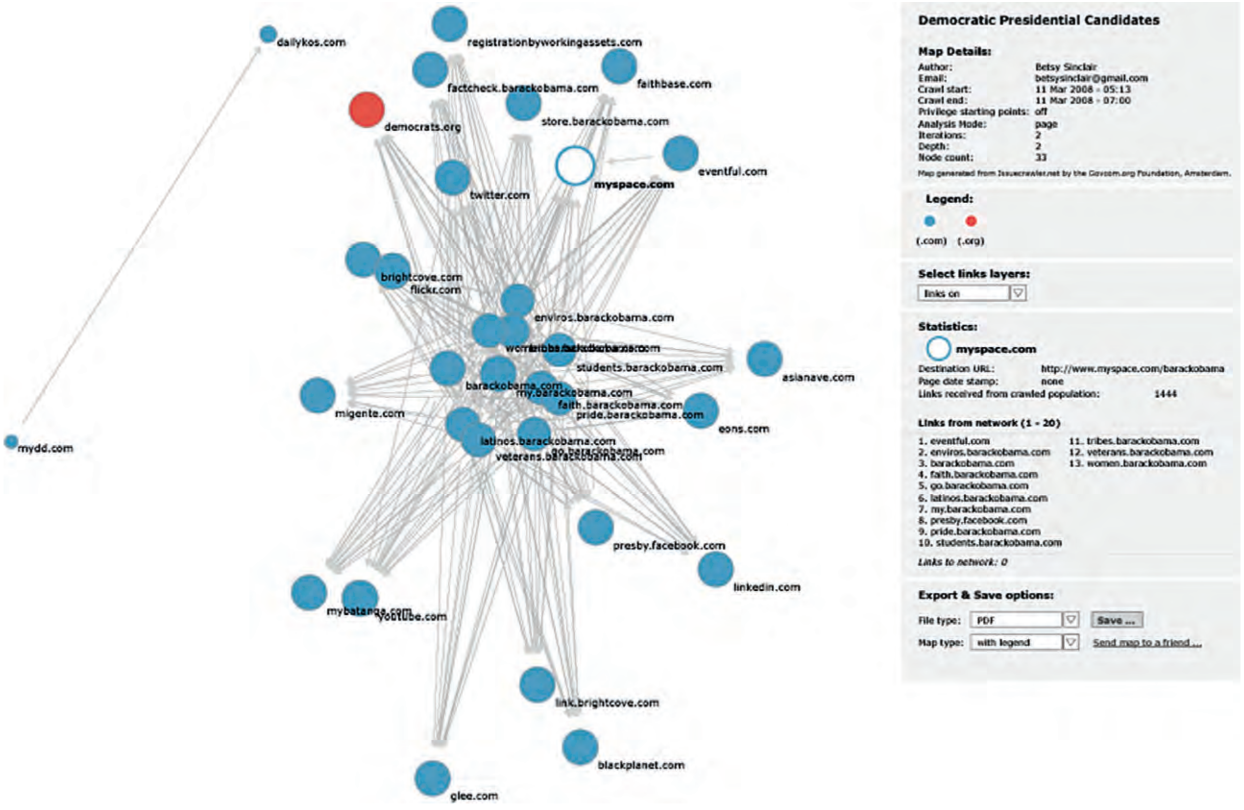
Issue Crawler is an online platform for the location, analysis and visualization of hyperlink networks on the Web. Launched in the early 2000s, Issue Crawler was intended to enable the location and analysis of 'issue networks' on the Web, as it uses hyperlink analysis to delineate sets of pages dealing with a common theme that are connected by hyperlinks (Marres and Rogers, 2008). But the tool has since been used in a variety of projects of online network analysis, including organizational networks (allowing organizations to answer questions such as 'how central are we in this area?') as well as the longitudinal study of online networks, as in the analysis of the rise of Obama and his social media campaign sites in the US democratic election network of 2008 (Borra, 2008; see Figure 3). Using the campaign sites of all democratic presidential candidates as starting points, this last study used Issue Crawler to conduct a series of scheduled crawls, which plotted the emergence of a highly ordered network on the Web, with Obama social media campaign sites dominating the entire network towards the end of the election period. Such network dynamics are arguably Web specific, insofar as the reconfiguration of material network relations can be analysed in real time. But the method on which Issue Crawler relies to demarcate hyperlink networks is based on a classic form of network analysis: co-citation analysis. As an implementation of this specific social research method, the design of Issue Crawler is informed by the context in which the platform was developed.[10]

In the late 1990s, as mentioned above, the rise of the Internet was widely interpreted as an opportunity to apply methods of citation analysis in the new medium, and to adapt this classic method for the analysis of hyperlink structures (Scharnhorst and Wouters, 2006). In this period, the rise of Google and its famous Pagerank algorithm, which relies on in-link measures to rank sources in its query return lists, made newly relevant debates about methods of citation analysis that had been developed by sociologists of science from the 1960s onwards. Larry Page's foundational article makes a specific argument about the re-mediation of citation analysis enabled by the Web, which in his view makes it possible to address *a shortcoming of this method:*

> There has been a great deal of work on academic citation analysis. Goffmann has published an interesting theory of how information flow in a scientific community is an epidemic process. [. . .] But the reason Pagerank is interesting is that there are many cases where simply citation counting does not correspond to our common sense notion of importance. For example, if a web page has a link of the Yahoo home page, it may be just one link but it is a very important one. This page should be ranked higher. Pagerank is an attempt to see how good an approximation to 'importance' can be obtained just from the link structure. (Page *et al.*, 1999)

Arguably, this issue of 'source authority' had already been discussed in citation analysis (MacRoberts and MacRoberts, 1988), and accordingly the degree to which Google's brand of hyperlink analysis contains an actual innovation can be debated. However, as methods of citation analysis were being re-invented as methods of hyperlink analysis, the question was also raised whether and how *critiques* of citation analysis transferred into the online environment. This was – in one of those stories one can tell about tools and methods – the question that Issue Crawler was made to address.

In the 1960s and 70s, sociologists had voiced concerns about citation analysis that now proved all too relevant to the methodological innovation proposed by Google. Citation analysis, it had been argued back then, enables a potentially



Democratic Presidential Candidates

Map Details:

Author:	Betsy Sinclair
Email:	betsysinclair@gmail.com
Crawl start:	11 Mar 2008 - 05:13
Crawl end:	11 Mar 2008 - 07:00
Privilege starting points:	off
Analysis Mode:	page
Iterations:	2
Depth:	2
Node count:	33

Map generated from Issuecrawler.net by the Govcom.org Foundation, Amsterdam.

Legend:

● (.com) ● (.org)

Select links layers:

links on ▽

Statistics:

○ **myspace.com**

Destination URL:	http://www.myspace.com/barackobama
Page date stamp:	none
Links received from crawled population:	1444

Links from network (1 - 20)

1. eventful.com
2. enviros.barackobama.com
3. barackobama.com
4. faith.barackobama.com
5. go.barackobama.com
6. latinos.barackobama.com
7. my.barackobama.com
8. presby.facebook.com
9. pride.barackobama.com
10. students.barackobama.com
11. tribes.barackobama.com
12. veterans.barackobama.com
13. women.barackobama.com

Links to network: 0

Export & Save options:

File type: PDF ▽ Save ...

Map type: with legend ▽ Send map to a friend ...

Figure 3: *Issue Crawler Map: The rise of Obama and Web 2.0 in the Democratic Presidential Candidates network, Betsy SinClair, March 2008*

perverse authority dynamic, according to which well-cited sources get cited more simply because they are already well-cited (Small, 1973; see also Callon *et al.*, 1983).[11] Any large number of citations tends to generate more of them, these now classic critiques proposed, resulting in a situation in which sources are considered authoritative simply by virtue of their authority, and accordingly processes of the valuation of knowledge are captured by social dynamics of popularity, and risk becoming divested from more substantive processes of valuation. This classic critique touched on issues of clear relevance to the new search engine algorithms, like Google's (Marres and Rogers, 2000): these algorithms, too, attributed authority to sources on the basis of the level of recognition implied by the overall number of hyperlinks they received, independent from content. In developing the methodology of Issue Crawler we then drew on this specific methodological critique of 'the authority of authority', in order to develop an alternative approach to hyperlink analysis, one that draws on co-citation analysis (Marres and Rogers, 2000).

In some respects, then, Issue Crawler simply transposed an old methodological solution into a new context. Co-citation analysis was developed in the 1960s as an alternative to the standard citation measure of the overall number of citations received. Rather than seeking to determine the overall authority of individual sources, co-citation analysis seeks to delineate clusters of relevant sources by identifying sources that are jointly linked by other sources. Applying this method to hyperlinks, Issue Crawler sought to introduce a substantive measure of relevance into hyperlink analysis. Issue Crawler deploys the method of co-link analysis in order to undercut the authority effects to which citation and network analysis are vulnerable: instead of assigning value to the overall number of links that sources receive, co-link analysis seeks to locate 'topical clusters' of sources, by identifying co-links in a thematic neighbourhood, or as we called them 'issue networks'. As is clear from the example in Figure 3, Issue Crawler has not necessarily been successful in foregrounding dynamics of relevance at the expense of dynamics of authority. Arguably, indeed, the more insightful issue networks located with Issue Crawler include a clear element of authority, though this is not always the case (Marres and Rogers (2008) discuss some exceptions).

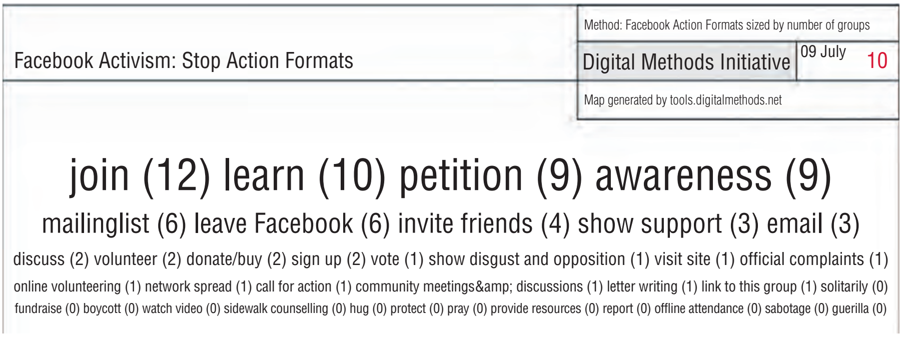
However, it is also important to note that in transposing co-citation analysis onto the Web, Issue Crawler transformed this classic method in some important respects. Before the Web, co-citation analysis was by its very nature limited to the analysis of scientific data-bases, most notably the Science Citation Index. Even as this method sought to challenge authority dynamics, it inevitably rendered itself dependent on institutional demarcations of the relevant fields, in this case scientific fields. For this reason, co-citation could *not* include all the sources to which citations directed it: the scope of its analysis was limited to the sets of sources included in official scientific databases. The Web, by contrast, presents us with *networks* of databases, and as such, it opens up the possibility of analysing a much broader array of sources in real time, generating data-sets that are much more *heterogeneous* than those of citation analysis (Marres and Rogers, 2000; Muniesa and Tchalakov, 2009).[12]

In using co-link analysis to locate thematic networks on the Web, Issue Crawler does not only transpose a particular method into the online environment, but also a specific *methodology critique*. In advocating co-citation analysis, sociologists did not only seek to address a problem with methods of citation analysis in themselves, or with questionable citation behaviours, whereby sources mainly recognize already authoritative sources, thus aggravating the popularity effect. In the pre-digital context, critics of citation analysis specifically targeted the ways in which citation analysis *amplified* these popularity effects: their concern was that science policy would increasingly rely on these methods, as research councils took up citation measures, in their attempt to render their modes of assessment more evidence-based (Leydersdorff, 1998). Similarly, the issue with search engines is not just that, in applying measures of in-link centrality, they help to generate more authority for already authoritative sources.[13] At issue is a whole complex of behaviours: by privileging sources with a high overall in-link count, search engines encourage linking behaviours that consolidate authority dynamics, and the modification of user trajectories to a similar effect (Introna and Nissenbaum, 2000; Vaidhyanathan, 2011). In networked environments, then, it is especially obvious that *multiple* agencies have a part to play in the enactment of 'social methods'.[14] To put it differently, in the digital context social methods must clearly be defined as a *distributed* accomplishment, and our attempts to intervene critically in this context must be informed by this circumstance.

The co-word machine: from co-word analysis to online issue profiling

In questioning the dominance of authority dynamics on the Web, back in the late 1990s, and the role of devices like Google in enabling this, however, I clearly had little idea of what we were up against. In retrospect it can seem naive to expect that a methodology developed by a minoritarian movement in the sociology of science, like co-citation analysis, could be rendered effective in digital networked spaces, which were just then emerging as key hubs of the global information economy. Indeed, recent developments in this area, most notably the rise of social media platforms like Facebook and Twitter, can be taken as evidence that the medium has gone 'the other way'. Reputational dynamics, whereby things become more widely liked by virtue of being liked, have become very much the currency of online media (Onnela and Reed-Tsochas, 2010; Gerlitz and Helmond, 2012). The social network, in which actor-alliances are formed largely independent from content – and *not* the 'issue network', with its topical dynamics of the thematic clustering of sources – has become the key organizational form associated with the Internet (for the distinction between issue- and actor-network, see Marres and Rogers, 2008). However, social media platforms also highlight the limits of our earlier argument in another, less ironic sense: social media have proven that networks driven by reputational logics are very well capable of organizing content, in ways that do *not* necessarily reproduce 'the tyranny of reputation'. The rise of these platforms has been accompanied by the proliferation of tools for the analysis and visualization of substantive dynamics. Figure 4, for example, provides a word frequency analysis of action terms on Facebook, showing the relative prominence of such terms in a selection of Facebook groups.

Social media, then, have proven to be no less adaptable to the purposes of content analysis than social network analysis. Nevertheless, I think that our initial intuition still holds: online digital environments are in need of alternative measures that can provide a counter-weight to dominant popularity metrics. On closer inspection, many current instruments of online content analysis, like tag clouding, have not really attenuated authority effects. They tend to rely on versions of the 'overall citation count' too: they bring into view what (or who) is most mentioned, followed, liked and so on, in a given data set at a given moment. Tag clouds, and other online applications of textual analysis and visualization perpetuate the preoccupation with the most cited or most popular, and these instruments can arguably be said to reproduce the authority effect in another form. After the rise of social media, the question then remains how to develop alternatives to reputational measures: the question is still that of how to move beyond 'purely social' mechanics of authority, popularity or celebrity, and get to more substantive dynamics of relevance. But in this context, too, existing sociological critiques of research methods may offer a useful resource: debates about the majoritarian bias in textual analysis, and the development of alternative forms of 'discourse analysis' have been ongoing in sociology for several decades. Here I would like to single out one such alternative method, namely *co-word analysis*, as this method was explicitly developed by sociologists



Facebook Activism: Stop Action Formats

Method: Facebook Action Formats sized by number of groups

Digital Methods Initiative

09 July 10

Map generated by tools.digitalmethods.net

join (12) learn (10) petition (9) awareness (9)

mailinglist (6) leave Facebook (6) invite friends (4) show support (3) email (3)

discuss (2) volunteer (2) donate/buy (2) sign up (2) vote (1) show disgust and opposition (1) visit site (1) official complaints (1)

online volunteering (1) network spread (1) call for action (1) community meetings&amp; discussions (1) letter writing (1) link to this group (1) solitarily (0)

fundraise (0) boycott (0) watch video (0) sidewalk counselling (0) hug (0) protect (0) pray (0) provide resources (0) report (0) offline attendance (0) sabotage (0) guerilla (0)

Figure 4: *Tag cloud analysis, Facebook is for joiners*
Source: Lonneke van der Velden and Clare Lee, Project Facebook, DMI Summerschool, July 2010 (https://wiki.digitalmethods.net/Dmi/TrainingProgramProjectFacebook).

of science and technology to enrich citation analysis and possibly by extension hyperlink analysis.

Co-word analysis was devised in the 1980s by the actor-network theorist Michel Callon and others as a way to expand the project of co-citation analysis. It was developed to locate 'pockets of innovation' in science, using textual analysis to locate especially active thematic clusters of sources in the scientific literature.[15] Co-word analysis did this by measuring the rise and fall of keywords, and the associations among them, in a corpus of scientific articles (Callon *et al.*, 1983; Whittaker, 1989). Analysing the keywords used to index articles in scientific databases, co-word analysis offered a way to determine which were the most 'active' key words, and word associations in the corpus. It provided a way to measure which keywords and keyword associations *varied* significantly in their mentioning and relations over a given period. In trying to determine the most 'happening' themes, this method was expressly designed to locate 'buzz' or 'live content' in the scientific literature, *but it did this without relying on popularity dynamics*. Indeed, terms that were mentioned with a constantly high frequency were automatically deleted from the set of active terms: the key indicator was not frequency of mentioning but *variation* in mentioning (and association) (Callon *et al.*, 1983).

In recent years, the method of co-word analysis has been transposed onto the Web, with various online applications deploying the methodology to visualize word associations in online data sets, such as corpi of email messages or twitter messages (Danowksi, 2009; www.infomous.com). In the online context, co-word analysis promises to offer an alternative to word frequency analysis, the method of which it seems fair to say spread like wildfire, also into the social sciences, on the back of tag clouding tools.[16] Co-word analysis determines the relevance of terms by measuring the strength and intensity of relations among them: only words that appear frequently *and* that appear together make it onto co-word maps. Co-word analysis, as mentioned, tries to purge its analysis of terms that are merely popular: it excludes terms that appear frequently but in random association with others. For this reason, co-word analysis seems to provide an alternative to the majoritarian logics of word frequency, which make a term appear larger and more visible the more often it appears. The method may help us move beyond the ranking or hit list, that most visible testimony to the tyranny of reputation, as is evidenced by the alternative visual format proposed by Callon and colleagues for co-word analysis (see Figure 5). Crucially, moreover, online co-word analysis does away with popularity without sacrificing *liveness,* or rather liveliness. Co-word analysis, too, aims to deliver the most *happening* content (see also Marres and Weltevrede, forthcoming). But it does this by deploying an alternative measure: not the safety of a large number of mentionings, but fluctuations in the presence of words and word associations is key.

Together with colleagues in Amsterdam, we are now working to develop a Co-Word Machine that deploys co-word analysis for the online location and visualization of 'issue language'. In transposing co-word analysis into the online

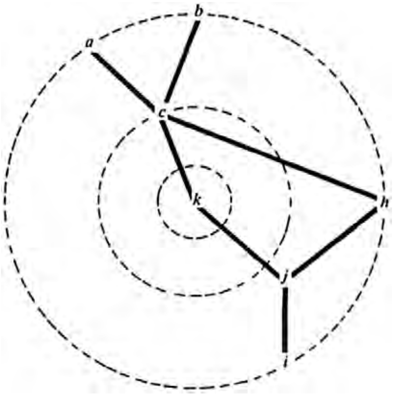


Figure 5: *Co-word visualization*
Source: Callon *et al.* (1983).

context, however, a number of issues arise which may either weaken or strengthen the analytical and critical capacities of this method, depending on how we deal with them, and how they will play out. First of all, online environments offer an opportunity which adherents of co-word analysis could only theorize in the 1980s. For Callon and his colleagues, the chief attraction of co-word analysis was its promise to help advance 'our search for the trans-disciplinary, trans-institutional problematic networks that we want to identify' (Callon *et al.*, 1983: 196) However, in the 1980s co-word analysts were frustrated in this project by the limits of the databases and genres to which they applied their method. As in the case of co-citation analysis, co-word analysis relied on scientific databases, and because the genre of the scientific article was so different from those current in other fields (the policy report, the newspaper article, the petition), there was no reliable way to track terms *across* discursive spheres. Online networked media provide a great opportunity to address this limitation, as one distinctive feature of these media is precisely the significant genre contamination across fields (which organization does not have a blog?). In this environment, co-word analysis, too, may be applied to far more *heterogeneous* data-sets (Marres and Rogers, 2000).

However, the Web also poses some serious challenges for co-word analysis, among others because of the widely divergent ways of indexing content prevalent in the medium. In this respect at least, classic co-word analysis had it easy, as it could rely on professional indices – keywords used by institutions like the Science Citation Index to index scientific articles – to locate emergent vocabularies. In online media, most applications rely on self-indexing – on keywords, or tags, provided by users marking up self-generated content. This inevitably raises issues of reliability and comparability, and in this respect, digital tagging practices drive home a basic but important point made by the American journalist Walter Lippmann (1997 [1922]) in his classic analysis of newspapers: any factual report is only as good as the sources from which it derives its findings (such as the National Office of Statistics). In this respect, co-word analysis certainly is *not* free of the problems associated with digital devices like tag clouding, which, as the name says, tend to rely on tags used to mark up online content, by bloggers and other users. In the case of co-word analysis as well, our results will only be as good as the classificatory practices on which we rely. We are returning, then, to the issue of the distributed accomplishment of digital methods: online textual analysis is likely to rely on the contributions of a whole host of agents, from the availability of tagging features, to the taggers who actually mark up online content, the analytical instruments used to analyse these tags, visualization modules, and so on. In order to intervene relevantly in online social research, we do well to recognize such assemblages of users, devices and informational practices, as the relevant unit of 'methodological innovation' in social research.

Conclusion

In online environments, the distributed nature of social research is especially hard to deny. User behaviours, information formats and digital devices that are embedded in the medium are likely to leave an imprint on social analysis. Something that applies to other research practices too is then rendered explicit in online social research: here, social research is *noticeably* marked by informational practices and devices not of its own making, from the analytic measures built into online platforms (eg numbers of links, number of mentionings, follower counts), to the visual forms embedded in visualization modules (the tag cloud). Online social research is visibly a distributed accomplishment. This circumstance, I have argued, does not only pose problems for social research but also offers opportunities for the development of social research methods. Digitization enables a broadening of the agencies playing an active role in the enactment of social methods, broadly conceived: in this context, a wide range of actors including platform users and analytic devices like search engines come to play a part in the collection, analysis and presentation of social data. And this redistribution of methods in digital social research opens up a space of intervention for social research.

Social methods, I have argued, are a key instrument with which wider divisions of roles in social research are being curated in online environments. Prominent digital devices like Google and Twitter and Facebook, and the users and developers enrolled by them, today actively inform the enactment of social methods online. The types of data platforms make available, the measures and formats on which they rely in communicating this data (rankings, follower counts and clouds), and the wider informational practices in which they are taken up (Facebook members visualizing the network of their Facebook friends): all of these elements inform the performance of 'social methods' in digital networked environments. The contours of these 'methodological spaces online' are not necessarily easy to determine, as platform settings change, and users change their allegiance to a new device. However, these assembled devices, settings and actors open up a particular space of intervention in digital social research: if specific digital social methods are a distributed accomplishment – such as the 'overall citation count' that is materialized in Google and other platforms – then sociological research may seek to intervene in the relations among entities that sustain these methods, by proposing alternative methods and distributions thereof. Web-based applications of co-link analysis and co-word analysis, the Issue Crawler and the Co-word machine currently under development, aim to do just this. In so doing, they extend some of the long-standing normative projects of sociological research into digital environments, such as the commitment to methods that privilege substantive dynamics of relevance over purely social or reputational ones, or what we could call 'post-social methods'.

As noted, there is a strong tradition in social research of seeking to bracket the effects of the methods deployed by 'the social actors themselves': many social researchers have become experts in devising tricks that make it possible to ignore the active contribution of research subjects to the organization of data and the framing of methods. But digital networked environments provide opportunities to explore different possible approaches to the distributed nature of social research and its methods. As online social research forces us to acknowledge the contributions of digital devices, practices and subjects, to the enactment of social research, it can be taken as an invitation to move beyond 'proprietary' concepts of methods, that is, beyond the entrenched use of method as a way to monopolize the representation of a given field or aspect of social reality. A redistributive approach to social research redefines methods as involving the combination and coordination of diverse competencies: classification, visual design, automated analysis, and so on. Behind debates about the unreliability of data generated by research subjects, and the 'mess' of self-indexed online content, there lies a debate about the redistribution of methods between researchers, devices, information and users, in online environments. Which is also to say, the debate about the digitization of social methods is perhaps most productively approached as a debate about *participatory* research methods.

Acknowledgements

I would like to thank Les Back, Erik Borra, Sabine Niederer, Nirmal Puwar, Richard Rogers, Emma Uprichard and Esther Weltevrede for their comments on earlier versions of this article.

Notes

[1] Both of these methods have been central to the development of actor-network theory and in focusing on the re-mediation of these methods, I am also exploring how online research

tools translate methods of actor-network theory into networked digital media. In doing so, I will join others in arguing that digitization offers opportunities for a generalization of this sociological research programme (Latour, 1998; Law, 2008; Savage, 2010; Latour *et al.*, 2012).

[2] 'New Yahoo! Clues Launches', posted 29 June 2011, http://www.ysearchblog.com/2011/06/29/new-yahoo-clues-launches/

[3] See http://twitterabused.com/2012/02/09/visualising-twitter-networks-john-terry-captaincy-controversy/

[4] The notion of the redistribution of social research in the digital context is both inspired by and deviates from the idea of the double social life of methods proposed by Savage *et al.* (2010). Whereas the latter proposes that social research methods are both deployed in social science and in society at large – as for instance 'the survey' – the idea of the redistribution of research directs our attention to shifting relations *between* agencies inside and outside the university.

[5] One redistributive issue requires special attention: digital social research entails a reshuffling of roles between human and technical elements, and as such it raises the question of which delegations of roles to new actors or devices are exactly occurring, and what their implications are for the analytic and empirical capacities of social and cultural research (Niederer and van Dijck, 2010; see also Bach and Stark, 2005).

[6] Andrew Abbott, 'Googles of the Past: Do Keywords Really Matter?', lecture, Department of Sociology, Goldsmiths, 15 March 2011.

[7] They note: 'For most of the past fifty years, the collection of network data has been confined to the field of social network analysis, in which data have to be collected through survey instruments that not only are onerous to administer, but also suffer from the inaccurate or subjective responses of subjects. People, it turns out, are not good at remembering who their friends are, and the definition of a 'friend' is often quite ambiguous in the first place' (Newman *et al.*, 2007: L-5).

[8] https://tools.issuecrawler.net/beta/scrapeGoogle/

[9] This notion of re-mediation was put forward by Bolter and Grusin (2000) in an effort to shift the debate about digital culture beyond yes/no exchange which pitched two sterile positions against one another: either new media merely offered old culture in a new jacket, or they enabled the invention of radically new forms of culture. Rejecting both positions, Bolter and Grusin proposed to focus instead on how older cultural forms underwent a process of *refashioning* in new media. I am proposing here that this notion can be usefully adapted to make sense of the digital social research methods.

[10] Issue Crawler was developed between 1999 and 2002 by the govcom.org foundation in Amsterdam, which is directed by Richard Rogers and of which the author was a founding member. www.govcom.org.

[11] This dynamic is in some ways similar to a classic sociological problematic, discussed by Tocqueville, of 'the tyranny of reputation'. According to this wider dynamic, ideas gain influence for the reason of being well regarded, a circular dynamic in which substantive considerations of the ideas in question do not necessarily enter.

[12] In some sense, online hyperlink analysis enabled a move beyond the database. In this respect, the technique of crawling the Web allows for a renewed engagement with a classic sociological concern of actor-network theory: the issue of the pre-ordering of data, as what prevents sociology from engaging with heterogeneous ontologies.

[13] Issue Crawler also engages with issues which in retrospect we can designate as issues of public sociology: its methodology concentrates on a publically accessible metric, hyperlinks, and the Issue Crawler archive of all located networks, dating back to 2001, is available to all users.

[14] Issue Crawler also seeks to put this situation to positive effect. The quality of its network maps depends on the knowledge implied in the hyperlinks that it analyses: Issue Crawler can only provide us with 'telling networks', if sources in the network link intelligently, ie if they identify issue-protoganists and alliances among them by way of hyperlinks.

[15] More specifically, co-word analysis was developed as a way to deal with the problem that co-link analysis reproduced a reputational logic in spite of itself. This problematics is all too relevant

in relation to Issue Crawler: this platform too can be said to reproduce popularity and authority effects, for various reasons: because of its demarcationist approach, because of hyperlinking reproducing authority effects, and because platform users want to know 'who is the most popular source'. In this respect, the reproduction of reputational dynamics by Issue Crawler is itself partly a social effect, ie it is a consequence of the distributed nature of digital social research: the effect can partly be traced back to 'reputational linkers', and the research agendas of the users of Issue Crawler.

[16] There are a number of related tools for visualizing word frequency analysis, like the Dorling visualization, and one of my favourites, the Bubble line.

References

Adkins, L. and Lury, C., (2009), 'Introduction to special issue "What is the empirical?" ', *European Journal of Social Theory*, 12: 5–20.

Bach, J. and Stark, S., (2005), 'Recombinant technology and the new geography of association', in R. Latham and S. Sassen (eds), *Digital Formations: IT and the New Global Realm*, 37–53, Princeton, NJ: Princeton University Press.

Back, L., (2010), 'Broken devices and new opportunities': re-imagining the tools of qualitative research', NCRM Working Paper Series, 08/10.

Bolter, J. and Grusin, R., (2000), *Remediation: Understanding New Media*, Cambridge: MIT Press.

Borra, E., (2008), 'The web as anticipatory medium, blogpost', http://erikborra.net/blog/2008/12/the-web-as-an-anticipatory-medium/#more-709.

boyd, d. and Crawford, K., (2011), 'Six provocations for big data', paper presented at Oxford Internet Institute's *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society,* 1–17.

Button, G., (1991), 'Introduction', in G. Button (ed.), *Ethnomethodology and the Human Sciences: A Foundational Reconstruction*, 1–9, Cambridge: Cambridge University Press.

Callon, M., Courtial, J., Turner W. and Bauin, S., (1983), 'From translations to problematic networks: an introduction to co-word analysis', *Social Science Information*, 22: 191–235.

Callon, M., Lascoumes, P. and Barthe, Y., (2009 [2001]), *Acting in an Uncertain World: An Essay on Technical Democracy*, Cambridge: MIT Press.

Carlson, S. and Anderson, B., (2007), 'What are data? The many kinds of data and their implications for data re-use', *Journal of Computer-Mediated Communication*, 12 (2), http://jcmc.indiana.edu/vol12/issue2/carlson.html

Danowski, J., (2009), 'Network analysis of message content', in K. Krippendorff and K. Bock (eds), *The Content Analysis Reader*, 421–430*,* Thousand Oaks, CA: Sage Publications.

Fielding, N., Lee, R. and Blank, G., (2008), 'The Internet as research medium: an editorial introduction to *The Sage Handbook of Online Research Methods*', in N. Fielding, R. Lee and G. Blank (eds), *The Sage Handbook of Online Research Methods*, 3–20, London: Sage.

Gerlitz, C. and Helmond, A., (2012), 'The like economy: social buttons and the data-intensive Web', *New Media and Society*, forthcoming.

Grandclément, C. and Gaglio, G., (2010), 'Convoking the consumer in person: the focus group effect', in D. Zwick and J. Cayla (eds), *Inside Marketing: Practices, Ideologies, Devices*, Oxford and Cambridge, MA: Oxford University Press, http://dx.doi.org/10.1093/acprof:oso/9780199576746.003.0005

Haraway, D., (1991), 'Cybermanifesto: science, technology and socialist-feminism in the late 20th century', in *Cyborgs, Simians and Women: The Reinvention of Nature*, 149–182, New York: Routledge.

Hine, C., (2002), *Virtual Ethnography*, London: Sage.

Hine, C. (ed.), (2005), *Virtual Methods: Issues in Social Research on the Internet*, Oxford: Berg.

Hubble, N., (2006), *Mass Observation and Everyday Life: Culture, History, Theory*, Basingstoke: Palgrave Macmillan.

Introna, L. and Nissenbaum, H., (2000), 'Shaping the Web: why the politics of search engines matters', *The Information Society*, 16 (3): 1–17.

Latour, B., (1988), *The Pasteurization of France*, trans. A. Sheridan and J. Law, Cambridge, MA: Harvard University Press.

Latour, B., (1998), 'Thought experiments in social science: from the social contract to virtual society', *1st Virtual Society? Annual Public Lecture*, http://www.artefaktum.hu/it/Latour.htm

Latour, B., (2005), *Reassembling the Social: An Introduction to Actor-network-theory*, Oxford: Oxford University Press.

Latour, B., Jensen, P., Venturini, T. and Boullier, D., (2012), 'The whole is always smaller than its parts: a digital test of Gabriel Tarde's Monads', *The British Journal of Sociology*, forthcoming.

Law, J., (2004), *After Method: Mess in Social Science Research*, London and New York: Routledge.

Law, J., (2008), 'On STS and sociology', *The Sociological Review,* 56 (4): 623–649.

Law, J., (2009), 'Assembling the world by survey: performativity and politics', *Cultural Sociology,* 3 (2): 239–256.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M., (2009), 'Computational social science', *Science*, 6 February, 5915 (323): 721–723.

Leavitt, A. (ed.), (2009), 'The Iranian election on Twitter: the first eighteen days', *Web Ecology Project*, http://www.webecologyproject.org/2009/06/iran-election-on-twitter/ (5 March 2011).

Lee, R. M., (2000), *Unobtrusive Methods in Social Research*, Buckingham: Open University Press.

Leydersdorff, L., (1998), 'Theories of citation?', *Scientometrics*, 43 (1): 5–25.

Lezaun, J., (2007), 'A market of opinions: The political epistemology of focus groups', *Sociological Review*, 55: 130–151.

Lippmann, W., (1997 [1922]), *Public Opinion*, New York: Free Press Paperbacks, Simon & Schuster.

Lury, C., (1996), *Consumer Culture*. Cambridge: Polity Press.

Lury, C., (2004), *Brands: The Logos of the Global Cultural Economy*, New York and London: Routledge.

MacRoberts, M. H. and MacRoberts, B. R., (1988), 'Problems of citation analysis: a critical review', *Journal of the American Society for Information Science*, 40 (5): 342–349.

Madsen, A. K., (2012), 'Web visions as controversy lenses', in A. Carusi, A. Sissel Hoel and T. Webmoor (eds), Special Issue on Computational Picturing, *Interdisciplinary Science Reviews*, 37 (1).

Marres, N., (2011), 'The cost of involvement: everyday carbon accounting and the materialization of participation', *Economy and Society*, 40 (4): 510–533.

Marres, N., (2012), 'The experiment in living', in C. Lury and N. Wakeford (eds), *Inventive Methods: The Happening of the Social*, 76–95, London: Routledge.

Marres, N. and R. Rogers (2000), 'Depluralising the Web and repluralising public debate: the case of the GM food debate on the Web', in R. Rogers (ed.), *Preferred Placement: Knowledge Politics on the Web*, 113–135, Maastricht: Jan van Eyck Editions.

Marres, N. and Rogers, R., (2008), 'Subsuming the ground: how local realities of the Ferghana Valley, the Narmada Dams, and the BTC pipeline are put to use on the Web', *Economy and Society*, 37 (2): 251–281.

Marres, N. and Weltevrede. E., (forthcoming), 'Scraping the social? Issues in real-time research', *Journal of Cultural Economy*.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H. and Kumar, S., (2011), 'Google Correlate Whitepaper', http://correlate.googlelabs.com/whitepaper.pdf

Muniesa, F. and Tchalakov, I., (2009), 'What do you think a simulation is, anyway? A topological approach to cultural dynamics', Working Paper, http://www.atacd.net/

Newman, M., Barabási, A. and Watts, D., (2007), *The Structure and Dynamics of Networks*, Princeton, NJ: Princeton University Press.

Niederer, S. and Van Dijck, J., (2010), 'Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system', *New Media and Society*, 12 (8): 1368–1387.

Onnela, J.-P. and Reed-Tsochas, F., (2010), 'Spontaneous emergence of social influence in online systems', *Proceedings of the National Academy of Sciences*, 107 (43): 18375–18380.

Page, L., Brin, S., Motwani, R., and Winograd, T., (1999), 'The PageRank Citation Ranking: bringing order to the Web. Technical Report', Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

Rheinberger, H.-J., (1997), *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*, Stanford, CA: Stanford University Press.

Rogers, R., (2009), *The End of the Virtual: Digital Methods*, Amsterdam: Amsterdam University Press.

Rogers, R., (2010), 'Internet research: the question of method', *Journal of Information Technology and Politics*, 7 (2/3): 241–260.

Savage, M., (2010), *Identities and Social Change in Britain since 1940: The Politics of Method*, Oxford: Oxford University Press.

Savage, M. and Burrows, R., (2007), 'The coming crisis of empirical sociology', *Sociology*, (41): 885–899.

Savage, M., Law, J. and Ruppert, E., (2010), 'The double social life of methods', CRESC Working Paper Series. No. 95, http://www.cresc.ac.uk/our-research/cross-theme-research/social-life-of-methods

Scharnhorst, A. and Wouters, P., (2006), 'Web indicators – a new generation of S&T indicators?' *Cybermetrics*, 10 (1).

Small, H., (1973), 'Co-citation in the scientific literature: a new measure of the relationship between two documents', *Journal of the American Society for Information Science*, 24 (4): 265–269.

Uprichard, E., Burrows, R. and Byrne, D., (2008), 'SPSS as an "inscription device": from causality to description?', *The Sociological Review*, 56 (4): 606–622.

Vaidhyanathan, S., (2011), *The Googlization of Everything (And Why We Should Worry)*, Berkeley, CA: University of California Press.

Weltevrede, E., (n.d.), 'Studying society, not Google: repurposing Google for social and cultural research', Department of Media Studies, University of Amsterdam, ms.

Whatmore, S. J., (2009), 'Mapping knowledge controversies: science, democracy and the redistribution of expertise', *Progress in Human Geography*, 33 (5): 587–598.

Whittaker, J., (1989), 'Creativity and conformity in science: titles, keywords, and co-word analysis,' *Social Science in Science*, 19: 473–496.

Woolgar, S., (2002), 'Introduction: five rules of virtuality', in S. Woolgar (ed.), *Virtual Society? Technology, Cyberbole, Reality*, Oxford: Oxford University Press.

Alexander von Humboldt
INSTITUT FÜR INTERNET
UND GESELLSCHAFT

HIIG Discussion Paper Series

Discussion Paper No. 2013-01

The Politics of Twitter Data[1]

23. Jan. 13

Cornelius Puschmann
cornelius.puschmann@oii.ox.ac.uk
Oxford Internet Institute (OII)
University of Oxford

Jean Burgess
je.burgess@qut.edu.au
ARC Centre of Excellence for Creative Industries and Innovation (CCI)
Queensland University of Technology

[1] This paper is a draft chapter from the forthcoming volume *Twitter and Society* (K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann, eds.) which will be available from Peter Lang Publishers, NYC, in spring 2013.

Abstract

Our paper approaches Twitter through the lens of “platform politics” (Gillespie, 2010), focusing in particular on controversies around user data access, ownership, and control. We characterise different actors in the Twitter data ecosystem: private and institutional end users of Twitter, commercial data resellers such as Gnip and DataSift, data scientists, and finally Twitter, Inc. itself; and describe their conflicting interests. We furthermore study Twitter’s Terms of Service and application programming interface (API) as material instantiations of regulatory instruments used by the platform provider and argue for a more promotion of data rights and literacy to strengthen the position of end users.

Keywords

Social media, Twitter, big data, users, platforms, regulation

Contents

1. The Big Data Moment

> *[D]ata is not free, and there's always someone out there that wants to buy it. As an end-user, educate yourself with how the content you create using someone else's service could ultimately be used by the service-provider. (Jud Valeski, CEO of Gnip, quoted in Steele, 2011, para 19)*

> *There are significant questions of truth, control, and power in Big Data studies: researchers have the tools and the access, while social media users as a whole do not. Their data were created in highly context-sensitive spaces, and it is entirely possible that some users would not give permission for their data to be used elsewhere. (boyd & Crawford, 2012, p. 12)*

Talk of Big Data seems to be everywhere. Indeed, the apparently value-free concept of 'data' has seen a spectacular broadening of popular interest, shifting from the dry terminology of lab coat-clad scientists to the buzzword *du jour* of marketers. In the business world, data is increasingly framed as an economic asset of critical importance, a commodity en par with scarce natural resources (Backaitis, 2012; Rotella, 2012), while in context with "open" public sector data there is a growing debate about digital information as an enabler of growth, transparency, and civic engagement.

It is social media that has most visibly brought the Big Data moment to media and communication studies, and beyond it, to the social sciences and humanities. Social media data is one of the most important areas of the rapidly growing data market (Manovich, 2012; Steele, 2011). Massive valuations are attached to companies that directly collect and profit from social media data, such as Facebook and Twitter, as well as to resellers and analytics companies like Gnip and DataSift. The expectation attached to the business models of these companies is that their privileged access to data and the resulting valuable insights into the minds of consumers and voters will make them irreplaceable in the future. Analysts and consultants argue that advanced statistical techniques will allow the detection of on-going communicative events (natural disasters, political uprisings) and the reliable prediction of future ones (electoral choices, consumption).

These predictions are made possible through cheap networked access to cloud-based storage space and processing power, paired with advanced computational techniques to investigate complex phenomena such as language sentiment (Thelwall, Buckley, & Paltoglou, 2011; Thelwall, to appear), communication during natural disasters (Sakai, Okazaki, & Matsuo, 2010), and information diffusion in large networks (Bakshy, Rosenn, Marlow, & Adamic 2012). Such methods are hailed as superior tools for the accurate modelling of social processes and have a growing base of followers among the proponents of "digital methods" (Rogers, 2009) and "computational social science" (Lazer et al., 2009). While companies, governments, and other stakeholders previously had to rely on vague forecasts, the promise of these new approaches is ultimately to curb human unpredictability through information. The traces created by the users of social media platforms are harvested, bought, and sold; as an entire commercial ecosystem is forming around social data, with analytics companies and services at the helm (Burgess & Bruns, 2012; Gaffney & Puschmann, to appear).

Yet, while the data in social media platforms is sought after by companies, governments and scientists, the users who produce it have the least degree of control over "their" data. Platform providers and users are in a constant state of negotiation regarding access to and control over information. Both on Twitter and on other platforms, this negotiation is conducted with contractual and technical instruments by the provider and with ad-hoc activism by some users.

The complex relationships among platform providers, end users, and a variety of third parties (e.g., marketers, governments, researchers) further complicates the picture. These nascent conflicts are likely to deepen in the coming years, as the value of data increases while privacy concerns mount and those without access feel increasingly marginalised.

Our paper approaches Twitter through the lens of "platform politics" (Gillespie, 2010), focusing in particular on controversies around user data access, ownership, and control. We characterise different actors in the Twitter ecosystem: private and institutional end users of Twitter, commercial data resellers such as Gnip and DataSift, data scientists, and finally Twitter, Inc. itself; and describe their conflicting interests. We furthermore study Twitter's Terms of Service and application programming interface (API) as material instantiations of regulatory instruments used by the platform provider and argue for a more promotion of data rights and literacy to strengthen the position of end users.

2. Twitter and the Politics of Platforms

The creation of social media data is governed by an intricate set of dynamically shifting and often competing rules and norms. As business models change, the emphasis on different affordances of the platform changes, as do the characteristics of the assumed end user under the aspects of value-creation for the company. Twitter has been subject to such shifts throughout its brief history, as the service adapts to a growing user community with a dynamic set of needs.

In this context, there has been a recent critique of a perceived shift from an 'open' Internet (where open denotes a lack of centralised control and a divergent, rather than convergent, software ecosystem), toward a more 'closed' model with fewer, more powerful corporate players (Zittrain, 2008). Common targets of this critique include Google, Facebook, and Apple, who are accused of monopolising specific services and placing controls on third-party developers who wish to exploit the platforms or contribute applications which are not in accordance with the strategic aims of the platform providers. In Twitter's case, the end of the Web 2.0 era, supposedly transferring power to the user (O'Reilly, 2005), is marked by the company's shift to a more media-centric business model relying firstly on advertising and corporate partnerships and, crucially for this paper, on reselling the data produced collectively by the platform's millions of users (Burgess & Bruns, 2012; van Dijck, 2012). This shift has been realised materially in the architecture of the platform—including not only its user interface, but also the affordances of its API and associated policies, affecting the ability of third-party developers, users, and researchers to exploit or innovate upon the platform.

There have been several recent controversies specifically around Twitter data access and control:

- the increasing contractual limitations placed on content through instruments such as the Developer Display Requirements (Twitter, 2012c), that govern how tweets can be presented in third-party utilities, or the Developer Rules of the Road (Twitter, 2012b), that forbid sharing large volumes of data;
- the requirement for new services built on Twitter to provide benefits beyond the service's core functionality;
- actions against platforms which are perceived by Twitter to be in violation of these rules, e.g. Twitter archiving services such as 140Kit and Twapperkeeper.com, business analytics services such as PeopleBrowsr, and aggregators like IFTTT.com;
- the introduction of the Streaming API as the primary gateway to Twitter data, and increasing limitation placed on the REST API as a reaction to growing volumes of data generated by the service;
- the content licensing arrangements made between Twitter and commercial data providers Gnip and Datasift (charging significant rates for access to tweets and other social media content); and
- the increasing media integration of the service, emphasizing the role of Twitter as "an information utility" (Twitter co-founder Jack Dorsey, quoted in Arthur, 2012).

In the following, we relate these aspects to different actors with a stake in the Twitter ecosystem.

3. Conflicting Interests in the Twitter Ecosystem

Lessig (1999) names four factors shaping digital sociotechnical systems: the market, the law, social norms, and architecture (code and data). The regulation of data handling by the service provider through the Terms of Service and the API is of particular interest in this context. As outlined above, Twitter seeks to regulate use of data by third parties through the Terms and the API, assigning secondary roles to the law (which the Terms frequently seek to extend) and social norms (which are inscribed and institutionalised in various ways through both the interface and widespread usage conventions).

3.1 Twitter, Inc.

Platform providers like Twitter, Inc. have a vested interest in the information that flows through their service, and as outlined above, these interests have become more pronounced over time, as the need for a plausible business model has grown more urgent. The users' investment of time and energy is the foundation of the platform's value and therefore growing and improving the service is of vital importance. In the case of Twitter, this strategy is exemplified by the changes made to the main page over the years. Whereas initially Twitter asked playfully, "What are you doing?," this invitation has long since been replaced by a more utilitarian and consumer-oriented exhortation to "Find out what's happening, right now, with the people and organizations you care about," stressing Twitter's relevance as a real-time information hub for business and the mainstream media.

Twitter's business strategy clearly hinges strongly on establishing itself as an irreplaceable real-time information source and on playing a vital part in the corporate media ecosystem of news propagation. Under its current CEO, Dick Costolo, Twitter has moved firmly towards an ad-supported model of "promoted tweets" similar to Google's AdWord model. Exercising tighter control over how users experience and interact with the service than in the service's fledgling days is a vital component of this strategy.

Data is a central interest of Twitter in its role as a platform provider, not solely because it aims to monetise information directly, but because the value of the data determines the value of the company to potential advertisers.

Increasing the relevance of Twitter as a news source is crucial, while maintaining a degree of control over the data market that is evolving under the auspices of the company.

3.2 End-users

Twitter's end users are private citizens, celebrities, journalists, businesses, and organisations; in other words, they can be both individuals and collectives, with aims that are strategic, casual, or a dynamic combination of both. What unites these different stakeholders is that they have an interest in being able to use Twitter free of charge and that data is merely a by-product of their activity, but not their reason for using the platform. They do, however, have an interest in controlling their privacy and being able to do the same things with their information that both Twitter and third-party services are able to do. While the Terms spell out certain rights that users have and constraints that they are under, the rights can only be exercised through the API, while the constraints are enforced by legal means (Beurskens, to appear).

End users have diverse reasons for wanting to control their data, including privacy concerns, impression management, fear of repressive governments, the desire to switch from one social media service to another, and curiosity about one's own usage patterns and behaviour. Giving users the ability to exercise these rights not only benefits users, but also platform providers, because it fosters trust in the service. The perception that platform providers are acting against users' interests behind their back can be successfully countered by implementing tools that allow end users greater control of "their" information.

3.3 Data traders and analysts

Both companies re-selling data under license from Twitter and their clients have interests which are markedly different from those of the company and platform end users. While Twitter seeks long-term profits guaranteed by controlled access to the platform and growing relevance, and end users may want to guard their privacy and control their information while being able to use a free service, data traders want access to vast quantities of data that allow them to model and predict user behaviour on an unprecedented scale. Access to unfiltered, real-time information (provided to them in the form of the

Streaming API) is vital, while to their clients the predictive power of the analytics is important. Neither is very concerned with the interests of end users, who are treated similarly to subjects in an experiment of gigantic proportions. Privacy concerns are backgrounded as they would reduce the quality of the analytics, and they are effectively traded for free access to the platform. What is also neglected is the ability to access historical Twitter data, as businesses by and large want to monitor their current performance, with only limited need to peer into the past.

A key aim of data traders is to commodify data and to guard it carefully against infringers operating outside the data market. In an interview, data wholesaler Gnip's CEO Jud Valeskii returns the responsibility back on end users, recommending they educate themselves about the public and commodified status of the data generated by their personal media use:

> *Read the terms of service for social media services you're using before you complain about privacy policies or how and where your data is being used. Unless you are on a private network, your data is treated as public for all to use, see, sell, or buy. Don't kid yourself. (Valeski, quoted in Steele, 2011, para 27)*

Two things stand out in this statement: the claim that data on Twitter is public and the inference that because it is public, it should be treated as "for all to use, see, sell, or buy." The public-private dichotomy applies to Twitter data only in the sense that what is posted there is accessible to anyone accessing the Twitter website or using a third-party client (with the exception of direct messages and protected accounts). But the question of access is legally unrelated to the issue of ownership—rights to data cannot be inferred from technical availability alone, otherwise online content piracy would be legal. In the same interview, Valeski also consistently refers to platform providers such as Twitter as "publishers" and warns of "black data markets."

4. Terms of Service and API as Instruments of Regulation

Since its launch in March 2006, Twitter has steadily added documents that regulate how users can interact with its service. In addition to the Terms

(Twitter, 2012a), two items stand out: the Developer Rules of the Road (Twitter, 2012b) and the Developer Display Requirements (Twitter, 2012c), which were added to the canon in September 2012. Twitter's Terms have changed considerably since Version 1, published when the platform was still in its infancy. In relation to data access, they lay out how users can access information, what rights Twitter reserves to the data that users generate, and what restrictions apply. Initially the Terms spell out the users' rights with respect to their data, i.e., each user's own personal content on the platform:

> *By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). (Twitter 2012a, para 5-1)*

This permission to use the data is supplemented with the permission to pass it on to sanctioned partners of Twitter:

> *You agree that this license includes the right for Twitter to make such Content available to other companies, organizations or individuals who partner with Twitter for the syndication, broadcast, distribution or publication of such Content on other media and services, subject to our terms and conditions for such Content use. (ibid, para 5-2)*

Third parties are also addressed in the Terms and encouraged to access and use data from Twitter:

> *We encourage and permit broad re-use of Content. The Twitter API exists to enable this. (ibid, para 8-2)*

However, the exact meaning of *re-use* in this context remains unclear, and reading the other above-mentioned documents, the impression is that data analysis is not the kind of re-use intended by the Terms. Neither is it made explicit whether the content referred to is still the users' own content or all data on the platform (i.e., the data of other users). Furthermore, it seems that it is no longer Twitter's users who are addressed, but third parties, as no referent is given. Reference to the API also suggests that a technologically savvy audience is addressed, rather than any typical user of Twitter.

The claim of encouraging broad re-use is further modified by the Developer Rules of the Road, the second document governing how Twitter handles data:

> *You will not attempt or encourage others to: sell, rent, lease, sublicense, redistribute, or syndicate access to the Twitter API or Twitter Content to any third party without prior written approval from Twitter. If you provide an API that returns Twitter data, you may only return IDs (including tweet IDs and user IDs). You may export or extract non-programmatic, GUI-driven Twitter Content as a PDF or spreadsheet by using 'save as' or similar functionality. Exporting Twitter Content to a datastore as a service or other cloud based service, however, is not permitted. (Twitter 2012b, para 8)*

Here, too, developers, rather then end-users are the implicit audience. Not only is the expression "non-programmatic, GUI-driven Twitter Content" fairly vague, the restrictions with regards to means of exporting and saving the data make the "broad re-use" that Twitter encourages in the Terms difficult to achieve in practice. They also stand in contradiction to the Terms which state:

> *Except as permitted through the Services (or these Terms), you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Content or Services. (Twitter 2012a, para 8-2)*

Thus, only by using the API and obtaining written consent from Twitter is it possible to redistribute information to others. This raises two barriers—requiring permission and having the technical capabilities needed to interact with the data—that must both be overcome, narrowing the range of actors able to do so to a small elite. In relation to this form of exclusion, boyd and Crawford (2012) speak of data "haves" and "have-nots," noting that only large institutions with the necessary computational resources will be able to compete. Studies such as those by Kwak, Lee, Park, and Moon (2010) and Romero, Meeder, and Kleinberg (2011) are only possible through large-scale institutional or corporate involvement, as both technical and contractual challenges must be met. While vast quantities of data are theoretically available via Twitter, the process of obtaining it is in practice complicated, and requires a sophisticated infrastructure to capture information at scale.

Actions such as the one against PeopleBrowsr, an analytics company that was temporarily cut off from access to the API, support the impression that Twitter is exercising increasingly tight control over the data it delivers through its infrastructure (PeopleBrowsr, 2012). PeopleBrowsr partnered with Twitter for over four years, paying for privileged access to large volumes of data, but as a result of its exclusive partnerships with specific data resellers, Twitter unilaterally terminated the agreement, citing PeopleBrowsr's services as incompatible with its new business model.

5. Data Rights and Data Literacy

Contemporary discussions of end user data rights have focused mainly on technology's disruptive influence on established copyright regimes, and industry's attempts to counter this disruption. Vocal participants in the digital rights movement are primarily concerned with copyright enforcement and Digital Rights Management (DRM), which, so the argument goes, hinder democratic cultural participation by preventing the free use, embellishment, and re-use of cultural resources (Postigo, 2012a, 2012b). The lack of control that most users can exercise over data they have themselves created in platforms such as Twitter seems a in some respects a much more pronounced issue.

Gnip's CEO Jud Valeski frames the "owners" of social media data to be the platform providers, rather than end users, a significant conceptual step forward from Twitter's own characterization, which endows the platform with the licence to reuse information, but frames end users as its owners (in Steele, 2011). Valeski's logic is based on the need to legitimise the data trade—only if data is a commodity, and if it is owned by the platform provider rather than the individual users producing the content, can it be traded. It furthermore privileges the party controlling the platform technology as morally entitled to ownership of the data flowing through it.

Driscoll (2012) notes the ethical uncertainties surrounding the issues of data ownership, access, and control, and points to the promotion of literacy as the only plausible solution:

> *Resolving the conflict between users and institutions like Twitter is difficult because the ethical stakes remain unclear. Is Twitter ethically bound to explain its internal algorithms and data structures in a language that its users can understand? Conversely, are users ethically bound to learn to speak the language of algorithms and data structures already at work within Twitter? Although social network sites seem unlikely to reveal the details of their internal mechanics, recent 'code literacy' projects indicate that some otherwise non-technical users are pursuing the core competencies necessary to critically engage with systems like Twitter at the level of algorithm and database. (p. 4)*

In the current state, the ability of individual users to effectively interact with "their" Twitter data hinges on their ability to use the API, and on their understanding of its technical constraints. Beyond the technical know-how that is required to interact with the API, issues of scale arise: the Streaming API's approach to broadcasting data as it is posted to Twitter requires a very robust infrastructure as an endpoint for capturing information (see Gaffney & Puschmann, to appear). It follows that only corporate actors and regulators—who possess both the intellectual and financial resources to succeed in this race—can afford to participate, and that the emerging data market will be shaped according to their interests. End-users (both private individuals and non-profit institutions) are without a place in it, except in the role of passive producers of data. The situation is likely to stay in flux, as Twitter must at once satisfy the interests of data traders and end-users, especially with regards to privacy regulation. However, as neither the contractual nor the technical regulatory instruments used by Twitter currently work in favour of end users, it is likely that they will continue to be confined to a passive role.

6. References

Arthur, C. (2012). Twitter too busy growing to worry about Google+, says Dorsey. *Guardian.co.uk*. Retrieved from http://www.guardian.co.uk/technology/2012/jan/23/twitter-dorsey

Backaitis, V. (2012). Data is the New Oil. *CMS Wire*.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The Role of Social Networks in Information Diffusion. *Proceedings of the 21st International*

Conference on the World Wide Web (WWW '12) (pp. 1–10). New York, New York, USA: ACM Press. doi:10.1145/2187836.2187907

Beurskens, M. (to appear). Legal questions of Twitter research. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Burgess, J. & Bruns, A. (2012). Twitter archives and the challenges of 'Big Social Data' for media and communication research. *M/C Journal*, *15*(5). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561

boyd, d. & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society* 15(5), 662-679.

Driscoll, K. (2012). From punched cards to "Big Data": A social history of database populism. *communication +1*, 1, Article 4. Retrieved from http://scholarworks.umass.edu/cpo/vol1/iss1/4

Gaffney, D., & Puschmann, C. (2012). Game or measurement ? Algorithmic transparency and the Klout score. *#influence12: Symposium & Workshop on Measuring Influence on Social Media* (pp. 1–2). Halifax, Nova Scotia, Canada.

Gaffney, D., Puschmann, C. (to appear). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347-364.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter , a social network or a news media? Categories and Subject Descriptors. *Proceedings of the 19th International Conference on the World Wide Web (WWW '10)* (pp. 591–600). Raleigh, NC.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Lessig, L. (1999). *Code and other laws of cyberspace*. New York, NY: Basic Books.

Manovich, L. (2012). Trending: The promises and the challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460-475). Minneapolis: University of Minnesota Press.

O'Reilly, T. (2005). What is Web 2.0? Design patterns and business models for the next generation of software. *O'Reilly Network*. Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

PeopleBrowsr (2012). PeopleBrowsr wins temporary restraining order compelling Twitter to provide firehose access. Retrieved from http://blog.peoplebrowsr.com/2012/11/peoplebrowsr-wins-temporary-restraining-order-compelling-twitter-to-provide-firehose-access/

Postigo, H. (2012a). Cultural production and the digital rights movement. *Information, Communication and Society*, *15*(8), 1165-1185.

Postigo, H. (2012b) *The digital rights movement*. Cambridge, MA: MIT Press.

Rogers, R. A. (2009). *The end of the virtual: Digital methods*. Amsterdam, the Netherlands: Amsterdam University Press.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. *Proceedings of the 19th World Wide Web Conference* (pp. 695–704). New York, NY: ACM.

Rotella, P. (2012). Is Data The New Oil? *Forbes*. Retrieved October 5, 2012, from http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users. *Proceedings of the 19th International Conference on the World Wide Web (WWW '10)* (pp. 1–10). New York, NY: ACM Press. doi:10.1145/1772690.1772777

Steele, J. (2011). Data markets aren't coming, they're already here. *O'Reilly Radar*. Retrieved from http://radar.oreilly.com/2011/01/data-markets-resellers-gnip.html

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science*, *62*(2), 406–418. doi: 10.1002/asi.21462

Thelwall, M. (to appear). Sentiment Analysis and Time Series with Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Twitter (2012a). Terms of Service. Retrieved from http://twitter.com/tos.

Twitter (2012b) Rules of the Road. Retrieved from https://dev.twitter.com/terms/api-terms.

Twitter (2012c) Developer Display Requirements. Retrieved from https://dev.twitter.com/terms/display-requirements.

van Dijck, J. (2011). Tracing Twitter: The rise of a microblogging platform. *International Journal of Media and Cultural Politics*, *7*(3), 333-348.

Zittrain, J. (2008). *The future of the Internet and how to stop it*. New Haven, CT: Yale University Press.

First Monday, Volume 17, Number 11 - 5 November 2012

first monday
PEER-REVIEWED JOURNAL ON THE INTERNET

The refraction chamber: Twitter as sphere and network

by Bernhard Rieder

Abstract

In this paper, we outline a study of the Twitter microblogging platform through a sample of French users. We discuss sampling methodology and compare three "issues" taken from the collected set of tweets. Based on the empirical findings we make a case for extending the notion of "information diffusion" to take into account questions of meaning, values, and ideology. We propose the concept of "refraction" to take a step toward this end.

Contents

Introduction

Since its creation in 2006, the Twitter microblogging service has emerged as a leading platform for short message communication and social networking. According to a recent study (comScore, 2011), Twitter reached one in ten Internet user at the end of 2011, after a year of strong growth (+59 percent). Perhaps even more significantly, Twitter has captured the public imagination due to its (strongly debated) role in political events such as the Iranian elections of 2009 (Morozov, 2009; Shirky, 2009) and what has come to be known as the "Arab spring" (Poell and Darmoni, 2012). These and other entanglements with "serious" matters have gone far in transforming Twitter's image from a system essentially used to share "pointless babble" (Pear Analytics, 2009) to a platform that allows for communication and coordination in significant social movements. For this reason, but also due to its relative openness in terms of data collection, Twitter has quickly become a favored research objects for scholars from various fields.

While most studies currently focus on English language activity, the empirical research this paper is based on is related to a larger project — Internet: Pluralité et Redondance de l'Information (IPRI) — that studies the question of how the Internet contributes to information plurality in the French language Internet (Marty, *et al.*, 2011). While digital networks have certainly been agents of globalization and transnationalization on different levels, the political news and debate sphere — our main object of concern — is, despite important developments over the last 20 years, still strongly organized around national actors, issues, and channels, even inside of the European Union (Wessler, *et al.*, 2008). Certain findings reported in this paper may well be generalizable beyond our empirical focus, but we estimate that certain national particularities do indeed come into play. As this article uses the empirical terrain first and foremost as a resource for conceptual work, we will not stress these particularities more than necessary. While Twitter's success in France is considerable, the microblogging service is not among the top five social networks in the country (comScore, 2011) and a study by AT

Internet (2011) observed that in March 2011 only 0.2 percent of the visitors for the 12 top news sites in France were relayed by the microblogging platform. These figures indicate, in line with similar data for the U.S. (Pew Research Center, 2011), that despite the high number of created accounts, the actual use of Twitter may well be lower than the high levels of media attention would indicate. At the same time, very large volumes of tweets are created and especially media professionals seem to be have embraced the platform with open arms.

This paper aims at making three contributions to research on Twitter that are in large part related to theory and method: first, we presents a research methodology that is based on user sampling rather than subject sampling, which allows us to study the platform as a *sphere* (suggesting both demarcation from the *outside*, and coherence on the *inside*) as well as a series of *conversations*; second, we analyze and compare three case studies in order to highlight specific characteristics of the particular national sphere we are focusing on; and third, based on our empirical findings, we argue that the "dominant paradigm" in Twitter research — *information diffusion* — needs to be extended to better account for dimensions of (shared) meaning, values, and norms. This, in short, is what the concept of "refraction" seeks to address.

Methodology

This paper follows the spirit of Joëlle Le Marec's contemplations on the nature of empirical research, which aim at resolving the supposed contradiction between theory and field work from a *science studies* angle and somewhat related to a *grounded theory* perspective. Here, "[the field] is not the reservoir of facts and social reality as it is spontaneously perceptible in its complexity and richness [...] but a set of operations, unprecedented situations, singular confrontation that occupy the researcher on a daily basis." (Le Marec, 2002) For Le Marec, the question is not whether "the theory" or "the field" takes precedent over the other but rather "what the field *does* to the concept" [our italics] by putting the researcher into a specific and complex epistemological situation that is characterized by a constant production of "surplus". This means that for example in the case of a large–scale communication platform such as Twitter, there is always "more", always an element of surprise — something that pushes beyond the concept and thereby pushes the concept. This is not a "reality check" where the facts correct an erroneous theory but rather a composite of methodology, data, and conceptual work caught up in the situated dynamics of research practice. In the context of this study, analytical methodology and conceptual apprehension had to be revised at several moments and we would like to make this "adjustment work" at least somewhat visible.

Studying Twitter through digital methods based approaches (Rogers, 2009) presents us with both opportunities and challenges. On the one side, different APIs (Application Programming Interfaces) provide relatively comprehensive access to user data and activity. While users can make accounts private, this is relatively rare — according to our tests about one in 10 accounts is protected — and most researchers are interested in the public face of Twitter communication in any case. On the other side, Twitter users now produce many millions of messages every day and such masses of data challenge the capabilities of even the most well funded projects. Every research project is therefore forced to decide, from the outset, on a method for creating a subset of data that will actually be analyzed. In the context of social media, the question of sampling is still far from being completely understood and we will therefore address it in some detail.

Sampling and data collection

In the empirical study of social and cultural phenomena, the question of how to create a corpus of "objects" to analyze is continuously present. In over 150 years of experimentation and debate in and around statistics, a set of standard sampling methods have stabilized and social scientists today have a good sense of the possibilities and limits of each approach. When studying large populations of people, quantitative approaches dominate and samples are most often stratified based on demographic information — usually census data — where individuals are selected in relation to the grid of categories (and their distributions inside of the population) these general surveys establish. When it comes to systems like Twitter, no such grid, *e.g.*, of socio–economic parameters, is available and this introduces the difficult question of how to negotiate between the practical logistical limits most research projects are subject to and the hope to be able to infer from the sample to a larger population. We argue that there are at least five factors that will weigh on decisions concerning sampling methods:

- The "epistemological outlook" of a project is decisive in the sense that different research questions but also different theoretical and methodological paradigms (quantitative, qualitative, representative, etc.) will lead to very different requirements and decisions.
- Many projects may simply not be interested in studying a platform on the whole but focus on geographical, linguistic, temporal or topical subgroups and questions.
- The technological capabilities of a project, which concerns both funding and members' experience of working in necessarily interdisciplinary teams, will limit possibilities in a very practical sense.
- Technical and legal limitations for data access (API restrictions, etc.) may have the effect that sample data cannot be compiled in the desired way.
- Ethical considerations and the increasing requirement to "green light" empirical research by "ethics committees" may infer with researchers' plans. These elements can vary strongly between individual institutions and national cultures.

With these contextual forces in mind, we can start looking at the practical options for sample construction. We can distinguish at least six methods:

- A *full sample* is a possibility, at least in theory: after having been white–listed by Twitter in 2009, Cha, *et al.* (2010) accessed 55M active user accounts with the help of 58 servers and recuperated 1.7B tweets. But the quickly growing data volumes on Twitter introduce massive logistics and even for the cited example, one could argue that the need to define an observation period makes this a partial sample as well. Working with very large amounts of data also introduces the problem that heterogeneous and skewed distributions may make procedures relying overly on averages simply meaningless and therefore require sophisticated analytical tools.
- A *random sample* can retain claims to representativeness and attenuate logistical requirements. Because it is, to our knowledge, simply not feasible to connect Twitter accounts to a census category grid, random selection is the only way to establish a representative sample. Twitter already provides different "statistically correct" data streams via its Streaming API and the incremental numbering scheme for account ids allows for direct sampling as well. The effects of non-normal distributions can produce problems here as well and because of the holes in the data, analysis on the micro level (*e.g.*, following a particular conversation) is no longer feasible.
- A *topic sample* is usually constructed by querying Twitter's Search API for certain keywords or hashtags. This is the most common method used by humanists and social scientists. It is generally much less demanding on the levels of logistics, but it is obviously very difficult to make any strong claims about the platform's uses beyond the studied subject.
- *Marker–based samples* can be compiled with the help of geographical, linguistic or technical pointers provided by the platform. While Twitter does not produce segmentations based on nationality, tweets can be searched on the basis of language and/or geographical location. But language detection is less than satisfactory and only a very low number of tweets (less than two percent according to our testing) are geotagged. Sampling based on technical markers such as the software used to post a message is more promising and may, for example, allow one to study mobile users only.
- *Graph–based sampling* usually proceeds by examining the friend/follower relationships and makes selections based on that data. Different methods from graph theory can be used to select certain dense zones in the network or only the most connected users. According to the selection method used, different biases weigh on the possibilities for interpretation.
- *Manual sampling* is interesting for smaller projects and localized populations. One could, for example, collect the accounts for a country's MPs or select particularly prominent individuals from a certain sector of society. This method is quite common and goes well with a more qualitative research outlook.

Every method implies a particular "epistemological spin" and will influence the kind of understanding that can actually be derived from the analysis of the gathered data. The fit between a research project and a sampling method will largely be negotiated around the five factors mentioned above.

In the context of our research project, we wanted to achieve four goals: first, to create a sample based on user accounts rather than subjects in order to be able to examine and compare a large variety of topics; second, to compile a "national" sample containing (mostly) French and French–language users; third, to focus on users interested in political subjects and current events; fourth, to have a sample that would be sufficiently large to claim at least a partial overview of the uses of Twitter in France. The goal to capture the most visible, the most *public* debates was privileged over representativeness, however. This decision is line with the aspiration to emphasize the *mass media* aspects of Twitter, its particular brand of "publicness", rather than its uses for interpersonal communication.

Our methodology started out with a manually selected "core" of 496 accounts selected by a group of researchers. This core consisted mainly of politicians from all major parties, as well as activists, bloggers, and media professionals that had achieved a certain visibility on the platform [1]. In a second step, we "snowballed" from the initial list by acquiring all users' friends and followers through the REST API [2], which lead us to a pool of 326,532 accounts. To keep numbers manageable, we kept only those accounts that were connected (in either direction) to at least 10 users in our core set. Of the 24,351 resulting accounts, 22,322 were unprotected and 17,361 actually posted at least one message during the observation period (15 February 2011 — 15 April 2011). All of the analyses were performed on this latter set: using the REST API — probably the most reliable data access to Twitter [3] — we stored all tweets posted by these active users over the observation period, 5,883,657 in total.

The collection of tweets also allowed us to confirm the validity of our sample retrospectively by observing a strong coherence between our user sample and the usernames mentioned in the messages. While the final number of users is significantly smaller than the number of accounts created in France, a study by OpinionWay (*Journal du Net*, 2010) estimated that there were only 225K users in France at the end of 2010 and Spintank (2011) indicated an even lower 30K–80K regular users. We are therefore confident that our sample allows for at least some generalizations about the French Twitter territory at the time of observation. While our method captured a certain number of "celebrity" and spam accounts, an analysis of user profiles and tweet language confirms a very strong French dominance, centered around Paris: 5,828 users, roughly one third of our active population, explicitly named "Paris" in their location field. Our sample is also very much concentrated on users working in media or politics related professions and users interested in these topics: 1,549 account descriptions (8.9 percent) had the word "journalist" in it, which is quite a significant percentage and confirms the often–made observation that media professionals have adopted Twitter with particular verve (Hermida, 2010).

Analytical methods and case studies

Our analytical toolkit included a wide variety of statistical, graph–theoretical, and content oriented methods (for a full empirical investigation see Rieder and Smyrnaios, 2012), but in this paper we focus on the third set and follow an approach that combines quantitative elements with a close reading of actual tweets. While we have investigated a larger number of subjects, in order to be able to discuss certain details we will focus on three issues that were tweeted about in significant volumes during our observation period.

The first case concerns the underwater earthquake that occurred on 11 March 2011 off the Japanese coast, which caused a tsunami that left nearly 16K dead and then to a nuclear accident at a power plant in the Fukushima prefecture. The second case can similarly be classified as an "event", but of a much smaller, mostly national scale: on 24 February 2011, Dior's chief designer John Galliano is arrested by police after an anti–Semitic rant in a Paris bar and subsequently first suspended and then fired by his employer. Both of these events were followed from the "breaking" to the point where tweet volumes dropped off significantly, 11 days in both cases. The third subject was analyzed over the full observation period of two month and centers around France's famous "three–strikes" anti–Internet piracy law known by the name of the institution charged with enforcing it: HADOPI (*Haute Autorité pour la diffusion des œuvres et la protection des droits sur internet*). There is no major "event" connected to this subject over the two–month period and it therefore provides a certain contrast to the other two.

We distinguished the subjects by means of a search query, which was relatively straightforward in all three cases, with "galliano" and "hadopi" as rather unambiguous issue identifiers and "japon" as the hashtag quickly established by French users to reference the events in Japan.

While this selection is by no means representative of the large variety of subjects that appeared in our sample — we counted a staggering 207K unique hashtags — they go far in showing variability and allow us to illustrate the more conceptual argument we will make further down.

Analysis

While this selection is by no means representative of the large variety of subjects that appeared in our sample — we counted a staggering 230K unique hashtags — they go far in showing variability and allow us to illustrate the more conceptual argument we will make further down.

Overview

When looking at our three case studies, we immediately see that they are of quite different scale and do not show the same intensity concerning the number and intensity responses they provoke. Table 1 gives an overview of a number of basic indicators.

Table 1: Quantitative overview of the three case studies.

Objects	2011 Japanese Earthquake	John Galliano	HADOPI
Period analyzed	10–20 March 2011 (11 days)	24 February–6 March 2011 (11 days)	15 February–15 April 2011 (60 days)
Query	"japon"	"galliano"	"hadopi"
Number of tweets	44,803	4,965	5,850
Number of users	6,657 (38.3%)	1,907 (11%)	1,548 (9%)
Average tweets per user	6.7	2.6	3.8
Tweets with URLs	56.6%	53.4%	68.2%
Number of hosts linked	2,399	398	349
Percentage of links to top five hosts	13%	21%	52%
Percentage of links to top five hosts	65%	46%	65%

As we may expect, the Japanese tsunami disaster provoked a much greater volume of tweets, from a significantly larger percentage (38.3 percent) of accounts [4], and a much higher number of tweets per user (6.7). Although the Galliano case is much ”closer to home‟ the magnitude of the events in Japan do not leave the French Twitter users indifferent. We can also see that the number and variety of domain names from linked URLs [5] is higher than in the other two cases and a closer examination shows many more non–French sources appearing. This is a global event after all. But even the less spectacular topics are far from insignificant, provoking messages from about 10 percent of our users in both cases. The much higher percentage of URLs in the HADOPI case can be interpreted as a first indicator for a more toned–down quality that can be ascribed to the absence of major variations in temporal intensity compared to the ”burstiness‟ of the two breaking events: much of the content posted around the anti–piracy institution constitutes information and documentation rather than expressions of outrage, sadness, or shock.

During our research, we found that while quantitative indicators did indeed provide an interesting first impression, a closer examination of message content was necessary to further understand differences and commonalities between the cases.

Message content

When analyzing the most popular retweets — a good starting point for characterizing the understandings surrounding a topic — one cannot help but notice the differences in tone and in particular the varying presence of humor, irony, and sarcasm. It may not be surprising that the 13 of the 20 most retweeted messages in the Galliano case are jokes or at least strongly ironic — the scandal does after all involve a video portraying the heavily intoxicated fashion designer declaring his love for Adolf Hitler. But even for the highly destructive tsunami, five out of 20 popular messages are humorous, such as the second most retweeted message:

> @Nain_Portekoi: Le pape attristé par le tremblement de terre au #Japon...Depuis quand il est autorisé à critiquer le boulot de son boss?
> [@Nain_Portekoi: The pope saddened by the earthquake in #Japan... Since when is he authorized to criticize his boss’s work?]
> 11 March 2011 — 22:34, https://twitter.com/Nain_Portekoi/status/46323031282421760

Surprisingly, the same analysis for the HADOPI subject yields only a single humorous message, a false takedown announcement for a popular blog on 1 April. The freedom of the Internet seems to be an issue that cannot be taken lightly, and this observation is further corroborated by the high number of URLs in the 20 most popular tweets (14), which is much lower for Japan (9) and Galliano (5). In the absence of scandal/catastrophe related excitement and the general gravity attributed to the subject, the HADOPI stream works like a highly attentive information network that is, at the one hand side, strongly dominated by a small number of specialized sources (the top two sources account for 46.8 percent of all links sent, a pattern of concentration that we have not observed anywhere else) and very active contributors but, on the other hand side, still receives attention from a relatively large number of users that write about or retweet it (1,548 users). When looking deeper into the contents of the stream, one finds that it provides meticulous information on the day–to–day developments of the subject matter. Users closely follow the subject on the levels of jurisdiction and lawmaking and the high percentage of URLs in tweets is a direct effect of the systematic referencing not only of news items and critical commentary — the author has not found a single positive appreciation of HADOPI — but also of legal materials and technical documents. New propositions or amendments proposed by MPs are duly reported and regularly retweeted.

Popular contents in the Japan earthquake stream are much more disparate: despite the fact that our sample is focused on French nationals living in France (and mostly Paris), an important type of message concerns the ‟coordination” functions often observed in the context of disasters (Bruns, 2011) and consists of important phone numbers, embassy contacts, calls for assistance (donations) and missing person inquiries, which mostly concern French expatriates or tourists, and their relatives in France, such as in this tweet, the third most retweeted:

> @francediplo: Numéro du Centre de crise pour les familles ayant des proches au #Japon : 01

> 43 17 56 46 #séisme
> [@francediplo: Number of the crisis center for families having close ones in #Japan: 01 43 17 56 46 #earthquake]
> 11 March 2011 — 16:42, https://twitter.com/francediplo/status/46234662711988224

A second type of message belongs to a “general information” category, which does however not consist so much of “general” news reporting — there is no lack of that on other channels after all — but rather of very specific developments and accounts (often linking to pictures or videos) as well as estimations of effect or “spectacular” facts, such as the displacement of the Earth’s axis due to the earthquake. A third category is made up by what we could call “repatriation” tweets, which comment on the event from an explicitly French perspective.

Here, we find the jokes and ironic remarks mentioned above, but also critiques of the Sarkozy government that use the catastrophe as a “hook”, France-related micro-scandals (*e.g.*, the price Air France charges for flights out of Japan), comments on comments by French public figures and messages containing “what if this would happen in France?” speculations. This tweet by one of the most popular Twitter personalities in our sample is quite emblematic of the rather irreverent tone:

> @Maitre_Eolas: Les japonais déclarent : “le séisme ok. Le tsunami passe encore. La fuite radioactive on assume. Mais là NON.” htt ...
> [@Maitre_Eolas: The Japanese declare: “the earthquake OK. The tsunami, still manageable. We can deal with the radioactive leak. But this, NO.” htt ...]
> 17 March 2011 — 12:58, https://twitter.com/Maitre_Eolas/status/48352496166518785

The URL then links to an article about a possible visit of Nicholas Sarkozy to Japan. In general, it is difficult to overstate the place critique and ridicule of the (now former) right-wing government in general and the President in particular occupies. We will have to come back to this phenomenon later in this paper.

In the Galliano case, there is, as noted before, a strong dominance of humor and irony (*e.g.*, links to the obligatory subtitled versions of scenes from Oliver Hirschbiegel film “Der Untergang”, which documents Hitler’s last days) but these messages are often very politicized in that they connect Galliano to the government and/or to another highly debated case of “public racism”, the multi-stage scandal around the journalist Eric Zemmour. This tweet, sent from a fake account for Nicholas Sarkozy’s wife, Carla Bruni-Sarkozy, became the fifth most retweeted messages and captures the atmosphere quite well:

> @_Carla_Bruni: Bichon tu pourrais proposer à John Galliano un poste de conseiller politique à l’Elysée non ?
> [@_Carla_Bruni: Darling you could propose a job to John Galliano as a political advisor at the president’s office no?]
> 2 March 2011 — 9:21, https://twitter.com/_Carla_Bruni/status/42862193405988864

Interestingly, the rather quick dismissal of Galliano by his employer, Dior, also became an occasion for what we propose to call “subject drift”, *i.e.*, the connection of one subject to another one in order to make a particular statement. In our case, users started to ask why Galliano’s dismissal could go over so swiftly when most other cases of public racism went unpunished, again, most notably that of Eric Zemmour. The Galliano case became the exception that allowed these users to voice their outrage over what they perceive as the norm: little or no accountability for many public figures when it comes to racist statements. Connecting a smaller event to larger threads of political debate (Galliano => racism, every subject imaginable => Nicholas Sarkozy) is certainly the most common form of subject drift. Finally, there are also a number of purely informational messages linking to news accounts of the incident that become quite popular but these are in a clear minority compared to the comment/ridicule category just described.

This relative lack of purely factual news accounts in our subjects and the broad political and cultural consensus that characterizes popular contents was indeed a surprising revelation that led us to a quite different interpretation of the inner workings of our sample than initially anticipated.

From diffusion to refraction

In the case of this particular empirical research, what "the field did to the concept" (Le Marec, 2002) was not only a drift in methodology toward a more qualitative, content–focused approach, but also a reevaluation of what has become a dominant paradigm in Internet research, in particular in the context of computational methods and "big data", namely the notion of "information diffusion". While studying a system like Twitter with a user sample certainly has its drawbacks in terms of representativeness and completeness, it allows us to examine the platform not only through a thematic slice or quantitative abstractions, but also in a fashion that is more sensitive to subject relations, commonalities, and trajectories of stabilization. We will therefore first discuss the limits of approaches relying solely on the information diffusion paradigm and then propose an extension that, in our view, allows for an interpretation that goes further in making sense of our findings.

Information diffusion

The spread of online platforms that are built on network architectures and that automatically produce analyzable data have been integral to (re)emergence of "diffusionist" approaches to communicative phenomena. Classic models or theories, such as the "two–step flow of communication" (Katz and Lazarsfeld, 1955) or the work on the "diffusion of innovations" (Rogers, 1962), which conceive of social relations as a *medium* through which ideas can spread, have experienced a second spring — the former theory was even found to be valid for Twitter (Wu, *et al.*, 2011). Even Tarde's (2001) idiosyncratic work on *imitation* as a cultural conduit has found a new generation of readers and commentators. Next to smaller fields such as "memetics" (Blackmore, 1999), a "new science of networks" (Watts, 2005) has emerged as the dominant way to study and model distributed communication in online platforms quantitatively. In combination, these developments exert a certain *gravitational pull* on both the conceptual and methodological levels towards a specific understanding of communication as diffusion of information in a network.

While there are considerable differences, diffusionist approaches share a conception of communication that makes a strong separation between an infrastructure and the (informational) "units" that circulate in it. These units can take different forms, from contents to behavior to ideas and opinions, and so can the infrastructure: transport networks, social networks, communication networks, all can be studied using the same conceptual and methodological toolkit. The vocabulary of diffusionist thinking has been widely adopted and terms like "spreading", "cascade", "percolation", "contagion", and so forth, are now commonly encountered in Internet research. The question of power is most often theorized either as *influence*, which goes back to early communication studies (Katz and Lazarsfeld, 1955), or as *access to resources* (often meaning access to information), which is associated with social exchange theory (Blau, 1964; Burt, 1992). Both elements are closely related to the question of network topologies and "powerful" actors — hubs, gatekeepers, influencers — that are thought to be located at structurally significant positions in the network. "Structure" here means something very different from the term's use in the context of structuralist thinking: while the latter conceives of structure as a set of rules and mechanisms that shape both meaning and practice before concrete actors even come into play, the diffusionist approach uses it to denote actor constellations and thereby as external to the actors themselves.

Such a conceptual outlook is highly compatible with computational and especially graph theoretical methods of analysis and papers on "information diffusion" abound with power law distributions of connectivity measures and network visualizations. These methods seem perfectly suited for platforms like Twitter that are equipped with technical features favoring diffusion ("retweeting") but they also resonate well with a contemporary understanding of political journalism organized around scandals, scoops, and information leaks, where the gesture of "unveiling" formerly unknown elements is indeed an act of diffusion. Notions like “real–time” play a significant role in that context and diffusion speed generally receives more attention than long–term dynamics.

While diffusionist approaches are very well suited for describing “bursty” forms of communication in distributed settings, in particular when the production of information is spread out over a territory, *e.g.*, in the context of popular protests, natural catastrophes, and so forth, they are less well equipped to account for more pervasive aspects of politics as long–term processes that are not limited to the question of what information is available, but rather organized around the production of shared understandings, values, and issue hierarchies. Because a network’s structure is the prime source of explanatory capacity, little attention is paid to things like content, interpretation, habitus, and other elements that are related to the question of meaning, such as ideology, cultural hegemony, normalization, trivialization, and so on. Diffusionist approaches are therefore vulnerable to certain elements of Gitlin’s (1978) classic critique of what he perceived as the “dominant paradigm” of the time in the communication field, namely Lazarsfeld functionalist and empiricist outlook.

While a more in–depth discussion about the strengths and weaknesses of diffusionist approaches would be particularly important at this point in the development of Internet research, it is beyond the scope of this paper. On a very naïve level, we simply have to observe that in the context of our empirical research “information spreading” in the sense of “sharing of previously unknown facts” is probably not the most common and certainly not the most significant practice of the Twitter users in our sample. We would therefore like to extend the notion of diffusion with that of “refraction” to be able to better take into account questions referring to meaning, rhetoric, and ideology.

Information refraction

In a somewhat different context from this research, Donna Haraway (1997) introduced the notion of “diffraction” as an “optical metaphor for the effort to make a difference in the world”, which “is about heterogeneous history, not about originals”. In a similar spirit, we would like to propose the metaphor of “refraction” as a way to further think about the space between identical reproduction and total heterogeneity. We do not use Haraway’s term because it suggests a level of heterogeneity and diversity that is simply not observable in our user sample; on the contrary, as we have started to see, political attitudes and moral coordinate systems are largely shared. Rather than the spreading out of waves in all directions that is suggested by diffraction, refraction refers to a singular change in direction for a wave passing through a surface, *e.g.*, transferring from air into water. When looking into a pond, we can still see the fish but our perception of both their size and position is skewed.

Especially when looking at the most popular tweets in our case studies, we find that neutral “reporting”, in the form of merely relaying or linking factual accounts without commentary, is the exception rather than the norm. The most successful tweets are most often those that add a “twist“ to the topic and “spin” it in a certain way, *i.e.*, that “refract” it. Let us consider the following messages from the Galliano case, which were both in the top 10 of the most retweeted messages over the 11–day observation period:

> @Le_Figaro: Alerte : Christian Dior suspend le couturier John Galliano de ses fonctions de directeur artistique http://tinyurl.com/6k ...
> [@Le_Figaro: Alert: Christian Dior suspends the fashion designer John Galliano from his functions as artistic director http://tinyurl.com/6k ...]
> 25 February 2011 — 15:06, https://twitter.com/Le_Figaro/status/41136918691454976

> @isaway: En fait #Galliano est complètement #hasbeen la mode est à haïr les musulmans enfin ! Pas les juifs !
> [@isaway: Actually #Galliano is completely #hasbeen come on, it’s fashionable to hate the Muslims! Not the Jews!]
> 2 March 2011 — 19:33, https://twitter.com/isaway/status/43016135066652672

This first message was posted on 25 Feburary, the day the news about the fashion designer’s insults “broke”, and this represents the kind of factual reporting that fits well into the diffusionist paradigm. One can easily locate the starting point(s) of the “previously unknown” information and then study how users “become informed” and, by spreading the news, “inform” others. The second is written five days later, on 2 March, when the affair is in full swing. While we can certainly think about it terms of diffusion as well (the message indeed spreads by being retweeted), the striking element is the subject drift to the question of Islam, which is as hotly debated in France as elsewhere. This is what we mean by refraction: the “issue” is commented upon, connected to a different issue or a specific detail is underscored.

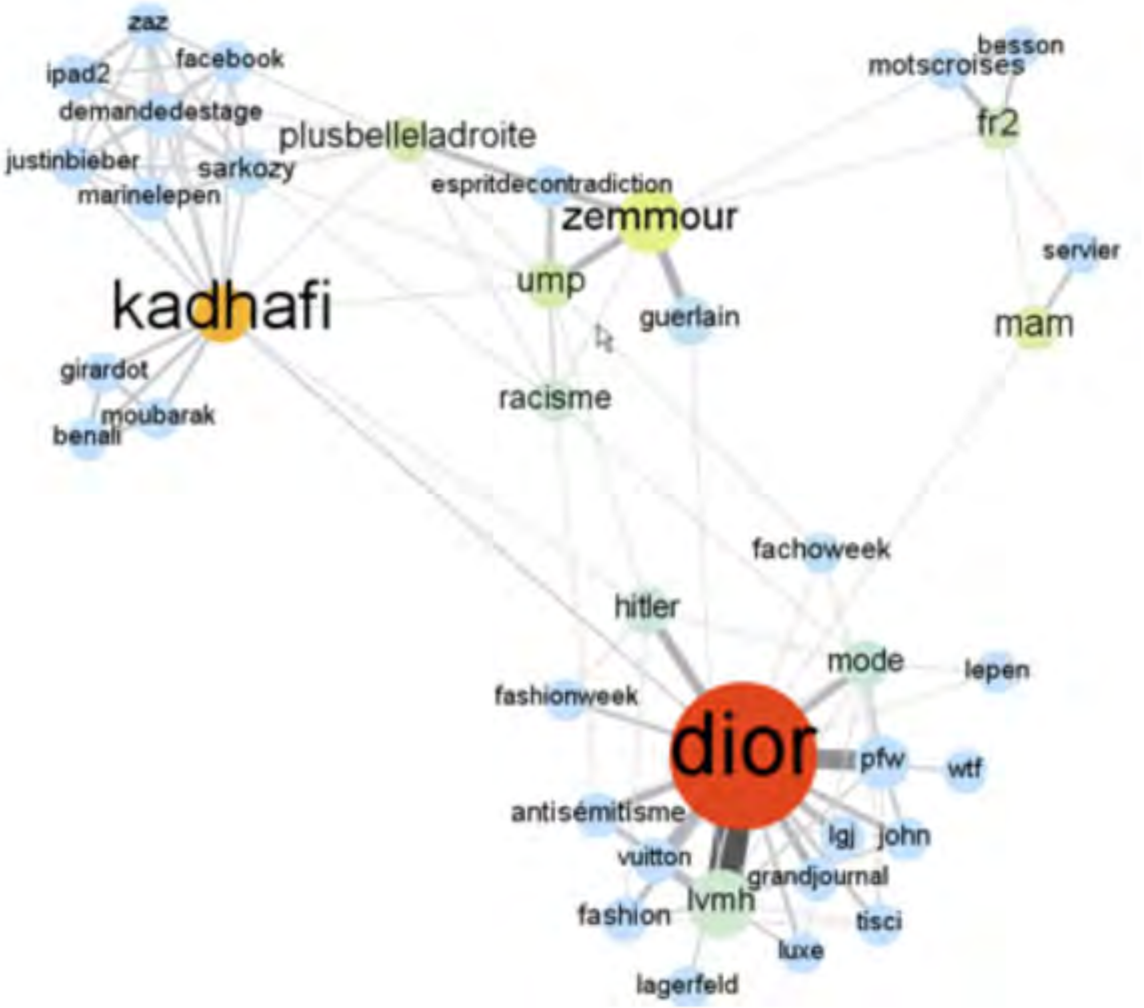


Figure 1: Co–word analysis, visualized with gephi, based on hashtags appearing at least 10 times in the “Galliano” dataset. Two hashtags are connected if they appear in the same tweet. Node size shows frequency, color (blue => yellow => red) shows betweenness centrality.

But this refraction is only possible in the length of a tweet because messages not only circulate in a network infrastructure, but also in a cultural *sphere*, a space of meaning that the tweet can mobilize to make its *argument*: if we follow Geertz (1973) in taking culture as “webs of significance”, we can see how even a very short message can convey complex meaning by drawing on a reservoir of shared ideas, debates, stereotypes, facts, trivia, and so on, which can be often be *evoked* with a single word. The second message plays with these webs on the level of medium specific conventions (“#hasbeen”), on the level of understandings about the fashion world where trends play a significant role, and on the level of the larger debates about racism. The tweet brings all of these elements together, drawing on them for its message and tying them together in the process. Shared meaning is both the condition and the outcome of the *work* the user does here. As Figure 1 shows, the joining element need not be profound in any way: the thematic "connector" between the Galliano incident and Muammar Gaddafi (here in French spelling) is a shared preference in headwear.

The category of messages we called "repatriation" tweets fit well into this line of interpretation if we consider the term to mean an anchoring not necessarily only into a national context but, more generally, into a familiar set of meanings and values. It is somewhat banal to underscore that users comment on issues from the perspective of their own immediate concerns; interestingly though, these concerns seem to be largely shared, even beyond subject matters. As indicated above, the dominant *frame* (Lakoff, 2009) in our sample is organized around criticism of the right–wing government in power at the time of our data collection, and in particular the person of President Sarkozy. This critique is virtually omnipresent and it is in this sense that refraction can be understood not necessarily as a multiplication of perspectives and opinions, but as the interpretation of many different events in terms of a limited set of largely shared concerns and ideas. If we follow boyd, *et al.* (2010) in that "retweeting can be understood both as a form of information diffusion and as a means of participating in a diffuse conversation", this conversation may be diffuse on the level of the dynamics of individual messages, but allows — at least in our sample — for the emergence of focus points — issues, norms, references — that give it structure in terms of meaning rather than topology.

As Marwick and boyd (2011) indicate, Twitter has become a strategic medium — in particular for certain professions — that is used to achieve "micro–celebrity" and the immediate feedback users can get from their precarious "network audience" (Marwick and boyd, 2011) may lead to mainstreaming in terms of the diversity of opinions represented. When looking at our sample, one could also make the observation that it represents what Bourdieu (1996) called the "journalistic field", the ensemble of media professionals that spend their time talking to and observing each other, progressively aligning their perspectives by imitating strategies and attitudes that "work" in terms of retweets, clicks, and other metrics. The "multi–faceted and fragmented news experience" Hermida (2010) speaks of may actually appear a lot less fragmented when we start looking at the communalities that form behind the microbursts that appear on our screen.

From a methodological standpoint, the question remains whether the concept of refraction can only be illustrated on a microscopic scale, by reading individual tweets, or if a more macroscopic approach is feasible. The following section proposes using co–word analysis as a means for the latter.

Mapping refraction

If we take Twitter hashtags to be a good conduit to study message content in a more condensed form, the topical diversity in our sample is staggering at first glance: we identified 207,059 unique elements in a pool of 2,217,937 hashtags posted. As we have argued, this diversity does not exclude concentration, however. The top five terms make up 7.6 percent and the top fifty 21.9 percent of all hashtags posted over the two month period. Focusing on a one–week period (28 February 2011 — 6 March 2011), over which 40,687 unique hashtags were used, an analysis of the 1,000 most used hashtags — accounting for 59.9 percent of all occurrences — allows us to make further observations.

While we are hesitant to statistically quantify the phenomenon we have labeled as refraction, mostly because we consider the concept to be *interpretative* (Geertz, 1973) rather than formal, co–word analysis (Callon, *et al.*, 1983) is uniquely suited to study complex textual material in a more rigorous manner. If we take hashtags to be equivalents of "macro–terms" that "crystallize and synthesize" [6] discourse, the analysis of the relationships between these terms allows us to map attempts at *drawing issues together.*

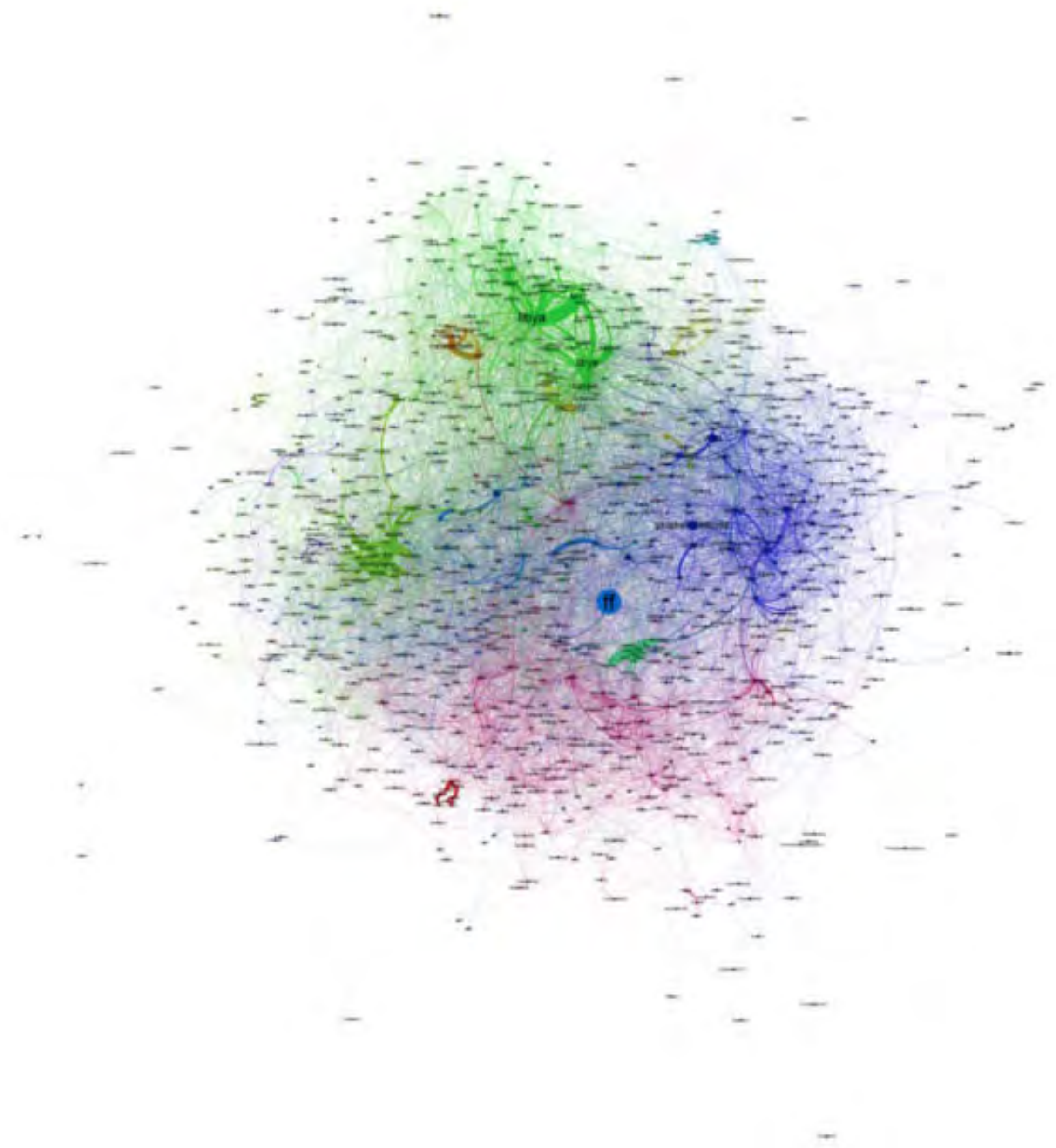


Figure 2: Co–word analysis, visualized with gephi, based on the 961 of the top 1,000 hashtags that form a giant component. Node size shows frequency and link width expresses the frequency of two tweets co–occurring in a same tweet. Colors are provided by gephi's community detection algorithm. For a higher resolution image, see http://bit.ly/YHP1cW.

Of the 1,000 most frequent hashtags, 961 appear in a connected component when we create a network of hashtags that are connected by co–occurring in a tweet. This giant component forms a small world, with a diameter of 6 and an average path length of only 2.66. Hashtags are very well connected, due to a high average degree (number of connections) of 19.95. This basically means that hashtags have a tendency to co–occur with a large variety of others, even if we consider that degree values are distributed quite unevenly — some hashtags are simply much more connected than others. If we take a closer look at the visual representation of the network, the structural organization of the network becomes clearer. There obviously are areas of thematic concentration: political debates on the right side, technology related topics on the left, and political upheavals in Africa and Asia at the top. These clusters are relatively well captured by the community detection algorithm provided by the gephi network visualization toolkit. If we see refraction merely as users making connections between different issues — and we would like to term to have a broader meaning that includes the other elements discusses above — the co–occurrence map shows to distinct levels: first, the high density in larger topic clusters indicates — unsurprisingly — that connections to "closer" issues are more frequent; this is particularly interesting in the upper center of the map, where events in Iran, Egypt, Lybia, Tunisia, and Ivory Coast ("civ2010") are frequently brought together. Second, even if one discounts hub nodes such as "ff" ("follow Friday") that signal platform conventions rather than issues, the full network holds well together, which means that connections between the different topic clusters is far from uncommon. Geertz' notion of culture as webs of significance is, in a sense, made visible here.

Conclusions

To conclude, we would like to underline the caveat Hermida (2010) adds to his account of the "fragmented news experience" provided by Twitter: "The value does not lie in each individual fragment of news and information, but rather in the mental portrait created by a number of messages over a period of time." What emerges from a content oriented examination of an admittedly small number of news subjects as they were discussed by a set of 17K French Twitter users is a "mental portrait" that takes the form of a *sphere* at the same time as that of a *network:* beyond the leveling layer of infrastructure, there is a production of *insides* and *outsides,* of borders (linguistic, cultural, political, etc.) and shared spaces. While a diffusion–oriented approach indeed shows a chaotic staccato of messages, bursts of attention, and the classic power–law distributions when it comes to connectivity and retweet frequency, a content–oriented perspective paints a much less heterogeneous picture. While our selected subjects do not share a common scale or temporality, shared values and concerns clearly shine through the most popular contents and lead us to diagnose the kind of commonality the image of the sphere evokes. Co–word analysis is certainly a means to bridge the gulf between the two concepts and methodological approaches: by modeling the relationships between hashtags as a network, we can begin to map the webs of significance manifesting in media spaces like Twitter and make claims about content in the face of very large amounts of data.

Whether the absence of political polarization, albeit often observed on Twitter when focusing on the U.S. (Conover, *et al.*, 2011), is an artifact of our sampling method or simply the reflection of the dominance of center–left positions in the media–savvy population active on Twitter in France cannot be fully decided — sampling on Twitter remains a deeply problematic exercise. However, a recent study (Harris Interactive, 2012) indicates a strong left leaning by French journalists active on Twitter and confirms our suspicion that the right is simply underrepresented in the media circles that dominate the platform. While we can, in line with An, *et al.* (2011), confirm the presence of a wide variety of sources, the refraction of these sources to a limited number of shared reference points suggests a lot less diversity on the level of opinions and values than initially anticipated — at least on the level of mainstreaming that we have focused on.

So why do we call Twitter a "refraction chamber" rather than simply follow Sunstein (2001) and speak of an "echo chamber"? Because we want to put the emphasis on something that the latter metaphor captures only imperfectly: instead of merely being exposed to like–mindedness, we consider that the users are the driving force behind the production of shared values and understandings. More than just following homophilic "urges" that result in biased source selection (*i.e.*, who to follow), refraction suggests that commonality is the result of labor on different levels and a *product* rather than an *effet pervers,* an unintended consequence. This difference may appear insignificant but it opens the door for interpretations that go further into the direction of ideological analysis.

While this research needs to be extended on virtually all levels, we hope to have shown that methodological and theoretical issues are supremely important when it comes to studying a complex communication system such as Twitter. We would also hope that a more intensive debate on methodology and theory in Twitter research would take place in the future, a debate that centers on the question how the impressive results from diffusionist approaches can be brought together with a perspective that goes further in accounting for matters of meaning and ideology.

About the author

Bernhard Rieder is an Assistant Professor at Amsterdam University's Media Studies department and a researcher and developer at the Digital Methods Initiative.
E–mail: rieder [at] uva [dot] nl

Acknowledgements

The empirical research presented here has profited from major contributions by Nikos Smyrnaios and Raphaël Velt. The author would also like to thank the participants of the Digital Methods Initiative's 2012 winter school for their valuable comments, in particular Anne Helmond, Catalina Iorga, Richard Rogers, and Michael Stevenson.

Notes

[1] The initial sample was constructed by three researchers — all long–term Twitter users — through field research at the end of 2010 and aimed at constructing an anchor point for capturing users interested in political subjects. The first step consisted of attempting to compile an *exhaustive* list of accounts by French politicians from all major parties with the help of the platform's search and navigation functions. This initial list of 380 users was complemented with 106 accounts maintained by activists, bloggers, and media professionals having achieved notoriety on political subjects on the platform. While we cannot guarantee representativeness — a near impossible task — particular attention was paid to include the full political spectrum in the initial set.

[2] For an explanation of Twitter's different APIs, see https://dev.twitter.com/docs/history-rest-search-api.

[3] In contrast to the often–used search API, the REST API provides access on a per user basis and is subject to fewer limitations: the last 3,200 tweets per user can be collected. Due to rate limiting, only a limited number of users can be accessed per day. Using a rotating set of access tokens, we were able to access all user accounts roughly twice per day.

[4] To provide some context, the three other main issues emerged around the events in Libya (140K messages over two months), Tunisia (40K messages over two months) and the cantonal elections in France (39K messages over two months).

[5] For this analysis, all shortened URLs sent in tweets were translated to their long form.

[6] Callon, *et al.*, 1983, p. 199.

References

Jisun An, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft, 2011. "Media landscape in Twitter: A world of new conventions and political diversity," *Proceedings of the Fifth International Conference on Weblogs and Social Media*, and at http://www.cl.cam.ac.uk/~jac22/out/twitter-diverse.pdf, accessed 2 November 2012.

AT Internet, 2011. "Sites médias: Une visite sur 6 issues de sites affluents vient de Facebook," at http://www.atinternet.com/Ressources/Etudes/Enjeux-web-marketing/Reseaux-sociaux-Mars-2011/index-1-1-4-233.aspx, accessed 13 March 2012.

Susan Blackmore, 1999. *The meme machine*. Oxford: Oxford University Press.

Peter M. Blau, 1964. *Exchange and power in social life*. New York: Wiley.

Pierre Bourdieu, 1996. *Sur la télévision: Suivi de L'emprise du journalisme*. Paris: Liber éditions.

danah boyd, Scott Golder, and Gilad Lotan, 2010. "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," *HICSS '10: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10.

Axel Bruns, 2011. "Towards distributed citizen participation: Lessons from WikiLeaks and the Queensland floods," *CeDEM11: Proceedings of the International Conference for E–Democracy* *and Open Government*, pp. 35–52.

Ronald S. Burt, 1992. *Structural holes: The social structure of competition*. Cambridge, Mass.: Harvard University Press.

Michel Callon, Jean–Pierre Courtial, William A. Turner and Serge Bauin, 1983. “From translations to problematic networks: An introduction to co–word analysis,” *Social Science Information*, volume 22, number 2, pp. 191–235.

Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi, 2010. “Measuring user influence in Twitter: The million follower fallacy,” *ICWSM ’10: Proceedings of the International AAAI Conference on Weblogs and Social Media*.

comScore, 2011. “It’s a social world: Top 10 need–to–knows about social networking and where it’s headed” (21 December), at http://www.comscore.com/it_is_a_social_world, accessed 13 March 2012.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer, 2011. “Political polarization on Twitter,” *ICWSM ’11: Proceedings of the Fifth International Conference on Weblogs and Social Media*.

Clifford Geertz, 1973. *The interpretation of cultures: Selected essays*. New York: Basic Books.

Todd Gitlin, 1978: “Media sociology: the dominant paradigm,” *Theory and Society*, volume 6, number 2, pp. 205–253.

Donna J. Haraway, 1997. *Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse*. New York: Routledge.

Harris Interactive, 2012. “Les journalistes présents sur Twitter et la campagne présidentielle de 2012,” at http://www.harrisinteractive.fr/news/2012/CP_HIFR_Medias_14062012.pdf, accessed 22 August 2012.

Alfred Hermida, 2010. “Twittering the news: The emergence of ambient journalism,” *Journalism Practice*, volume 4, number 3, pp. 297–308.

Journal du Net, 2010. “Twitter rassemblerait 225 000 utilisateurs en France” (12 November), at http://www.journaldunet.com/ebusiness/le-net/membres-twitter-en-france-1110.shtml, accessed 13 March 2012.

Elihu Katz and Paul Lazarsfeld, 1955. *Personal influence: The part played by people in the flow of mass communications*. Glencoe, Ill.: Free Press.

George Lakoff, 2009. *The political mind: Why you can’t understand 21st–century politics with an 18th–century brain*. New York: Viking.

Joëll Le Marec, 2002. *Ce que le “terrain” fait aux concepts: Vers une théorie des composites*. Paris: Université Paris Diderot — Paris 7, at http://sciences-medias.ens-lyon.fr/scs/IMG/pdf/HDR_Le_Marec.pdf, accessed 13 March 2012.

Emmanuel Marty, Nikos Smyrnaios, and Franck Rebillard, 2011. “A multifaceted study of online news diversity: Issues and methods,” *Proceedings of the ECREA Journalism Studies Section and 26th International Conference of Communication*, pp. 228–242.

Alice E. Marwick and danah boyd, 2011. “I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience,” *New Media & Society*, volume 13, number 1, pp. 114–133.

Evgeny Morozov, 2009. “How dictators watch us on the Web,” *Prospect*, number 165, at http://www.prospectmagazine.co.uk/magazine/how-dictators-watch-us-on-the-web/, accessed 13 March 2012.

Pear Analytics, 2009. “Twitter study — 2009,” at http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf, accessed 13 March 2012.

Pew Research Center, 2011. “Navigating news online” (9 May), at http://www.journalism.org/analysis_report/navigating_news_online, accessed 13 March 2012.

Thomas Poell and Kaouthar Darmoni, 2012. “Twitter as a multilingual space: The articulation of the Tunisian revolution through #sidibouzid," *NECSUS_European Journal of Media Studies*, volume 1, number 1, at http://www.necsus-ejms.org/twitter-as-a-multilingual-space-the-articulation-of-the-tunisian-revolution-through-sidibouzid-by-thomas-poell-and-kaouthar-darmoni/, accessed 22 August 2012.

Bernhard Rieder and Nikos Smyrnaios, 2012. "Pluralisme et infomédiation sociale de l'actualité: Le cas de Twitter," Réseaux, forthcoming.

Everett M. Rogers, 1962. *Diffusion of innovations*. New York: Free Press of Glencoe.

Richard Rogers, 2009. *The end of the virtual: Digital methods*. Amsterdam: Amsterdam University Press.

Clay Shirky, 2009. "The net advantage," *Prospect*, number 165, at http://www.prospectmagazine.co.uk/magazine/the-net-advantage/, accessed 13 March 2012.

Spintank, 2011. "Twitter en France, au–delà de l'écume" (3 January), at http://www.spintank.fr/twitter-en-france-etat-des-lieux-chiffres-2011/, accessed 13 March 2012.

Cass Sunstein, 2001. *Republic.com*. Princeton, N.J.: Princeton University Press.

Gabriel Tarde, 2001. *Les lois de l'imitation*. Paris: Seuil.

Duncan J. Watts, 2005. "The 'new' science of networks," *Annual Review of Sociology*, volume 30, pp. 243–270.

Harmut Wessler, Bernhard Peters, and Michael Bruggemann, 2008. *Transnationalization of public spheres*. New York: Palgrave Macmillan.

Shaomei Wu, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts, 2011. "Who says what to whom on Twitter," *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, at http://www.www2011india.com/proceeding/proceedings/p705.pdf, accessed 2 November 2012.

Editorial history

Received 9 August 2012; accepted 15 October 2012.

The refraction chamber: Twitter as sphere and network
by Bernhard Rieder
First Monday, Volume 17, Number 11 - 5 November 2012
http://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/4199/3359
doi:10.5210/fm.v17i11.4199

Reading Salon #3: We Take All (Network) Shapes and Sizes

Moderators: Anne Helmond and Esther Weltevrede

Elmer, Greg, and Ganaele Langlois. 2013. "Networked Campaigns: Traffic Tags and Cross Platform Analysis on the Web." Information Polity 18 (1) (January 1): 43–56.

Highfield, Tim. 2012. "Talking of Many Things: Using Topical Networks to Study Discussions in Social Media." Journal of Technology in Human Services 30 (3-4): 204–218.

Manjoo, Farhad. 2012. "The End of the Echo Chamber." Slate, January 17.

Ruppert, E, J Law, and M Savage. "Reassembling Social Science Methods: the Challenge of Digital Devices." Theory, Culture & Society (May 14, 2013).

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The Role of Social Networks in Information Diffusion. Proceedings of the 21st International Conference on the World Wide Web (WWW '12) (pp. 1–10). New York, New York, USA: ACM Press.

The Role of Social Networks in Information Diffusion

Eytan Bakshy*
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
ebakshy@fb.com

Itamar Rosenn
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
itamar@fb.com

Cameron Marlow
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
cameron@fb.com

Lada Adamic
University of Michigan
105 S. State St.
Ann Arbor, MI 48104
ladamic@umich.edu

ABSTRACT

Online social networking technologies enable individuals to simultaneously share information with any number of peers. Quantifying the causal effect of these mediums on the dissemination of information requires not only identification of who influences whom, but also of whether individuals would still propagate information in the absence of social signals about that information. We examine the role of social networks in online information diffusion with a large-scale field experiment that randomizes exposure to signals about friends' information sharing among 253 million subjects in situ. Those who are exposed are significantly more likely to spread information, and do so sooner than those who are not exposed. We further examine the relative role of strong and weak ties in information propagation. We show that, although stronger ties are individually more influential, it is the more abundant weak ties who are responsible for the propagation of novel information. This suggests that weak ties may play a more dominant role in the dissemination of information online than currently believed.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Experimentation, Measurement, Human Factors

Keywords

social influence, tie strength, causality

1. INTRODUCTION

Social influence can play a crucial role in a range of behavioral phenomena, from the dissemination of information, to the adoption of political opinions and technologies [23, 42], which are increasingly mediated through online systems [17, 38]. Despite the wide availability of data from online social networks, identifying influence remains a challenge. Individuals tend to engage in similar activities as their peers, so it is often impossible to determine from observational data whether a correlation between two individuals' behaviors exists because they are similar or because one person's behavior has influenced the other [5, 32, 39]. In the context of information diffusion, two people may disseminate the same information as each other because they possess the same information sources, such as web sites or television, that they consume regularly [3, 38].

Moreover, homophily – the tendency of individuals with similar characteristics to associate with one another [1, 28, 34] – creates difficulties for measuring the relative role of strong and weak ties in information diffusion, since people are more similar to those with whom they interact often [22, 34]. On one hand, pairs of individuals who interact more often have greater opportunity to influence one another and have more aligned interests, increasing the chances of contagion [11, 27]. However, this commonality amplifies the potential for confounds: those who interact more often are more likely to have increasingly similar information sources. As a result, inferences made from observational data may overstate the importance of strong ties in information spread. Conversely, individuals who interact infrequently have more diverse social networks that provide access to novel information [12, 22]. But because contact between such ties is intermittent, and the individuals tend to be dissimilar, any particular piece of information is less likely to flow across weak ties [14, 37]. Historical attempts to collect data on how often pairs of individuals communicate and where they get their information have been prone to biases [10, 33], further obscuring the empirical relationship between tie strength and diffusion.

Confounding factors related to homophily can be addressed using controlled experiments, but experimental work has thus far been confined to the spread of highly specific information within limited populations [6, 13]. In order to understand how information spreads in a real-world environment, we wish to examine a setting where a large population of individuals frequently exchange information with their peers. Facebook is the most widely used social networking service in the world, with over 800 million people using the service each month. For example, in the United States, 54% of adult Internet users are on Facebook [26].

*Part of this research was performed while the author was a student at the University of Michigan.

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

Those American users on average maintain 48% of their real world contacts on the site [26], and many of these individuals regularly exchange news items with their contacts [38]. In addition, interaction among users is well correlated with self-reported intimacy [18]. Thus, Facebook represents a broad online population of individuals whose online personal networks reflect their real-world connections, making it an ideal environment to study information contagion.

We use an experimental approach on Facebook to measure the spread of information sharing behaviors. The experiment randomizes whether individuals are exposed via Facebook to information about their friends' sharing behavior, thereby devising two worlds under which information spreads: one in which certain information can only be acquired external to Facebook, and another in which information can be acquired within or external to Facebook. By comparing the behavior of individuals within these two conditions, we can determine the causal effect of the medium on information sharing.

The remainder of this paper is organized as follows. We further motivate our study with additional related work in Section 2. Our experimental design is described in Section 3. Then, in Section 4 we discuss the causal effect of exposure to content on the newsfeed, and how friends' sharing behavior is correlated in time, irrespective of social influence via the newsfeed. Furthermore, we show that multiple sharing friends are predictive of sharing behavior regardless of exposure on the feed, and that additional friends do indeed have an increasing causal effect on the propensity to share. In Section 5 we discuss how tie strength relates to influence and information diffusion. We show that users are more likely to have the same information sources as their close friends, and that simultaneously, these close friends are more likely to influence subjects. Using the empirical distribution of tie strength in the network, we go on to compute the overall effect of strong and weak ties on the spread of information in the network. Finally, we discuss the implications of our work in Section 6.

2. RELATED WORK

Online networks are focused on sharing information, and as such, have been studied extensively in the context of information diffusion. Diffusion and influence have been modeled in blogs [2, 20, 25], email [31], and sites such as Twitter, Digg, and Flickr [8, 21, 29]. One particularly salient characteristic of diffusion behavior is the correlation between the number of friends engaging in a behavior and the probability of adopting the behavior. This relationship has been observed in many online contexts, from the joining of LiveJournal groups [7], to the bookmarking of photos [15], and the adoption of user-created content [9]. However, as Anagnostopoulos, et al. [4] point out, individuals may be more likely to exhibit the same behavior as their friends because of homophily rather than as a result of peer influence. Statistical techniques such as permutation tests and matched sampling [5] help control for confounds, but ultimately cannot resolve this fundamental problem [39].

Not all diffusion studies must infer whether one individual influenced another. For example, Leskovec et al. [30] study the explicit graph of product recommendations, Sun et al. [41] study cascading in page fanning, and Bakshy et al. [9] examine the exchange of user-created content. However, in all these studies, even if the source of a particular contagion event is a friend, such data does not tell us about the relative importance of social networks in information diffusion. For example, consider the spread of news. In Bradley Greenberg's classsic study of media contagion [24], 50% of respondents learned about the Kennedy assassination via interpersonal ties. Despite the substantial word-of-mouth spread, it is clear that all of the respondents would have gotten the news at a slightly later point in time (perhaps from the very same media outlets as their contacts), had they not communicated with their peers. Therefore, a complete understanding of the importance of social networks in information diffusion not only requires us to identify sources of interpersonal contagion, but also requires a counterfactual

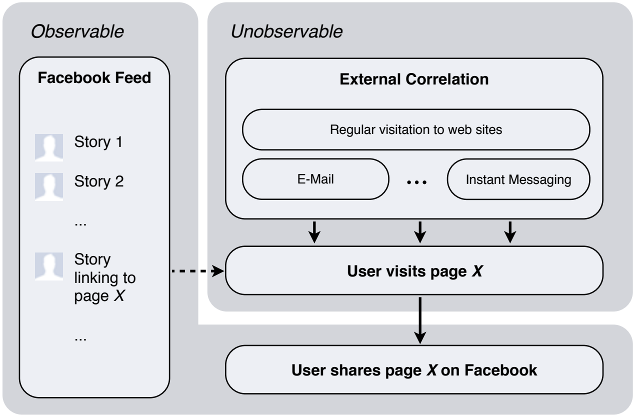
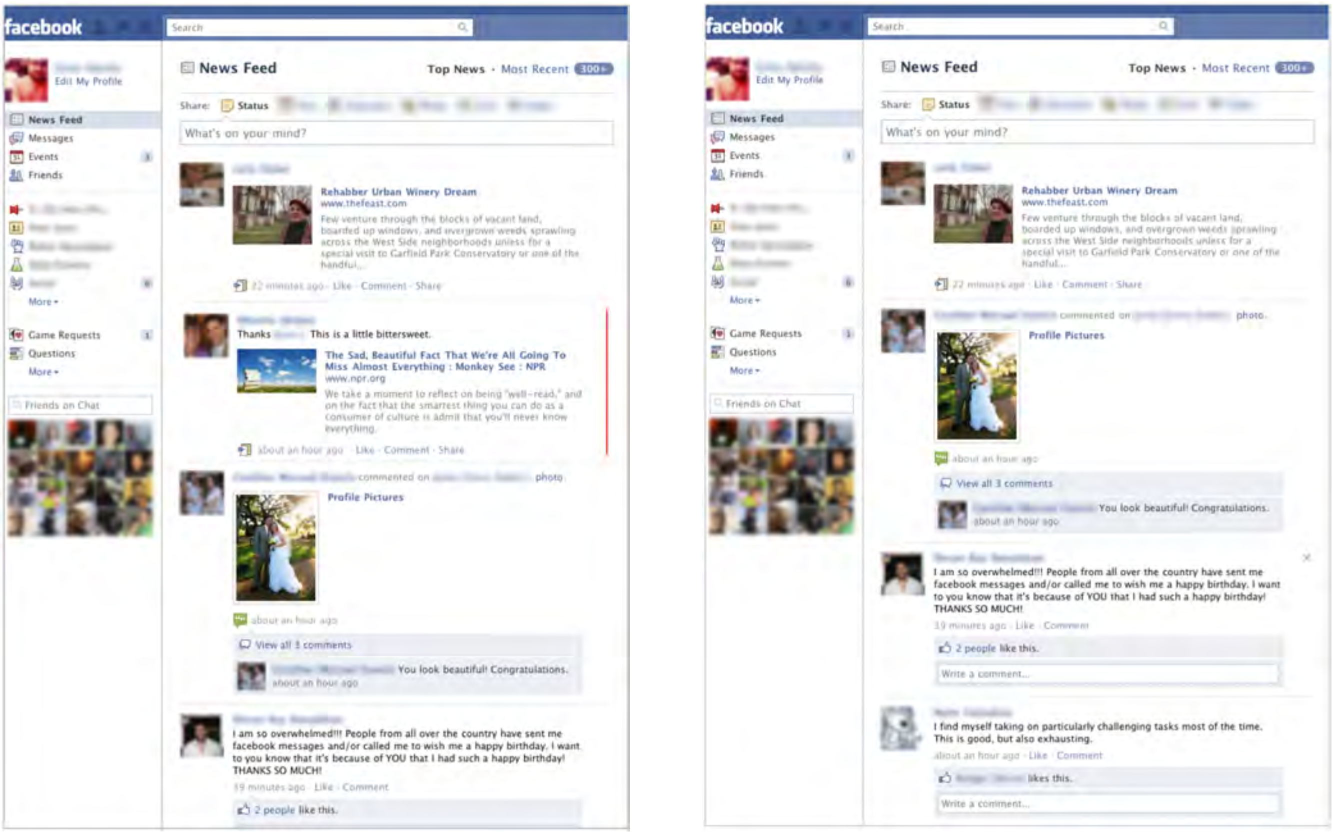


Figure 1: Causal relationships that explain diffusion-like phenomena. Information presented in users' news feeds and other sharing behavior on facebook.com are observed. External events that cause users to be exposed to information outside of Facebook cannot be observed and may explain their sharing behavior. Our experiment blocks the causal relationship (dashed arrow) between the Facebook newsfeed and user visitation by randomly removing stories about friends' sharing behavior in subjects' feeds. Thus, our experiment allows us to compare situations where both influence via the feed and external correlations exist (the *feed* condition), to situations in which only external correlations exist (the *no feed* condition).

3. EXPERIMENTAL DESIGN AND DATA

Facebook users primarily interact with information through an aggregated history of their friends' recent activity (stories), called the News Feed, or simply feed for short. Some of these stories contain links to content on the Web, uniquely identified by URLs. Our experiment evaluates how much exposure to a URL on the feed increases an individual's propensity to share that URL, beyond correlations that one might expect among Facebook friends. For example, friends with whom a user interacts more often may be more likely to visit sites that the user also visits. As a result, those friends may be more likely to share the same URL as the



Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...
22 minutes ago · Like · Comment · Share

Thanks This is a little bittersweet.
The Sad, Beautiful Fact That We're All Going To Miss Almost Everything : Monkey See : NPR
www.npr.org
We take a moment to reflect on being "well-read," and on the fact that the smartest thing you can do as a consumer of culture is admit that you'll never know everything.
about an hour ago · Like · Comment · Share

commented on photo.
Profile Pictures
about an hour ago
View all 3 comments
You look beautiful! Congratulations.
about an hour ago

I am so overwhelmed!!! People from all over the country have sent me facebook messages and/or called me to wish me a happy birthday. I want to you know that it's because of YOU that I had such a happy birthday! THANKS SO MUCH!
19 minutes ago · Like · Comment
2 people like this.

(a)

Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...
22 minutes ago · Like · Comment · Share

commented on photo.
Profile Pictures
about an hour ago
View all 3 comments
You look beautiful! Congratulations.
about an hour ago

I am so overwhelmed!!! People from all over the country have sent me facebook messages and/or called me to wish me a happy birthday. I want to you know that it's because of YOU that I had such a happy birthday! THANKS SO MUCH!
19 minutes ago · Like · Comment
2 people like this.
Write a comment...

I find myself taking on particularly challenging tasks most of the time. This is good, but also exhausting.
about an hour ago · Like · Comment
likes this.
Write a comment...

(b)

Figure 2: An example of the Facebook News Feed interface for a hypothetical subject who has a link (highlighted in red) assigned to the (a) *feed* or (b) *no feed* condition.

user before she has the opportunity to share that content herself. Additional unobserved correlations may arise due to external influence via e-mail, instant messaging, and other social networking sites. These causal relationships are illustrated in Figure 1. From the figure, one can see that all unobservable correlations can be identified by blocking the causal relationship between the Facebook feed and sharing. Our experiment therefore randomizes subjects with respect to whether they receive social signals about friends' sharing behavior of certain Web pages via the Facebook feed.

3.1 Assignment Procedure

Subject-URL pairs are randomly assigned at the time of display to either the *no feed* or the *feed* condition. Stories that contain links to a URL assigned to the *no feed* condition for the subject are never displayed in the subject's feed. Those assigned to the *feed* condition are not removed from the feed, and appear in the subject's feed as normal (Figure 2). Pairs are deterministically assigned to a condition at the time of display, so any subsequent share of the same URL by any of a subject's friends is also assigned to the same condition. To improve the statistical power of our results, twice as many pairs were assigned to the *no feed* condition. Because removal from the feed occurs on a subject-URL basis, and we include only a small fraction of subject-URL pairs in the *no feed* condition, a shared URL is on average delivered to over 99% of its potential targets.

All activity relating to subject-URL pairs assigned to either experimental condition is logged, including feed exposures, censored exposures, and clicks to the URL (from the feed or other sources, like messaging). Directed shares, such as a link that is included in a private Facebook message or explicitly posted on a friend's wall, are not affected by the assignment procedure. If a subject-URL pair is assigned to an experimental condition, and the subject clicks on content containing that URL in any interface other than the feed, that subject-URL pair is removed from the experiment. Our experiment, which took place over the span of seven weeks, includes 253,238,367 subjects, 75,888,466 URLs, and 1,168,633,941 unique subject-URL pairs.

3.2 Ensuring Data Quality

Threats to data quality include using content that was or may have been previously seen by subjects on Facebook prior to the experiment, content that subjects may have seen through interfaces on Facebook other than feed, spam, and malicious content. We address these issues in a number of ways. First, we only consider content that was shared by the subjects' friends only after the start of the experiment. This enables our experiment to accurately capture the first time a subject is exposed to a link in the feed, and ensures that URLs in our experiment more accurately reflect content that is primarily being shared contemporaneously with the timing of the experiment. We also exclude potential subject-

Demographic Feature (% of subjects)	feed	no feed
Gender		
Female	51.6%	51.4%
Male	46.7%	47.0%
Unspecified	1.5%	1.5%
Age		
17 or younger	12.8%	13.1%
18-25	36.4%	36.1%
26-35	27.2%	26.9%
36-45	13.0%	12.9%
46 or older	10.6%	10.9%
Country (top 10 & other)		
United States	28.9%	29.1%
Turkey	6.1%	5.8%
Great Britain	5.1%	5.2%
Italy	4.2%	4.1%
France	3.8%	3.9%
Canada	3.7%	3.8%
Indonesia	3.7%	3.5%
Philippines	2.1%	2.3%
Germany	2.3%	2.3%
Mexico	2.0%	2.1%
226 Others	37.5%	37.7%

Table 1: Summary of demographic features of subjects assigned to the *feed* ($N = 160,688,092$) and *no feed* ($N = 218,743,932$) condition. Some subjects may appear in both columns.

URL pairs where the subject had previously clicked on the URL via any interface on the site at any time up to two months prior to exposure, or any interface other than the feed for content assigned to the *no feed* condition. Finally, we use the Facebook's site integrity system [40] to classify and remove URLs that may not reflect ordinary users' purposeful intentions of distributing content to their friends.

3.3 Population

The experimental population consists of a random sample of all Facebook users who visited the site between August 14th to October 4th 2010, and had at least one friend sharing a link. At the time of the experiment, there were approximately 500 million Facebook users logging in at least once a month. Our sample consists of approximately 253 million of these users. All Facebook users report their age and gender, and a user's country of residence can be inferred from the IP address with which she accesses the site. In our sample, the median and average age of subjects is 26 and 29.3, respectively. Subjects originate from 236 countries and territories, 44 of which have one million or more subjects. Additional summary statistics are given in Table 1, and show that subjects are assigned to the conditions in a balanced fashion.

3.4 Evaluating Outcomes

The assignment procedure allows us to directly compare the overall probability that subjects share links they were or were not exposed to on the feed. The causal effect of exposure via the Facebook feed on sharing is simply the expected probability of sharing in the *feed* condition minus the expected probability in the *no feed* condition. This quantity, known as the average treatment effect on the treated (or alternatively, the absolute risk increase), can vary when conditioning on other variables, including the number of friends and tie strength, which are analyzed in Sections 4 and 5. Alternatively, the difference in probabilities can be viewed as a ratio (the relative risk ratio), which quantifies how many times more likely an individual is to share as a result of being exposed to content on the feed.

Although the assignment is completely random, subjects and URLs may differ in ways that impact our measurements. For example, certain users may be highly active on Facebook, so that they are assigned to experimental conditions more often than other users. If these users were to vary significantly in terms of their information sharing propensities, such as sharing or re-sharing greater or fewer links than others, the disproportionate inclusion of these users may bias our measurements and threaten the population validity of our findings. Similarly, very popular URLs may also introduce biases; they may be more or less likely to be re-shared because of their inherent appeal or more likely to be discovered independently of Facebook because of their relative popularity amongst friends.

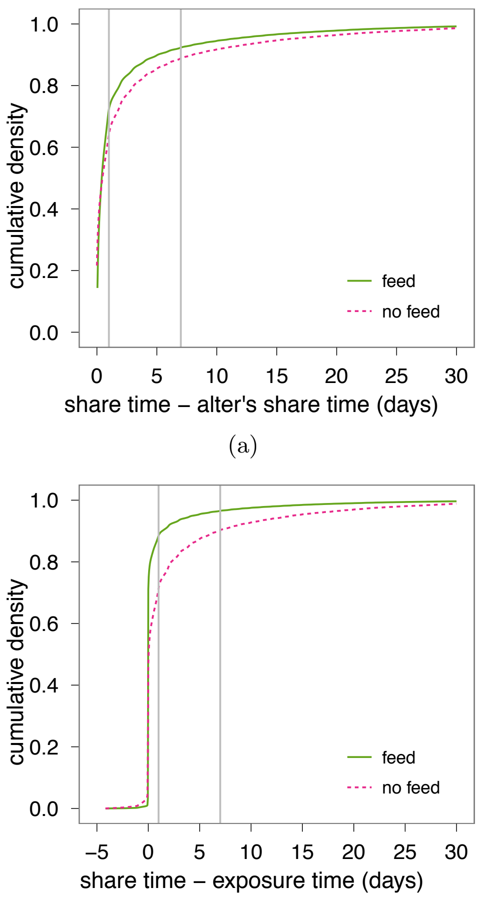
To provide control for these biases, we use bootstrapped averages clustered by the subject or URL. We find that in all of our analyses, clustering by the URL rather than the subject yields nearly identical probability estimates that have marginally wider confidence intervals, so we have chosen to present our results using means and 95% confidence intervals clustered by URL. Risk ratios are obtained using the 95% bootstrapped confidence intervals of likelihood of sharing in the *feed* and *no feed* conditions. To compute the lower bound of the ratio, we divide the lower bound of the probability of sharing in the *feed* condition by the upper bound for the *no feed* condition. The upper bound of the ratio is computed by dividing the upper bound in the *feed* condition by the lower bound of the *no feed* condition. The additive analog of the same procedure is used to obtain confidence intervals for probability differences.

4. HOW EXPOSURE TO SOCIAL SIGNALS AFFECTS DIFFUSION

We find that subjects who are exposed to signals about friends' sharing behavior are several times more likely to share that same information, and share sooner than those who are not exposed. To measure the relative increase in sharing due to exposure, we compute the risk ratio: the likelihood of sharing in the *feed* condition (0.191%) divided by the likelihood of sharing in the *no feed condition* (0.025%), and find that individuals in the *feed* condition are 7.37 times more likely share (95% $CI = [7.23, 7.72]$). Although the probability of sharing upon exposure may appear small, it is important to note that individuals have hundreds of contacts online who may see their link, and that on average one out of every 12.5 URLs that are clicked on in the *feed* condition are subsequently re-shared.

4.1 Temporal Clustering

Contemporaneous behavior among connected individuals is commonly used as evidence for social influence processes (e.g. [4, 9, 8, 15, 16, 19, 20, 25, 29, 36, 43]). We find that subjects who share the same link as their friends typically do so within a time that is proximate to their friends' sharing time, even when no exposure occurs on Facebook. Figure 3 illustrates the cumulative distribution of information lags between the subject and their first sharing friend, among



(a)

(b)

Figure 3: Temporal clustering in sharing the same link as a friend in the *feed* and *no feed* conditions. (a) The difference in sharing time between a subject and their first sharing friend. (b) The difference between the time at which a subject was first to exposed (or was to be exposed) to the link and the time at which they shared. Vertical lines indicate one day and one week.

subjects who had shared a URL after their friends. The top panel shows the latency in sharing times between the subject and their friend for users in the *feed* and *no feed* condition. While a larger proportion of users in the *feed* condition share a link within the first hour of their friends, the distribution of sharing times is strikingly similar. The bottom panel shows the differences in time between when subjects shared and when they were (or would have been) first exposed to their friends' sharing behavior on the Facebook feed. The horizontal axis is negative when a subject had shared a link after a friend but had not yet seen that link on the feed. From this comparison, it is easy to see that users in the *feed* condition are most likely to share a link immediately upon exposure, while those who share it without seeing it in their feed will do so over a slightly longer period of time.

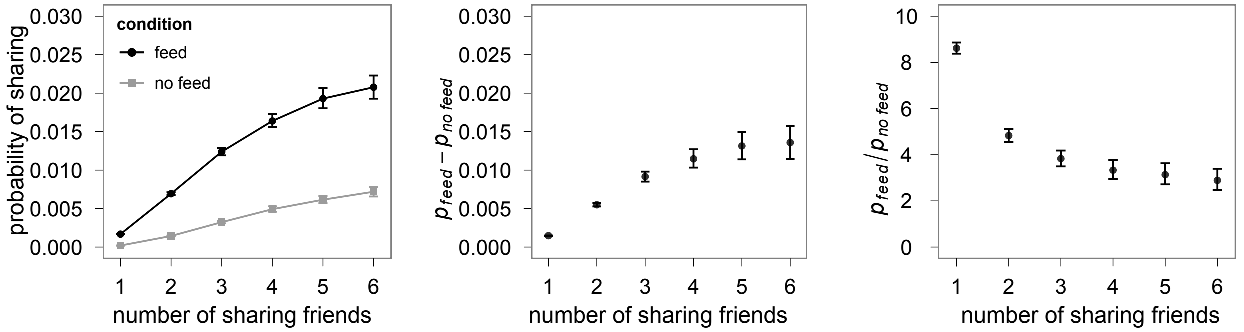
To evaluate how exposure on the Facebook feed relates to the speed at which URLs appear to diffuse, we consider URLs that were assigned to both the *feed* and *no feed* condition. We first match the share time of each URL in the *feed* condition with a share time of the URL in the *no feed* condition, sampling URLs in proportion to their relative abundances in the data. From this set of contrasts, we find that the median sharing latency after a friend has already shared the content is 6 hours in the *feed* condition, compared to 20 hours when assigned to the *no feed* condition (Wilcoxon rank-sum test, $p < 10^{-16}$). The presence of strong temporal clustering in both experimental conditions illustrates the problem with inferring influence processes from observations of temporally proximate behavior among connected individuals: regardless of access to social signals within a particular online medium, individuals can still acquire and share the same information as their friends, albeit at a slightly later point in time.

4.2 Effect of Multiple Sharing Friends

Classic models of social and biological contagion (e.g. [23, 35]) predict that the likelihood of "infection" increases with the number of infected contacts. Observational studies of online contagion [4, 9, 15, 30] not only find evidence of temporal clustering, but also observe a similar relationship between the likelihood of contagion and the number of infected contacts. However, it is important to note that this correlation can have multiple causes that are unrelated to social influence processes. For examle, if a website is popular among friends, then a particularly interesting page is more likely to be shared by a users' friends independent of one another. The positive relationship between the number of sharing friends and likelihood of sharing may therefore simply reflect heterogeneity in the "interestingness" of the content, which is clustered along the network: the more popular a page is for a group of friends, the more likely it is that one would observe multiple friends sharing it.

We first show that, consistent with prior observational studies, the probability of sharing a link in the *feed* condition increases with the number of contacts who have already shared the link (solid line, Figure 4a). But the presence of a similar relationship in the *no feed* condition (grey line, Figure 4a) shows that an individual is more likely to exhibit the sharing behavior when multiple friends share, even if she does not necessarily observe her friends' behavior. Therefore, when using observational data, the naïve conditional probability (which is equivalent to the probability of sharing in the *feed* condition) does not directly give the probability increase due to influence via multiple sharing friends. Rather, such an estimate reflects a mixture of internal influence effects and external correlation.

Our experiment allows us to directly measure the effect of the feed relative to external factors, computed as either the difference or ratio between the probability of sharing in the *feed* and *no feed* conditions (Figure 4bc). While the difference in sharing likelihood grows with the number of sharing friends, the relative risk ratio falls. This contrast suggests that social information in the feed is most likely to influence a user to share a link that many of her friends have shared, but the relative impact of that influence is highest for content that few friends are sharing. The decreasing relative effect is consistent with the hypothesis that having multiple sharing friends is associated with greater redundancy in information exposure, which may either be caused by homophily in visitation and sharing tendencies, or external influence.



(a) (b) (c)

Figure 4: Users with more friends sharing a Web link are themselves more likely to share. (a) The probability of sharing for subjects that were (*feed*) and were not (*no feed*) exposed to content increases as a function of the number sharing friends. (b) The causal effect of the feed is greater when subjects have more sharing friends (c) The multiplicative impact of the feed is greatest when few friends are sharing. Error bars represent the 95% bootstrapped confidence intervals clustered on the URL.

5. TIE STRENGTH AND INFLUENCE

Next, we examine the relationship between tie strength, influence, and information diversity by combining the experimental data with users' online and offline interactions. Following arguments originally proposed by Mark Granovetter's seminal 1973 paper, *The Strength of Weak Ties* [22], empirical work linking tie strength and diffusion often utilize the number of mutual contacts as proxies of interaction frequency. Rather than using the number of mutual contacts, which can be large for pairs of individuals who no longer communicate (e.g. former classmates), we directly measure the strength of tie between a subject and her friend in terms of four types of interactions: (i) the frequency of private online communication between the two users in the form of Facebook messages[1]; (ii) the frequency of public online interaction in the form of comments left by one user on another user's posts; (iii) the number of real-world coincidences captured on Facebook in terms of both users being labeled by users as appearing in the same photograph; and (iv) the number of online coincidences in terms of both users responding to the same Facebook post with a comment. Frequencies are computed using data from the three months directly prior to the experiment. The distribution of tie strengths among subjects and their sharing friends can be seen in Figure 5.

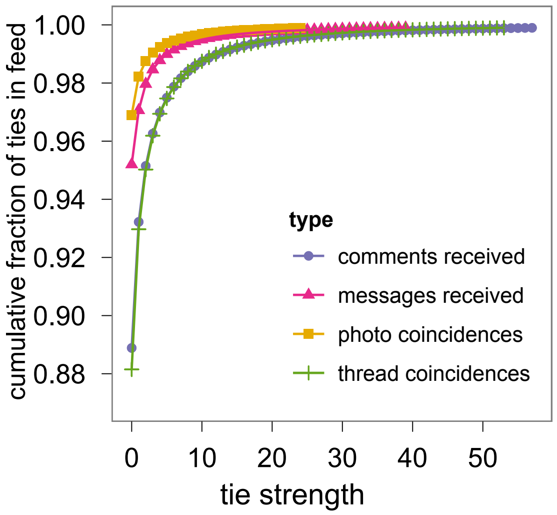
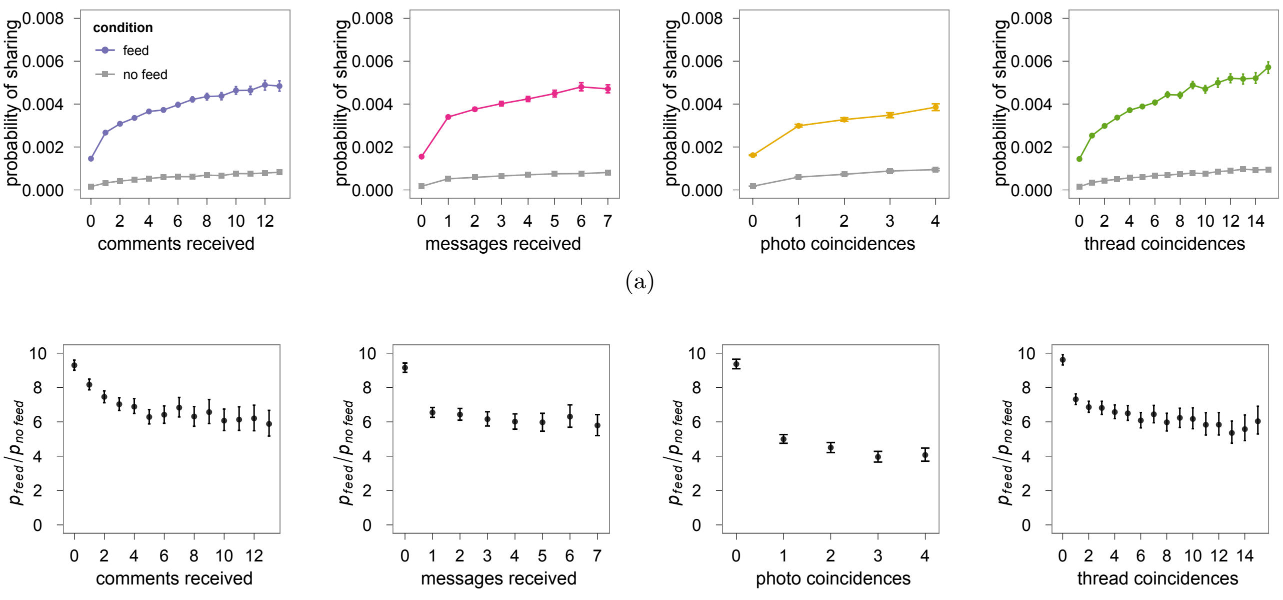


Figure 5: Tie strength distribution among friends displayed in subjects' feeds using the four measurements. Points are plotted up to the 99.9^{th} percentile. Note that the vertical axis is collapsed.

5.1 Effect of Tie Strength

We measure how the difference in the likelihood of sharing a URL in the *feed* versus *no feed* conditions varies according to tie strength. To simplify our estimate of the effect of tie strength, we restrict our analysis to subjects with exactly one friend who had previously shared the link. In both conditions, a subject is more likely to share a link when her sharing friend is a strong tie (Figure 6a). For example, subjects who were exposed to a link shared by a friend from whom the subject received three comments are 2.83 times more likely to share than subjects exposed to a link shared by a friend from whom they received no comments. For those who were not exposed, the same comparison shows that subjects are 3.84 times more likely to share a link that was previously shared by the stronger tie. The larger effect in the *no feed* condition suggests that tie strength is a stronger predictor of externally correlated activity than it is for influence on feed. From Figure 6a, it is also clear that individuals are more likely to be influenced by their stronger

[1] We quantify message and comment interactions as the number of communication events the subject received from their friend. The number of messages and comments sent, and the geometric mean of communications sent and received, yielded qualitatively similar results, so we plot only the single directed measurement for the sake of clarity.



(a)

(b)

Figure 6: Strong ties are more influential, and weak ties expose friends to information they would not have otherwise shared. (a) The increasing relationship between tie strength and the probability of sharing a link that a friend shared in the *feed* and *no feed* conditions. (b) The multiplicative effect of feed diminishes with tie strength, suggesting that exposure through strong ties may be redundant with external exposure, while weak ties carry information one might otherwise not have been exposed to.

ties via the feed to share content that they would not have otherwise spread.

Furthermore, our results extend Granovetter's hypothesis that weak ties disseminate novel information into the context of media contagion. Figure 6b shows that the risk ratio of sharing between the *feed* and *no feed* conditions is highest for content shared by weak ties. This suggests that weak ties consume and transmit information that one is unlikely to be exposed to otherwise, thereby increasing the diversity of information propagated within the network.

5.2 Collective Impact of Ties

Strong ties may be individually more influential, but how much diffusion occurs in aggregate through these ties depends on the underlying distribution of tie strength (i.e. Figure 5). Using the experimental data, we can estimate the amount of contagion on the feed generated by strong and weak ties. The causal effect of exposure to information shared by friends with tie strength k is given by the average treatment effect on the treated:

$$ATET(k) = p(k, feed) - p(k, no\ feed)$$

To determine the collective impact of ties of strength k, we multiply this quantity by the fraction of links displayed in all users' feeds posted by friends of tie strength k, denoted by $f(k)$. In order to compare the impact of weak and strong ties, we must set a cutoff value for the minimum amount of interaction required between two individuals in order to consider that tie strong. Setting the cutoff at $k = 1$ (a single interaction) provides the most generous classification of strong ties while preserving some meaningful distinction between strong and weak ties, thereby giving the most influence credit to strong ties.

Under this categorization of strong and weak ties, the estimated total fraction of sharing events that can be attributed to weak and strong ties is the average treatment effect on the treated weighted by the proportion of URL exposures from each tie type:

$$T_{weak} = ATET(0) * f(0)$$

$$T_{strong} = \sum_{i=1}^{N} ATET(i) * f(i)$$

We illustrate this comparison in Figure 7, and show that by a wide margin, the majority of influence is generated by weak ties[2]. Although we have shown that strong ties are individually more influential, the effect of strong ties is not large enough to match the sheer abundance of weak ties.

6. DISCUSSION

Social networks may influence an individual's behavior, but they also reflect the individual's own activities, interests, and opinions. These commonalities make it nearly impossible to determine from observational data whether any particular interaction, mode of communication, or social environment is responsible for the apparent spread of a behavior through a network. In the context of our study, there are three possible mechanisms that may explain diffusion-like phenomena: (1) An individual shares a link on Facebook,

[2] Note that for the purposes of this study, it is not necessary to model the effect of tie strength for users with multiple sharing friends, since stories of this kind only constitute 4.2% of links in the newsfeed, and their inclusion would not dramatically alter the balance of aggregate influence by tie strength.

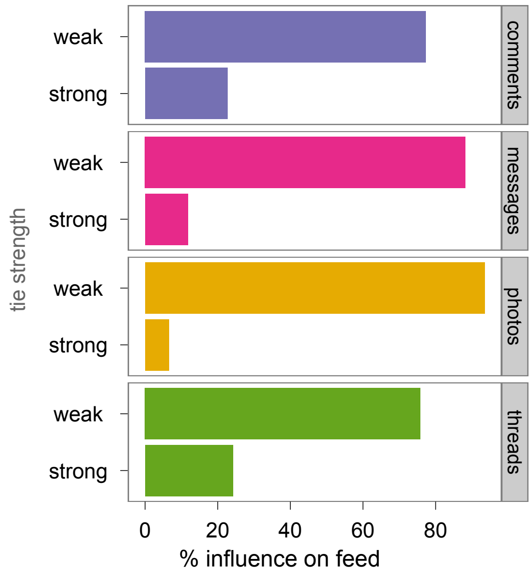


Figure 7: Weak ties are collectively more influential than strong ties. Panels show the percentage of information spread by strong and weak ties for all four measurements of tie strength. Although the probability of influence is significantly higher for those that interact frequently, most contagion occurs along weak ties, which are more abundant.

and exposure to this information on the feed causes a friend to re-share that same link. (2) Friends visit the same web page and share a link to that web page on Facebook, independently of one another. (3) An individual shares a link within and external to Facebook, and exposure to the externally shared information causes a friend to share the link on Facebook. Our experiment determines the causal effect of the feed on the spread of sharing behaviors by comparing the likelihood of sharing under the *feed* condition (possible causes 1-3) with the likelihood under the *no feed* condition (possible causes 2-3).

Our experiment generalizes Mark Granovetter's predictions about the strength of weak ties [22] to the spread of everyday information. Weak ties are argued to have access to more diverse information because they are expected to have fewer mutual contacts; each individual has access to information that the other does not. For information that is almost exclusively embedded within few individuals, like job openings or future strategic plans, weak ties play a necessarily role in facilitating information flow. This reasoning, however, does not necessarily apply to the spread of widely available information, and the relationship between tie strength and information access is not immediately obvious. Our experiment sheds light on how tie strength relates to information access within a broader context, and suggests that weak ties, defined directly in terms of interaction propensities, diffuse novel information that would not have otherwise spread.

Although weak ties can serve a critical bridging function [22, 37], the influence that weak ties exert has never before been measured empirically at a systemic level. We find that the majority of influence results from exposure to individual weak ties, which indicates that most information diffusion on Facebook is driven by simple contagion. This stands in contrast to prior studies of influence on the adoption of products, behaviors or opinions, which center around the effect of having multiple or densely connected contacts who have adopted [6, 7, 14, 13]. Our results suggest that in large online environments, the low cost of disseminating information fosters diffusion dynamics that are different from situations where adoption is subject to positive externalities or carries a high cost.

Because we are unable to observe interactions that occur outside of Facebook, a limitation of our study is that we can only fully identify causal effects within the site. Correlated sharing in the *no feed* condition may occur because friends independently visit and share the same page as one another, or because one user is influenced to share via an external communication channel. Although we are not able to directly evaluate the relative contribution of these two potential causes, our results allow us to obtain a bound on the effect on sharing behavior within the site. The probability of sharing in the *no feed* condition, which is a combination of similarity and external influence, is an upper bound on how much sharing occurs because of homophily-related effects. Likewise, the difference in the probability of sharing within the *feed* and *no feed* condition gives a lower bound on how much on-site sharing is due to interpersonal influence along any communication medium.

The mass adoption of online social networking systems has the potential to dramatically alter an individual's exposure to new information. By applying an experimental approach to measuring diffusion outcomes within one of the largest human communication networks, we are able to rigorously quantify the effect of social networks on information spread. The present work sheds light on aggregate trends over a large population; future studies may investigate how properties of the individual, such as age, gender, and nationality, or features of content, such as popularity and breadth of appeal, relate to the influence and its confounds.

7. ACKNOWLEDGMENTS

We would like to thank Michael D. Cohen, Dean Eckles, Emily Falk, James Fowler, and Brian Karrer for their discussions and feedback on this work. This work was supported in part by NSF IIS-0746646.

8. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

[2] E. Adar and A. Adamic, Lada. Tracking information epidemics in blogspace. In *2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne University of Technology, France, 2005.

[3] E. Adar, J. Teevan, and S. T. Dumais. Resonance on the web: web dynamics and revisitation patterns. In *Proceedings of the 27th International Conference on Human factors in Computing Systems*, CHI '09, pages 1381–1390, New York, NY, USA, 2009. ACM Press.

[4] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th Internal Conference on*

Knowledge Discover & Data Mining, pages 7–15, New York, NY, USA, 2008. ACM Press.

[5] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci.*, 106(51):21544–21549, December 2009.

[6] S. Aral and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, Aug. 2011.

[7] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM.

[8] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *3rd ACM Conference on Web Search and Data Mining*, Hong Kong, 2011. ACM Press.

[9] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the tenth ACM conference on Electronic commerce*, pages 325–334. ACM, 2009.

[10] H. R. Bernard, P. Killworth, D. Kronenfeld, and L. Sailer. The problem of informant accuracy: The validity of retrospective data. *Annu. Rev. Anthropol.*, 13:495–517, 1984.

[11] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *J. Consumer Research*, 14(3):pp. 350–362, 1987.

[12] R. S. Burt. *Structural holes: The social structure of competition.* Harvard University Press, Cambridge, MA, 1992.

[13] D. Centola. The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996):1194–1197, September 2010.

[14] D. Centola and M. Macy. Complex contagions and the weakness of long ties. *Am. J. Sociol.*, 113(3):702–734, Nov. 2007.

[15] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[16] N. A. A. Christakis and J. H. H. Fowler. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.*, 357(4):370–379, July 2007.

[17] S. Fox. The social life of health information. Technical report, Pew Internet & American Life Project, 2011.

[18] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, CHI '09, pages 211–220, New York, NY, USA, 2009. ACM.

[19] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference.* Little Brown, New York, 2000.

[20] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1019–1028, New York, NY, USA, 2010. ACM.

[21] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 241–250, New York, NY, USA, 2010. ACM.

[22] M. S. Granovetter. The strength of weak ties. *Am. J. Sociol.*, 78(6):1360–1380, May 1973.

[23] M. S. Granovetter. Threshold models of collective behavior. *Am. J. Sociol.*, 83(6):1420–1443, 1978.

[24] B. S. Greenberg. Person to person communication in the diffusion of news events. *Journalism Quarterly*, 41:489–494, 1964.

[25] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

[26] K. Hampton, L. S. Goulet, L. Rainie, and K. Purcell. Social networking sites and our lives. Technical report, Pew Internet & American Life Project, 2011.

[27] S. Hill, F. Provost, and C. Volinsky. Network-Based marketing: Identifying likely adopters via consumer networks. *Stat. Sci.*, 21(2):256–276, May 2006.

[28] G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network. *Am. J. Sociol.*, 115(2):405–450, September 2009.

[29] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[30] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM.

[31] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633, 2008.

[32] C. F. Manski. Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.*, 60(3):531–42, July 1993.

[33] A. Marin. Are respondents more likely to list alters with certain characteristics? Implications for name generator data. *Social Networks*, 26(4):289–307, Oct. 2004.

[34] M. McPherson, L. S. Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.*, 27(1):415–444, 2001.

[35] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, Jul 2002.

[36] J.-P. Onnela and F. Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, 2010.

[37] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási.

Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.

[38] K. Purcell, L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead. Understanding the participatory news consumer. Technical report, Pew Internet & American Life Project, 2010.

[39] C. R. Shalizi and A. C. Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods and Research*, 27:211–239, 2011.

[40] T. Stein, E. Chen, and K. Mangla. Facebook Immune System. In *EuroSys Social Network Systems*, 2011.

[41] E. S. Sun, I. Rosenn, C. A. Marlow, and T. M. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the 3rd Int'l AAAI Conference on Weblogs and Social Media*, San Jose, CA, 2009. AAAI.

[42] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[43] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *ACM Conference on the World Wide Web*, Hyderbad, India, 2011. ACM Press.

Information Polity 18 (2013) 43–56
DOI 10.3233/IP-2011-0244
IOS Press

Networked campaigns: Traffic tags and cross platform analysis on the web

Greg Elmer[a,*] and Ganaele Langlois[b]
[a]*Ryerson University, Toronto, Canada*
[b]*Department of Communication, University of Ontario Institute of Technology, Oshawa, Canada*

Abstract. This article defines a new methodological framework to examine emerging forms of political campaigning on and across Web 2.0 platforms (i.e. Facebook, Youtube, Twitter) in the North-American context. The proposed method seeks to identify the new strategies that make use of campaign texts, users, keywords, information networks and software code to spread a political communications and rally voters across distributed, and therefore seemingly unmanageable spheres of online communication. The proposed method differentiates itself from previous Web 1.0 methods focused on mapping hyperlinked networks. In particular, we pay attention to the new materiality of the Web 2.0 as constituted by shared objects that circulate across modular platforms. In this paper we develop an object-centered method through the concept of *traffic tags* – unique identifiers that by enabling the circulation of web objects across platforms organize political activity online. By tracing the circulation of traffic tags, we can map different sets of relationships among uploaded and shared web objects (text, images, videos, etc.), political actors (online partisans, political institutions, bloggers, etc.), and web based platforms (social network sites, search engines, political websites, blogs, etc.).

1. The challenge of 2.0 networking

Politics has always been about networking. Well before seeking office prospective candidates are advised to identify well-connected individuals – those who can help raise funds, make insider connections in party circles, and otherwise "open doors". And while political networking today still requires face-to-face meetings, it now also requires a virtual dimension, one that raises significant opportunities and pitfalls for campaigns and political life in general. For candidates, political party strategists and communications staff, social media (such as *Facebook*, *Twitter*, and *Youtube*) offer distinct opportunities to reach segmented communities and to narrowcast messages to party members in specific electoral ridings and districts, regionally, or nationally, at particular times of the day, and for specific purposes (campaign stops, stump speeches, fundraising, candidate nomination meetings, leadership contests, etc.). Yet at the same time social networking sites also challenge the ability to control and otherwise manage so-called talking points[1], election policy and platforms, and more broadly overarching election campaign "scripts". Indeed, A message, image, or video can be shared with political opponents and remixed or critiqued in very short order. Networked political communication, in other words, has become mutable, evasive, and much more difficult to manage in the social media universe. For political actors the sheer

[*] Corresponding author. E-mail:gelmer@ryerson.ca.

[1] See for example the publication of in-camera party "talking points" on new 2.0 sites such as "wikileaks". https://secure.wikileaks.org/wiki/Canadian_Conservative_Party_May_Constituency_Week_Caucus_Pack%2C_May_2009<accessed May 16, 2009>.

number of new media spaces, 2.0 platforms, social networks, and information aggregators , complicate the ability to deploy contemporary political campaigns. Where does one start? What should one share? Or reserve solely for party supporters? How should one respond to political attacks and rumours on social media? Gone are the days when political strategists focused exclusively on editorial boards of newspapers, briefing notes and stump speeches for media campaign buses and planes, and fund raising letters.

In light of such radical changes in the information and communications sphere, political scientists and communication scholars have sought to develop new experimental methods of understanding this new digitally networked terrain of politics [7]. Web 1.0 studies – those primarily concerned with political communication, organizing and networking on the world wide web – sought to develop methods of mapping hyperlinked relationships among web sites for political candidates, parties, and civil society organizations, to name but a few. Building upon earlier forms of social network analysis, hyperlink networking methods and tools [29], notably programmable "crawlers" that jumped from link to link, sought to identify key political actors or "hubs" in networked hyperlink diagrams [19,22,31]. Such research sought to locate the most influential political actors on the web through identifying the most linked-to web pages. For our purposes here we refer to such forms of analysis as *http methods*, in recognition of their use of one sole form of code that link together html documents on the world wide web: the HREF (or hyperlink) command [15].

The reliance upon the hyperlink as sole indicator of techno-political association both online and offline [15,32, p. 38], however, has not been without its skeptics [23]. Elmer [11] has argued that hyperlink mapping faces numerous technological hurdles as web servers often crash and need to go-offline for routine maintenance. Web sites and pages are often blocked for a host of other reasons, politically, inter-regionally or otherwise. Thus, researching political connections and associations on the web requires one to also recognize disconnected or disrupted forms of networked computing. Whatever one thinks of early forms of hyperlink analysis, such methods clearly contributed to innovative forms of data visualization, attempts to more accurately – or perhaps more creatively – represent distributed forms of political networking[2]. New data visualization software[3], some representing a seemingly infinite number of hyperlinks[4], however, often produce undecipherable, death star-like maps of hubs and spokes, posing significant challenges to meaningful forms of analysis [5]. Hyperlink maps, furthermore, only render and visualize f*unctional* hyperlinks and websites at specific moments. In other words, where are the network maps for example that denote disconnections, server timeouts and crashes, and deleted links between sites? Such positivism, in both senses of the term – meaning successful, and empirically verifiable links – in the absence of various forms of disconnectivity and dysfunction, in our opinion, reify political networking *as successful forms of connectivity*. Political networking (much like computer networking) is, however, often quite the opposite: laboured, unstable, precarious, unverifiable, sometimes unconscious, and hidden. How might such forms of research therefore acknowledge such qualitative distinctions in and across such networks?

This paper sets the stage for another approach to the study of internet politics and networking, one that addresses the impact that new web 2.0 interactive platforms have had upon what we refer to as the conditions of networked connectivity. By *conditions*, we again suggest that connectivity itself has

[2] Cf < http://manyeyes.alphaworks.ibm.com/manyeyes/>.

[3] For a list of representative software see www.visualcomplexity.com.

[4] Cf this representation of hyperlinks among political blogs in June 2008: <http://simoncollister.typepad.com/.shared/image.html?/photos/uncategorized/2008/06/26/polblogo.jpg>.

been largely understudied, or worse – interpreted as either a sign of political alliance, support, or merely "successful" connection. In contrast, this paper offers the building blocks for methods that attempt to account for connection failures, disruptions, and roadblocks, some "accidental", others the obvious result of restrictive terms of use encoded into web 2.0 platforms (and their application programming interface, protocols and algorithms). By focusing on the conditions of connectivity, we seek to integrate user based experiences and of course their shared, remixed, and uploaded digital objects[5] into the broader research paradigm[6]. This involves mapping networking (file sharing, etc.) opportunities and restrictions on the one hand, and dysfunctions and incompatibilities on the other.

In the process of developing new methods for studying the relationship between political actors, objects, and platforms online, this paper first offers a brief "meta-tag" analysis of political keywords (text) on the world wide web as a test case for demonstrating how non-hyperlink forms of software code can also provide insight into networked political campaigns on the world wide web. After some initial reflections and analysis of our "tag" based study of political networking, we will then discuss how such "tags" operate in the much more complex world of the web 2.0, where users are increasingly called upon to self-categorize (through titles, keywords, hash tags, etc.) their online contributions (images, blog posts, tweets, comments, videos, etc.). The paper concludes with an initial effort at further expanding and analyzing how a plethora of 2.0 based forms of user and automatically generated software code can be harnessed to better understand the possibilities and constraints of political networking across a number of web sites and 2.0 platforms (eg. twitter, Facebook, Youtube, blogs, etc.). The ultimate goal of this longer term project is to offer methods and tools that might diagnose the possible reach of online political campaigns, communications, and networks. Our approach seeks to determine the constitution and constraints afforded by different sets of relationships among uploaded and shared web objects (text, images, videos, etc.), political actors (online partisans, political institutions, bloggers, etc.), and web based platforms (social network sites, search engines, political websites, blogs, etc.).

To this end, and in moving from so-called web1.0 http or html approaches to 2.0 cross platform based methods, this paper is particularly interested in harnessing, methodologically speaking, user-generated forms of classification – or *tags* to use the net-vernacular. Such forms of text/keywords are commonly used by social media partisans and activists to associate their online contributions (blog posts, *Youtube* videos, etc.) to likeminded political and social debates, actors, sites, platforms, and other online objects. To identify the relationships – the networks – forged by objects, actors, and platforms, however, this paper also makes the case for identifying discrete forms of communication and networking *in motion*, that is as internet network *traffic*. While http based hyperlink analysis offered a means of identifying relationships among web sites and their assumed owners/webmasters, our *traffic tag* approach seeks to determine the multiplicity of avenues (across web 2.0 platforms) – or conversely dead ends – that limit the reach and political possibilities of online campaigns. Only through tracking the unique forms of ID associated with platforms (eg. through their URLs), online political actors (eg. their accounts, usernames, etc.), or networked objects (titles, URLs, etc.) can we begin to diagnose the possibilities and pitfalls of 2.0 political networking, communications, and campaigning.

2. Trafficking political rhetoric – "Stand up for Canada"

In this section of the paper, we offer a brief analysis of how meta-tag keywords on the world wide web can be harnessed and analyzed to understand the reach and circulation of political campaigns on

[5] Most notably videos, digital images, blog posts, twitter posts, shared hyperlinks, etc.

[6] Cf. Hindman's [22] *The Myth of Digital Democracy* for a good overview of http based methods of network analyses.

the internet. The study offers a glimpse into why subsequent 2.0 forms of analysis need to take into consideration the role that self and automatically generated tags play in the generation of possible avenues for networked political content (objects) and actors across a number of popular 2.0 platforms. So as not to overstate the novelty of our proposed method of research, of mapping political networks, issues, actors, and objects across the 2.0 universe, it is important to note that the building blocks of a more nuanced, 2.0 enabled form of network mapping or "traffic tags" approach to the study of political campaigns, were to a much lesser degree present on the world wide web. While HTML encoded web pages offered HREF tags (hyperlinks) as conduits for network mapping, http header meta-tag keywords and other meta data have also afforded other opportunities for qualifying and expanding network analysis[7]. One such line of inquiry has focused on the relationship between websites and their visibility and ranking via industry leading search engines. Google's indexing bots, for example, "read" the http header keywords of html web pages so that they can be better integrated into Google's archival, page ranking, information aggregation, commercial advertising, and user profiling functions [21]. Webmasters thus encode their websites' header keywords to sufficiently represent their sites' content, enabling accurate indexing from Google and other information aggregators. Such keywords thus link web sites to web aggregators, most notably Google via its "page rank" algorithm [4].

Political consultants and campaign staff in the most recent American presidential election were quick to recognize the many different techniques that campaigns could use to better "optimize" their candidate's visibility on the web by refining titles and other keywords in the headers of campaign web pages [8]. Similarly, the home page for the Conservative party of Canada includes rather obvious meta tag keywords such as "conservative party" and "Stephen Harper" (the Canadian prime minister). However, reviewing the http header – that one can easily do by choosing the "view > page source" pull down menu on most web browsers – also reveals the strategic insertion of a recent election campaign slogan "Stand up for Canada", and a short list of political issues and buzzwords: "trade, transit, accountability, childcare, etc."[8]. While Conservatives in Canada strategically use such tags to brand their political campaigns and messages, web masters as a whole can dream up and encode their http header with any sequence of keywords, tactically deployed to gain greater Google-visibility (higher ranking), resulting in increased traffic to their site[9]. In lieu of considering these connections between websites as networked associations then, we should also consider the view that such keywords serve to self-identify web pages and cultivate new sources of traffic. The "tagging" of one's content – through the use of keywords – suggests a degree of self-promotion, a form of publicity, that from time to time stretches the indexical purpose of such meta tag keywords[10].

[7] The British Liberal democrat party encodes a geo-tag in their http header that Identifies their location as Westminister, UK.

[8] <http://www.conservative.ca/> , under view>page source option. Accessed April 8, 2009.

[9] This tactic is often referred to as "meta-tag stuffing", it falls under the less subjective term "search engine optimization". The topic has been vigorously debated by lawyers worldwide [28].

[10] The most blatant example of so-called "meta-tag stuffing" therein refers to nefarious attempts to try to latch on to popular or trendy keywords that users use as search words on Google to increase internet traffic to web sites – a form of traffic spam if you will. The de-regulated nature of meta tag html page encoding thus raises broad questions and concerns about the over-promotion of certain content (porn, dubious credit cards, etc.) and the burying of perhaps more socially relevant information. Ira S. Nathenson (1998) draws a rather clever yet frustrating analogy of a "spamdexed" network:
Imagine a never-ending traffic jam on a ten-lane highway. Road signs can't be trusted: the sign for Exit 7 leads to Exit 12, the sign for Cleveland leads to Erie. If you ask the guy at the Kwik-E-Mart how to get to I-79, he gives you directions to Route 30. To top it off, when you ask for a Coke, he gives you a Pepsi. Enough already. You stop at a pay phone to call directory assistance for the number to the local auto club, and instead get connected to "Dial-a-porn." (p. 45).

A brief search of the "Stand up for Canada" phrase, using the Google search engine, offers an glimpse into the circulation and adoption of such politically loaded and "genetically" encoded[11] words from the Conservative Party's website. Google results for the Conservative's phrase "Stand up for Canada", for example, provides an intriguing picture of the numerous web sites and 2.0 platforms that repeat, adopt, or otherwise circulate the phrase[12]. In addition to a page from the Conservative Party website that re-uses the phrase as a generic headline for political reaction to a constitutional crisis that emerged shortly after the Canadian federal election in 2008, Google also returns the following results:

Web platform	Content
1. Conservative Party Website	Political content, using phrase as headline
2. Conservative Party Website	Home page, phrase used as main header-banner
3. Childcare resource center	Archive of Conservative party platform document that used the phrase in its title
4. Youtube	2 Youtube videos, i) critical of the PM, using phrase in title and in content ii) phrase included in title and description of video critical of North American Union policies
5. United Steelworkers website	"Stand up for Canada: Save Manufacturing" advocacy article.
6. Political website	Uses the phrase to critique a wide set of government policies.
7. personal blog	"Time for CRTC to Stand up for Canada" title for blog post
8. Prime Minister's Facebook page	Headline to same article as #1 result, reproduced for Facebook.
9. Government Web page	Speaking notes for government minister that uses phrases in title and 3 times in body of speech
10. PM's Myspace page	Reproduction of #1 and #8.

[13]

Through this brief glimpse of meta-tag keywords one can make a series of preliminary though important methodological conclusions and claims, the most broadest of all supporting our contention that certain web based tags – words inserted into a HTTP header by webmasters in this 1.0 case study – can be used in much the same way that hyperlink analysis has been deployed, that is to track the relationship between and dissemination of digital objects, issues-language, coordinated campaigns, and lastly, political actors. While the nature of digital objects tends to multiply exponentially in a 2.0 web environment, a keyword and tag based method of analysis conducted above, is largely restricted to the study of plain text, political keywords or short catch phrases used to symbolize ideologies, policies, and legislative priorities. However, by tracking, albeit rather simplistically, the dissemination of such keywords across the web – as aggregated by *Google* – we can also catch a glimpse of the spread and adoption of such political keywords

[11] By using this biological term we mean to suggest that such http headers tags and keywords serve to implicate and reproduce both political languages and possible sites for articulating, networking, and organizing political agendas.

[12] Since this search was conducted in July 2009, the results discussed here offers a significant "time delayed" picture of the Conservative slogan – one that provides, perhaps, a more steeped view of the spread, adoption, and reuse of the phrase.

[13] A lit of the URLs for the "Stand up for Canada" search (July 9, 2009).
1. www.conservative.ca/EN/2459/107759, 2. <www.conservative.ca>, 3. http://action.web.ca/home/crru/rsrcs_crru_full.shtml?x=84178&AA_EX_Session=c8b1cacfb93b7da41cf1b4f974865afd>, 4. i) <http://www.youtube.com/watch?v=Dgp7-XjQ7rg> ii) < http://www.youtube.com/watch?v=9CrR0UYrnq8>, 5. http://www.uswa.ca/program/content/4606.php, 6. <http://www.titanrainbow.com/garydavidson/betrayed.html>, 7. <http://harveyoberfeld.ca/blog/time-for-crtc-to-stand-up-for-canada/>, 8. <http://www.facebook.com/note.php?note_id=42004436572>, 9. http://www.hrsdc.gc.ca/eng/corporate/newsroom/speeches/blackburnjp/070925.shtml>, 10. <http://blogs.myspace.com/index.cfm?fuseaction=blog.view&friendId=405845189&blogId=453480096>.

and slogan (e.g. "Stand up for Canada"), by whom (actors), in what political context (coordinated campaign, or political retort), and across specific platforms (2.0 social network sites or otherwise). For example, five of the top ten results of the phrase aggregated and ranked by Google emanated from either the government of Canada or its ruling political party (the Conservative party of Canada). The keywords are most commonly associated with three identical texts, a political document circulated by the Conservative party of Canada that attacks Canada's opposition parties. Results #1, 8, and 10, in other words clearly demonstrate a coordinated, cross-platform campaign by the Conservative party to utilize a title ("Stand Up for Canada"), to frame a word for word verbatim attack on their political opponents. Result #2, furthermore suggests that the party is also using the heading as a more generic keyword to frame its broader P.R. strategy. The ninth result, where the phrase is found in the full text of a speech delivered by a Conservative government minister, demonstrates that the phrase "Stand up for Canada" is also used not only for partisan purposes, but also as a key political phrase repeated in public and policy settings. The third result for the phrase also points to the phenomenon of third parties, in this case a Childcare resource center, archiving certain government and political documents for, presumably, their own political use, such as lobbying purposes and internal membership campaigns. Opponents and critics of the Conservative government are equally accounted for in Google's top ten results for the meta-tag phrase "Stand up for Canada". Two user-generated 2.0 sites, a blog, and a Youtube account clearly attempt to usurp the government phrase for critical purposes, as does to a lesser extent a manufacturing advocacy piece from the website of the United Steelworkers union.

From this brief analysis of embedded html keywords then, one can clearly see that this political phrase "Stand Up for Canada" is a contested one online, bringing together party communications staff, government departments-ministers, interest groups covering industrial and social issues (steel workers, and childcare advocates), and social media users. This brief analysis shows that the phrase circulates across established HTML web sites, to blogs, top English language social networking sites Facebook and Myspace, and the popular Youtube social media aggregator. Objects, actors and political campaigns become increasingly remediated across social media and web 2.0 platforms, and as such the need to develop a traffic tag approach to the study of political networking takes on an even greater sense of urgency.

3. Social media: The sharing of objects

Since much of this paper presumes a radical shift in web operability (from 1.0 to 2.0), some important conceptual remarks on social media are required to establish the building blocks of a 2.0 method of researching political networking. This is particularly urgent for, as a concept, Web 2.0 feels a bit like a black hole: everything gets trapped within its porous boundaries, from commercial and private social networks to the collaborative site *Wikipedia*, from the latest online social networking craze *Twitter* to the one of the first and enduring successful online business model, *Amazon.com* (O'Reilly, 2005). That said, mainstream discourse about Web 2.0 often refers to a projected perception of the contemporary state of the World Wide Web as correcting the shortcomings of the previous Web 1.0 era and fostering a democratically infused and dis-intermediated commercial sector [2,14]. Thus, while *YouTube*, *Facebook*, and *Wikipedia* each emphasize different functions, media, and business models, all are intensely reliant upon user-generated content. To clarify, Web 2.0 largely relies on users to not only produce and upload content, but more importantly, to share and circulate it across friends networks of like-minded individuals and groups. Social networks on sites like *Facebook*, *Myspace*, *Bebo*, *Cyworld* and others are in effect produced by the sharing of objects on their sites. Without such trafficking of objects (links, images, videos, text, etc.), the owners of such sites would be unable to aggregate and data mine personal information from users and their like-minded friends. Similarly, the popularity of *YouTube* is not simply linked to its capacity to act as a repository or archive of videos – rather it continues to grow as a result of its ability to the share, through embedded code, videos on a number of platforms across the web [20]. Web 2.0, in other words, relies upon shared objects – and avenues for circulating said objects – that link together individual users and their networking affinities. We like to think of such avenues and objects as "friendly traffic", of course not to downplay the fact that such sites subsequently aggregate user's psychographics, profiles and online behaviours to sell "targeted" advertising [34]. The focus on such friend-based traffic – the sharing of objects on and across social media platforms – thus calls into question the architecture of social media, as themselves objects of research and analysis. Political research on social media, in other words, must take into account not only users (be they political partisans, or institutions) but also the possibilities that social media platforms afford on their sites – the opportunities and roadblocks of uploading, of sharing, and networking across the web, hand-held devices, and beyond.

4. 2.0 Networking: From universal protocols to unique identifiers

To begin to map and analyze the circulation of objects, actors, and broader networked campaigns on the web today, we argue for a cross-platform approach – a method that seeks to determine the networking opportunities and limitations among and across so-called web 2.0 sites. A methodology that would witness the unfolding of the circulation of virtual political campaigns and networks via Web 2.0 platforms would be of considerable benefit in terms of identifying specific networking opportunities, limitations, and pitfalls in the political sphere. The first step in developing such a perspective requires a move beyond, and below the user interface. That is, we need to challenge our perception of the Web as rooted within the visual aesthetics of the user interface. This is all the more crucial and challenging on proprietary and closed websites such as *Facebook*, the interface becomes a limiting factor as our only point of entry is through the customized or, should we say, personalized (1st person, that is) perspective of our own networked environment. Web 2.0 social networking is in other words by definition an intensely personalized medium, no two *Facebook* interfaces and accompanying "friend" networks are the same. We all see – and operate within – *Facebook* through the contours of our own social networks. Such networks bias, and to a degree determine, the searches we perform via *Facebook's* search window, skewing the results to highlight our own aggregated friend-network-profiles. No two search results via *Facebook*, in other words, are alike – even for the exact same search term. Thus we can never have access to the totality or even common set of information available on *Facebook* via the interface – and as network researchers this always-already personalized interface and algorithm complicates our ability to analyze from third person perspectives, that is from the "outside". Indeed, the user perspective creates an oddly narcissistic worldview of Web 2.0 – one individuated through a me-centric (and thus uncannily familiar) network-interface. Adopting a cross-platform perspective, however, helps to overcome the limitation of the user worldview by disaggregating objects, actors, and networks from 2.0 user accounts.

Web 2.0 protocols are largely concerned with managing users and user-generated content 'objects', connections that enable relationships that populate networks across Web 2.0. In other words, Web 2.0 platforms set up the channels through which information can circulate. Our proposed method, in turn, seeks to develop tools to track, map and visualize such channels or traffic routes. Such an approach has roots in the critical aesthetics of software studies – for instance, Fuller's *Webstalker* [17], an alternative Web browser that simply sought to represent the linked relationships between websites, a browser devoid of any aggregated information or iconic graphics. Our critical approach to Web 2.0 platforms likewise requires a process of disaggregating the relationship between interface and back-end code and protocols, a form of reverse engineering, if you will. The building blocks of a disaggregated net, as previously stated, begin with a process of identifying the key components in political/computer networking – actors, objects, and platforms – each of which contain unique forms of ID, including user-generated tags. Once we can identify each of these actors and objects on the net, we can then map the traffic or the routes of such IDs-tags, to determine how and where political campaigns circulate across the web.

Such "traffic tags" serve to not only organize cross-platform communication but also to enable connections across different actors and to organize online activity. Our focus on traffic tags emerged from a realization that there is a need to include the beyond and below the discursive dimension of online content, and from an acknowledgement that what used to be discrete Web objects have morphed into entities capable of enabling different forms of connection simultaneously at different levels. By beyond and below the discursive dimension of online content, we mean the material aspects and social effects of political content networked across Web 2.0 platform. Below content encompasses the data processes and network routes through which content is circulated and published. Beyond content refers to the capacity of content to not simply represent, but more crucially in the online political context, to organize and spur action (i.e. voting, fundraising, protesting). Furthermore, the acknowledgement of the morphing of web objects into traffic tags offers a methodological incentive to pay closer attention to the beyond and below aspects of online content.

Let's use Barack Obama's famous political phrase "Yes We Can" by way of example. "Yes We Can", as a rallying cry, a lasting rhetoric gesture, and as the summation of an expansive, and expensive political campaign, should be considered as a brand, that is, as a "platform for the patterning of activity, a mode of organizing activity in time and space" [26, p. 1]. What are the aspects of patterning and organization expressed through the online circulation of "Yes We Can"? First, the online "Yes We Can" is a multi-dimensional Web object: it is a rhetorical logos, a cultural symbol to which are associated a range of media objects (official texts, videos and pictures, citizen responses, critiques and parodies) It is also, as a link object, a deictic or pointer [9,10,33] to different platforms (the official campaign website, the Facebook page, other websites).Under its repurposable form as a button that can be embedded in individual Facebook pages, blogs and websites, it is a form of political action to declare allegiance and vote intention. As an application, especially a Facebook application developed by Obama campaign staff, it serves as a covert polling technology – to cull more information on supporters and would-be voters. As such, "Yes We Can" is a multilevel traffic tag that serves to organize and centralize different types of activity. From the point of view of the user, "Yes We Can" is both a content and a deictic pointer to a broader community of like-minded individuals. At the political level, the importance of the "Yes We Can" logo is not simply that it is a symbolic rallying cry, but that is also an operative one that can quantify its effects by being turned into a tracking device and enable precise quantification of the reach of a message. From a computer-networking point of view, "Yes We Can" is the user-understandable facet of a range of data processing that aims to identify and link relevant information, according to different platform logics. For instance, while the Google search engine logic aims to identify the most relevant material for the large population of users, the Facebook search engine will operate through a logic of personalization, such as friends' preference, and geographic proximity. Traffic tags are thus operators that allow for the conjunction of multiple modes of organization, of connection of different actors – for instance political rallying and web tracking. As such, they express multiple practices that aim to organize political relationships, political discourse and informational networks. For this reason, traffic tags should be considered as objects of analysis to better understand political activity across Web platforms, as well as analytical objects through which we can derive new methodologies for tracking the unfolding of online political campaigns, communications, and networks.

"Traffic tags" can be human-generated, such as the title of video, or the formal name of a user as they appear on the user-interface, or the user tags that describe how an object belongs to a class of object (i.e. 'X's wedding' or 'election 2008'). Traffic tags are also computer-generated: unique identification numbers are assigned to a *YouTube* video, as well as to users on *Facebook*. Traffic tags allow for the identification of objects across the Web, most notably through search engines, but also through application programming interfaces (APIs), which, as we have already noted, govern how objects circulate within and sometimes across most web platforms. For instance, when a user clicks on the 'Share on *Facebook*' button after watching a video on YouTube, the ID number of the video will reappear in the *Facebook* source code of the user's page. The current challenge thus lies in identifying and following traffic tags associated with Web objects so as to see how information circulates within and across Web 2.0 platforms. This process of tracking the migration of object or actor-specific-code will provide us with clues as to how cultural processes that are traditionally only visible at the level of the user-interface are governed by the largely commercial imperatives of APIs (particularly on the larger and more popular platforms like Facebook and Youtube).

5. The taxonomy of traffic tags

While meta-tags offer an important contrasting view to the use of hyperlinks as indicators of political associations and networks, their use has been vastly complicated and expanded in the web 2.0 universe. In fact, as we have argued elsewhere [24], such forms of user-generated content serve as a key component in the production of web 2.0 sites – since they are almost entirely rely upon user-generated content to function and thrive. However, the task of developing methods for tracking individual users and networked political objects across platforms is a complex one, in large part because each platform has its own set of protocols that disrupt the more free flowing aspects of web 1.0 (or html based forms of publishing and networking). In the remainder of this paper we identify new forms of code and software functions that might allow one to track objects and users across web 2.0 sites. Such software artifacts serve as possible sites of 2.0 research, though, as we detail below, this de-centered method of analysis, which begins with objects and users, as opposed to networks, communities or other digital collectivities, will inevitably raise questions about one's choice of a starting point – that is, the rationale for what objects one begins to track, and what sequence and series of information aggregators one deploys to view the dissemination of said "traffic tags". Lastly, before we move on to discuss such new sites of research, we should reiterate that "traffic tags" typically come in two forms – both of which are required to track objects and map routes of networked content, and relationships between users, content, and other users – namely code that individually identifies specific users/objects and code that facilitates the circulation of shared objects. In many respects this method is not entirely new, as it also duplicates, albeit with some differences of course, the techniques and technologies that are deployed to diagnose the circulation of commodities, consumers, and services in today's economy [11]. In lieu of traffic tags discussed below then, such networked objects, users, and routes, have employed well know technologies such as barcodes, RFID tags, and more broadly "just-in-time-delivery systems", for many decades now.

While inevitably incomplete, we have identified a number of traffic tags that exemplify our search for code that can be employed in a object centered method of web 2.0 analysis:

- plain language (text)
- individual user IDs
- APIs

- tags that accompany user-generated objects (self generated, auto-generated)
- hyperlinks
- spam-strings
- RSS feeds
- object title
- file formats
- usernames
- formal names
- IP addresses
- copyright code (eg. creative commons)
- email addresses.

This list is of course not exhaustive, but is rather meant to offer a starting point for discussion. That said, we would argue that plain language or text is one the most overlooked forms of traffic tags on the web. As we argue elsewhere, with respect to the re-use and circulation of Wikipedia entries [25], one can take formal language and deploy it in a series of net information aggregators (search engines for example) to identify the dissemination of similar or exact duplicates of sentences, and paragraphs. Plain language is a particularly cogent form of traffic tag as they double of course as both semiotic and deitic signs [12,33], meaning that they provide researchers with the rhetorics of networked politics, as well as to how terms are used to, literally in the case of hyperlinked words, take users to other documents and web platforms.

APIs, or application program interfaces are similarly pivotal in our proposed research perspective since, as we noted earlier, they sit "in-between" interfaces and back-end code, often providing more savvy users with an ability to data-mine specific platforms for information on users and objects. So perhaps to qualify a bit here, APIs serve as search engines of sorts, as they link together users with objects and particulars spaces on platforms like Facebook (eg. on groups, or "causes" pages, etc.). One can "query" an API for example, for various data associated with a particular user[14] or group of users. That said, APIs can also be used to better understand how networked political objects move across, are slightly modified, or become the domain of specific 2.0 platforms – to the degree that their sharing becomes more difficult.

Really Simple Syndication or RSS feeds similarly offer researchers a universally recognizable code embedded on many political websites, blogs, and media sites, that serve in many respects as a content portal, a mega hyperlink in 1.0 language, to the extent that it creates a gateway from which almost all content and indeed some meta-tags and information for specific website entries, stories, or posts (date stamps, bylines, etc.) can be collected and used for comparative cross platform analysis. Much like API's, in other words, RSS feeds serve to *traffic* meta-tagged content. Our own analyses of political blogs in Canada used the RSS feeds from partisan blogs to performs various forms of content analysis across the Canadian political blogosphere [13]. A slightly modified version of these traffic-focused tags and code includes the creative commons logos and tags, signs and code that govern, classify, and enable access to various forms of multimedia on the web (Flickr images for example). Content, actors, and platforms associated with creative commons licenses speak directly to the rules concerning the ability to publicly use, reuse, remix, and broadly share digital objects. Searching for creative commons code across platforms using search engines, APIs, and RSS feeds thus provide helpful sets of data that provide insight

[14] See, for instance, the API test console on Facebook: http://developers.facebook.com/tools.php.

onto the various forms of digital ownership and subsequently trafficking of content that takes place across the internet. Such issues are of increasing importance for the political sphere as various jurisdictions around the world move to more open source models of information management and access[15].

Identifying and tracking the contributions of political actors (partisan bloggers, vloggers, political staff, journalist-bloggers, etc.) is perhaps one of the easiest components of our suggested method of inquiry. In large part because almost all web 2.0 sites require some form of user registration, individual IDs are common place. These IDs are of course then platform specific, which can help when trying to determine the success of failure of various cross-platform political campaigns. User accounts almost always require registrants to register a unique username, thus making it relatively easy to track all of the content and objects uploaded, remixed, commented upon, etc. by specific users. We might also extend this logic to less formalized definitions of usernames, for example "AXXO" a well-known user of peer-to-peer software known to upload and circulate DVD ripped material on bittorrent networks[16]. Email addresses, likewise, offer opportunities to identify the circulation and contribution of individuals across platforms and time, though with important caveats that speak to the limits of political networking – both as a practice and site of research. Emails listed on *Facebook* pages for instance are not retrievable through interface searches or through the platform's API, thus making it harder to analyze and also circulate calls to action posted on *Facebook* that often end in an organizer's email address. IP addresses are similarly one of the more reliable and unique forms of identifying specific users, or in the case of "whois" searches, the unique address where a computer is registered. Journalists often turn to such "whois" searches during election campaigns to determine the owners of specific attack or parody websites – a daunting task as according to one estimate over 2,357 sites were registered for candidate Obama[17]. In terms of identifying specific internet users or actors, formal names of course, while less specific, can also be used in conjunction with other IDs to track the contributions of specific users or 2.0 platform accounts, an important caveat again as often multiple techniques of identifying actors are required when searching for networked campaigns and content across 2.0 platforms.

The last set of traffic tags discussed herein speak more to the qualification and characterization of digital objects, a means by which posts, images, and videos are "tagged" typically using keywords, hash tags, and other content related indices. Such user-generated forms of classification of course serve a central role in various projects seeking to monitor trends on social media platforms like twitter or in the blogosphere, for instance as aggregated by the platform specific search engine Technorati. Such tags serve particularly those in the fields of information science, information retrieval and library science, to complicate objective means of classifying, controlling and circulating documents and media objects. The emergence of the folksonomy epistemology, conversely, can also be overly celebrated as the ultimate freeing of information, wherein citizens not only produce and circulate their own political campaign objects, but also play a pivotal role in classifying their contributions to a networked political landscape.

6. Conclusions

While we recognize that this paper has only begun to enumerate a new 2.0 inspired approach to the study of online networks and political networking, there are clear examples in the political sphere that

[15] The decision by the Obama whitehouse to switch to an open source "Drupal" website management suite was widely lauded by information activists. < http://drupal.org/node/375843>.

[16] Cf. < http://www.mininova.org/user/aXXo> for an online list of files uploaded by this "username". Wikipedia also provides an interesting overview of this "internet alias" <http://en.wikipedia.org/wiki/AXXo>.

[17] <http://inside.123-reg.co.uk/archives/domain-names-the-web-and-the-us-election>.

suggest we are on the right track. Journalists now routinely seek to track the original of digital objects that seek to anonymously attack or parody public figures and politicians[18]. Such forms of political research is also practiced by party staff. One of Canada's most social media savvy reporters, for example, recently noted that political staff in the Canadian capital matched the exact software code for a shade of blue used by the governing party on its political/party website (Pantone #333399!) to a government website in an effort to argue that the current administration was politicizing – through similar branding – various government programs[19].

Our paper has similarly attempted to provide examples of code that can be analyzed to track political campaigns and communication across web 2.0 platforms. However, much remains to be done. First, a road map of sorts is required to understand how – and under what conditions – an actor (political party, blogger, or other user) can best take advantage of the routes in, through, and across social media sites. Certain opportunities to network content between two platforms are routinely prohibited. Youtube videos can be embedded on blogs, but up until recently not on Facebook or Twitter. Blog posts can be linked to Facebook friend feeds, but not Youtube, etc. Such distinctions are important to recognize when studying the effectiveness of online political campaigns, yet the speed at which such networked platforms emerge, and then later change their back-end code and APIs however makes such network mapping always and already out-of-date.

Studying political networking across the web 2.0 thus requires a commitment to experimenting with numerous traffic tags in the process of trying to track the uploading, spread, reuse or remixing of various digital objects. Some sites provide for easy data collection with RSS feeds (such as blogs or information aggregators like Google), while others like Facebook and Youtube require an engagement with their API to collect large data sets. And again just when one thinks that a sound method has been achieved to collect and track youtube videos to blogs or Facebook their API is changed (as was the case in 2009), forcing researchers to readjust their methods again.

While the broader task of tracking inter-relationships between platforms is fraught with pitfalls, the concept of traffic tags is still a fundamentally sound one if one wants to understand the relationship between objects, users (actors), and social media platforms. Shared 2.0 objects, like internet packets, need unique identifiers to distinguish themselves from each other, in addition to providing the glue which binds together not only users (as "friends" on Facebook, for example), but also between and among social media or 2.0 platforms. Without uploading, sharing, commenting, and remixing there would be no networked media to map or take advantage of. Blog posts, comments, videos, and photos serve as molecular objects, always moving among the larger networked apparatus.

What is needed then is a road map that can point to how one can not only identify users and objects, but also how these can be tracked across any two social media platforms – a process, that requires constant updating, to include to new platforms, new functions, and new APIs. Such maps of traffic tags would consequently move research on political network away from implied definitions of political connections or associations online, through an overreliance on hyperlink mapping research, to a much richer understanding of what practices and sets of objects, users/actors, and 2.0 sites make for an effective (or botched!) networked campaign.

[18] Our own work with the Canadian Broadcasting Corporation focused heavily on determining the source of social media barbs and dirty tricks during the 2008 federal election in Canada. <http://www.cbc.ca/news/canadavotes/campaign2/ormiston/> accessed January 25, 2010.

[19] http://davidakin.blogware.com/blog/_archives/2009/10/27/4363377.html <accessed January 25, 2010.

Acknowledgements

This article was made possible through support from the Social Science and Research Council of Canada, the National Center for e-Social Science, University of Manchester and Govcom.org, University of Amsterdam. The authors would also like to acknowledge the input and feedback from other members of the Infoscape research lab, particularly Zach Devereaux, Fenwick McKelvey, Isabel Pedersen, and Kenneth C. Werbin.

References

[1] J. Abbate, *Inventing the Internet*, Cambridge: MIT Press, 1999.

[2] Y. Bakos, The Emerging Role of Electronic Marketplaces on the Internet, *Communications of the ACM* **41**(8) (1998), 35–42.

[3] J.D.. Bolter and D. Gromala, *Windows and Mirrors: Interaction Design, Digital Art, and the Myth of Transparency*, Cambridge: MIT Press, 2005.

[4] Brin and Page, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report. Stanford Infolab, 1998.

[5] C. Bourret, Callon & Rabeharisoa, ANT 2.0? Ou: Les methods d'analyse des reseaux peuvent-elles nous delivrer de la tyrannie des reseaux? Unpublished ms, 2007.

[6] J. Burgess, and G. Joshua, *Youtube: Online Video and Participatory Culture*, London: Polity Press, 2009.

[7] A. Chadwick and P.N. Howard, *Handbook of Internet Politics*, London: Routledge, 2008.

[8] B. Easter, Search Showdown: Barack Obama vs. John McCain, *Promotionworld.com*, 26 September 2008. <http://www.promotionworld.com/se/articles/article/080926SearchShowdownBarackObamavsJohnMcCain.html>.

[9] G. Elmer, U.S. Energy Policy: Mapping the Cyber-Stakeholders, *The Communication Review* **9**(4) (2006), 297–320.

[10] G. Elmer, Re-tooling the Network: Parsing the Links, Codes, and Commands of the Web World, *Convergence: The International Journal of Research in New Media Technologies* **12**(1) (2006).

[11] G. Elmer, *Profiling Machines: Mapping the Personal Information Economy*, Cambridge: MIT Press, 2004.

[12] G. Elmer, Spaces of Surveillance: Indexicality and Solicitation on the Internet, *Critical Studies in Mass Communication* **9**(1) (1997), 182–191.

[13] G. Elmer, G. Langlois, Z. Devereaux and F. McKelvey, Blogs I read: Partisan Recommendations in the Blogosphere, *Journal of Information Technology and Politics* **6**(2) (2009), 156–165.

[14] G. Eysenbach, From intermediation to disintermediation and apomediation: new models for consumers to access and assess the credibility of health information in the age of Web 2.0, *Studies in Health Technology and Informatics* **129**(Pt 1) (2007), 162–166.

[15] K. Foot and S. Steven, *Web Campaigning*, Cambridge: MIT Press, 2006.

[16] A. Freidberg, *The Virtual Window: From Alberti to Microsoft*, Cambridge: MIT Press, 2006.

[17] M. Fuller, Behind the Blip: Essays on the Culture of Software, New York: Autonomedia, 2003.

[18] A. Galloway, *Protocol: Or how Control Exists After Decentralization*, Cambridge: MIT Press, 2004.

[19] M. Garrido and A. Halavais, Applying social-network analysis to study contemporary social movements, in: *Cyberactivism: Online Activism in Theory and Practice*, M. Martha and M.D. Ayers, eds, Cambridge: Routledge, 2003.

[20] J. Green and J. Burgess, *Youtube: Online Video and Participatory Culture*, Cambridge: Polity Press, 2009.

[21] A. Halavais, *Search Engine Society*, Cambridge: Polity Press, 2008.

[22] M. Hindman, *The Myth of Digital Democracy*, Princeton: Princeton University Press, 2009.

[23] N. Jankowski and S. Martine, Internet-based political communication research: Illustrations, challenges & innovations, *Javnost – The Public* **15**(2) (2008).

[24] G. Langlois, F. McKelvey, G. Elmer and K. Werbin, Mapping Commercial Web 2.0 Worlds: Towards a New Critical Ontogenesis, *Fibreculture* **14** (2009).

[25] G. Langlois and G. Elmer, Wikipedia leeches? The promotion of traffic through a collaborative web format, *New Media & Society* **11**(5) (2009). http://nms.sagepub.com/cgi/content/abstract/11/5/773.

[26] C. Lury, *Brands: the Logos of the Global Economy*, London: Routledge, 2004.

[27] A. Mackenzie, *Cutting Code: Software and Sociality*, Peter Lang: New York, 2006.

[28] T.J. McCarthy, Trademarks, Cybersquatters and Domain Names, *DePaul-LCA Journal of Art and Entertainment Law* **10** (1999).

[29] H.W. Park and T. Mike, Hyperlink Analyses of the World Wide Web: *A Review, Journal of Computer Mediated Communication* **8**(4) (2003). <http://jcmc.indiana.edu/vol8/issue4/park.html>.

[30] R. Rogers, *Information Politics on the Web*, Cambridge: MIT Press, 2006.

[31] R. Rogers, Towards a Live Social Science on the Web, *EASST Review* **21**(3/4) (2002), 8–11.

[32] R. Rogers and N. Marres, Landscaping climate change: A mapping technique for understanding science & technology debates on the World Wide Web, *Public Understanding of Science* **9**(2) (2000), 141–163.

[33] R. Shields, Hypertext Links: The Ethic of the Index and Its Space-Time Effects, in: *The World Wide Web and Contemporary Cultural Theory*, A. Hermann and T. Swiss, eds, London: Routledge, 2000, pp. 145–160.

[34] M. Warschauer and G. Douglas, Audience, Authorship, and Artifact: The Emergent Semiotics of Web 2.0, *Annual Review of Applied Linguistics* (Issue 27) (2008), 1–23.

Journal of Technology in Human Services
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/wths20

Talking of Many Things: Using Topical Networks to Study Discussions in Social Media

Tim Highfield [a]

[a] Queensland University of Technology/Curtin University, Australia

Published online: 06 Dec 2012.

To cite this article: Tim Highfield (2012): Talking of Many Things: Using Topical Networks to Study Discussions in Social Media, Journal of Technology in Human Services, 30:3-4, 204-218

To link to this article: http://dx.doi.org/10.1080/15228835.2012.746894

Journal of Technology in Human Services, 30:204–218, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1522-8835 print/1522-8991 online
DOI: 10.1080/15228835.2012.746894

Talking of Many Things: Using Topical Networks to Study Discussions in Social Media

TIM HIGHFIELD
Queensland University of Technology/Curtin University, Australia

This article outlines a method for studying online activity using both qualitative and quantitative methods: topical network analysis. A topical network refers to "the collection of sites commenting on a particular event or issue, and the links between them" (Highfield, Kirchhoff, & Nicolai, 2011, p. 341). The approach is a complement for the analysis of large data sets enabling the examination and comparison of different discussions as a means of improving our understanding of the uses of social media and other forms of online communication. Developed for an analysis of political blogging, the method also has wider applications for other social media websites such as Twitter.

KEYWORDS *blogs, hyperlinks, issue publics, public debate, social media, topical networks*

INTRODUCTION

Social media facilitate the development of conversations online around particular events or topics of interest, where participation is not necessarily limited by geographical or social factors. A message posted publicly on Twitter, for example, is potentially visible to all users of the site, and indeed

Received June 5, 2012; revised November 1, 2012; accepted November 2, 2012.

The initial research informing this article comes from my doctoral research; I thank Axel Bruns and Jason Sternberg for their supervision of this project, and Lars Kirchhoff and Thomas Nicolai for collecting the blog data discussed here. I also thank the anonymous reviewers for their comments and suggestions for this article.

Address correspondence to Tim Highfield, ARC Centre for Creative Industries and Innovation, Queensland University of Technology, Z1–515, Musk Avenue, Kelvin Grove, QLD 4059, Australia. E-mail: t.highfield@qut.edu.au

to people without Twitter accounts themselves. Such web-based communication platforms offer ways for opinions, messages, and content to be shared and repurposed quickly and easily. While we might refer to Twitter, Facebook, or the blogosphere as singular entities to identify where discussions are taking place, though, the individuals using these platforms are not the same (and, indeed, the likes of Facebook and Twitter are not used in isolation); their motivations for using these sites vary, and so does their respective interest in a given subject of conversation. The group of bloggers responding to a specific political issue, for example, may be different from the group discussing the previous weekend's sports results. Different discussions will also take varying forms, although occurring within the same space. For instance, responding to crises, publishing live commentary on televised events, or taking part in a conference backchannel may involve different users, interactions, and types of information, but they coexist within the overall activity hosted on sites such as Twitter.

To study how discussion takes place within, and across, social media platforms, researchers might establish long-term projects, tracking a group of users over time. This approach provides important cumulative data for identifying patterns of use—such as how many posts were published by bloggers over time, or which bloggers posted most often. These overall, baseline data are useful for examining *what* the research has found—the overall posting patterns, and the most and least active users, for example. However, it does not easily explain the patterns discovered. As boyd and Crawford (2012) note in their discussion of studies involving "Big Data," the analysis of large data sets from online sources, such as Twitter archives, can lead researchers to find "patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions" (p. 668).

To provide additional insight into online activity, this article promotes the study of *topical networks*: "the collection of sites commenting on a particular event or issue, and the links between them" (Highfield, Kirchhoff, & Nicolai, 2011, p. 341). Using such units of analysis within large data sets is not intended to replace "Big Data"-type studies, but to supplement them by examining the tracked activity in greater detail. Identifying topical networks using these large data sets enables researchers to determine *why* and *when* connections were made, and the context for the discussion of particular topics. This method may also be employed within smaller data sets too, of course; in response to critiques of the quantitative focus of "Big Data," though, this approach can also provide some qualitative exploration of sections of large data sets.

Topical network analysis follows Rogers's (2009) promotion of investigating the "online groundedness" of online activity, where research follows a particular online medium, to track "its dynamics, and makes grounded claims about cultural and societal change" (p. 8). The specific methods used for analyzing data will vary from project to project, depending on the tools used. Due to these differences, this article does not aim to set out a step-by-step process. Instead, it argues at the conceptual level for a mixed-methods approach to gain further value from large, rich data sets. The following sections provide an initial overview of topical networks and the methods for their identification and analysis, and the advantages and limitations of this approach. An example from the Australian political blogosphere is used to illustrate this process. I also outline connections between topical networks and concepts developed around both public communication and online activity, such as issue publics and web spheres, which provide theoretical grounding for this analysis. Finally, I note further directions and applications of this approach.

TOPICAL NETWORKS

Topical networks were initially identified within a long-term research project comparing political blogging in Australia and France (Highfield, 2011), as a way of locating and comparing specific discussions within these blogospheres. The definition cited earlier applied to bloggers' coverage of particular themes and their linking to other blogs and web sources. However, topical networks are not restricted to the blogosphere alone. Rather than referring to "sites," the definition can be expanded to encompass multiple social media platforms, or to concentrate on activity on a single website, such as Twitter. In this latter case, the topical network could feature the different users commenting on a particular issue, such as through a central hashtag. The websites involved will vary between studies, adapting the method in the process as topical networks are examined within a range of contexts, including politics, economics, popular culture, health, and education. Regardless of the research focus, though, the resulting topical network will be oriented around a specific thematic discussion, often within a longer term study of a wider population of sites or users.

In this article, I draw on research using web-based, publicly accessible data, captured from blog posts. However, the same analytical approaches may be used on other data sets. From an education perspective, for example, a collection of *Blackboard* bulletin board posts may be categorized by the subjects covered in the text, providing the initial basis for topical networks within the data set. While this article examines explicit network data through hyperlinks, implied connections may also be used to demonstrate the links between users. Such implied links might appear through replies to other discussion board posts, which might not have a hyperlink to signify the connection. Even without "networked" data, the "topical" approach may be used to examine different types of online communication.

The comparative topical network approach discussed here was developed in response to studies of large, long-term data sets, as a means of examining specific thematic discussions within the wider data collected. The analysis of several months or years worth of data provides valuable information about patterns of use for different websites, such as the extended coverage of Arabic and Persian blogging by, respectively, Etling, Kelly, Faris, and Palfrey (2010) and Kelly and Etling (2008). However, the wider analysis alone does not explain the behaviors tracked, for example, what topics were discussed during a spike or lull in activity or the context for links to external websites.

To answer these questions, topical network analysis takes a multiprocess, mixed-methods approach. First, relevant keywords are used to identify and isolate from the wider data set the data pertaining to a chosen topic, such as blog posts, discussion board contributions, or tweets commenting on specific public figures, organizations, or events. The selected data then form the basis of the topical network. Following the identification of relevant content, a series of quantitative and qualitative processes may be used in combination to examine the discussions and activity represented within the topical network. For example, quantitative methods are used, as with the overall data sets, to determine patterns of activity. Such patterns include the number of contributions per week, day, or hour, the total contributions per user, blogger, or website, and any noticeable spikes or troughs in the discussions.

Depending on the type of data represented within the topical network (tweets, blog posts, and so on), different processes can be used to further analyze the coverage of the chosen topic. Hyperlink network mapping, for example, draws on the networked aspect of the data in question, through explicit hyperlinks to other online sources. Visualizing these connections as network maps can then demonstrate which sources are common references for the participants contributing to the topical network. The visualization process can also help to identify any clusters of users and sources within the overall network, where users in a smaller group link to each other or a distinct collection of websites that are not cited, at least not as frequently, by the rest of the network. However, it should also be noted that while network maps provide important visual cues around the connections between users, and help to make sense of the links present within large data sets, visualization by itself does not provide an explanation as to *why* these connections are made. Similarly, while hyperlinks are often used as indicators of connections between different websites, not all links are the same (see Adamic, 2008; Halavais, 2008).

Topical network analysis also makes use of approaches such as textual analysis to determine the context for the studied discussions. These methods allow studies to take into account different aspects of social media that might not be possible with large-scale, automated data processing, such as differentiating between link type—such as links in blog posts, blogrolls, or comments on posts—and to examine what these links can tell us about online communication. This approach can also negate the question surrounding the longevity of connections between participants in the network. Links featured within blog posts are not necessarily permanent indicators of affiliation or endorsement. A blogger may cite another's work once, in reference to a specific subject, but then never again in his or her later posts. The link from one blogger to another would still appear within the overall data set collected, yet analyzing the total patterns does not provide any context for this connection. Bruns (2012) raises a similar conceptual question around the life span of content and links posted on Twitter, asking how long the connections between users linked by @replies last. The answer to this temporal dilemma is beyond the scope of this article; as is the case for other aspects of these studies, though, the wider context for these links will be important (e.g., the rate of posting per user and overall, the time period covered by the discussion, and any repetition of the links).

Further processes involve analyzing the text of relevant posts individually to provide a qualitative view of the topical network data. Rather than treating the network as a like-minded whole, covering the chosen topic to an equal degree, the textual analysis demonstrates the different responses to the topic, and the context for these comments. While each blog post or tweet contains a relevant keyword for the topical network, the surrounding text might have a different subject as its focus, or the topic in question might be framed around an alternative context. These distinct ways of commenting on a given topic are not as easily identified within quantitative analysis alone, highlighting the value of examining at a qualitative level the activity captured within large data sets to understand the topical networks.

The different discussions tracked by topical networks might also take varying forms depending on the type of event or issue covered. The live-blogging or -tweeting of a sporting event or televised debate may lead to a topical network that is completely dissimilar to that formed in response to a crisis or scandal, with different patterns of posting, linking, and sharing information—even when drawn from the same overall data set. Similarly, the coverage of the same issue on different websites may also vary, depending on such factors as the number of people contributing to discussions and their personal or professional interest in the topic at hand. For example, tracking health-related issues on a specialist forum or discussion board, where health is the main subject to be covered, may depict a debate that is dissimilar to that captured from more general hashtag or keyword archives on Twitter.

Elmer (2006) notes that using a combination of qualitative and quantitative methods allows researchers to analyze in greater detail the dynamics of different discussions (p. 15). This idea guides topical network analysis and its mixed-methods approach to studying large data sets. This method enables the researcher to compare numerous discussions taking place at different points in the data, providing a means for contextualizing overall patterns and accounting for possible variations within "Big Data" projects. Most importantly, topical networks enable researchers to examine how the coverage of a given issue plays out within the wider data, such as how the discussion of a particular person or subject develops over time.

CONCEPTUAL REVIEW

Topical networks have their theoretical roots in several concepts concerning the shape of public debate. Some of these are directly applicable to online media, while others were developed independently and subsequently adapted to this context. Topical networks depict discussions taking place around specific issues. The connections here are not necessarily permanent associations, and the discussions may develop and decline quickly. This idea links to the notion of multiple, temporary issue publics appearing within a more constant, wider scope public sphere (or public spheres). These assemblages may overlap, with people contributing to more than one debate, but each issue public is centered on particular topics or themes (see Dahlgren, 2009). The shape of an issue public will change over time, and the context for each group means these publics will also take different forms in comparison with each other. Different contributors will comment on a range of topics, with no requirement to contribute to all or any debates.

The type and frequency of comments by each person involved in the topical network will also vary based on a number of professional and personal factors. For example, Jang and Park (2012) note the presence of "issue specialists" within discussions, based on the subject in question being an issue of personal relevance. In addition, individuals with a professional background in aspects of the topics covered by the network may be among the most active contributors to the discussion. Within the Australian political bloggers, for instance, different groups of specialists were identified within the wider blogosphere, who would contribute to political debate by adding new interpretations of the issues at hand based on their own economics or polling data analyses (Highfield, 2011).

The discussions featured in this article do not necessarily focus on one particular interpretation of an issue. As Marres (2006) notes, the presence of a group of people in conversation does not mean that participants agree with each other. A great number of voices contribute to public debate overall, with smaller, topical debates taking place among a subset of these participants, each of whom has a varying level of engagement with the topics in question.

While issue publics may develop away from computer-mediated communication, there are Internet-specific concepts that also help develop the ideas behind topical networks. These include web spheres (Schneider & Foot, 2005) and issue networks (Marres, 2006), both of which are formed around issue- or event-driven debates, and include both the individuals contributing to the debate and the resources used within these discussions. Such groups can be platform specific. Bruns and Burgess (2011) suggest that the use of hashtags within tweets "facilitates the *ad hoc* emergence of issue publics made up of interested *Twitter* users around these topics" (p. 38).

Topical network analysis provides researchers with the capability to compare patterns and citations across different events and over time. Each discussion sees the creation of a temporary issue public within the larger group represented in the larger data set, but there is no way of predicting which users will comment on which subject. In a data set containing a known number of contributors, such as tracking the output of several Twitter accounts, topics might be covered by any, all, or none of the individuals concerned. Topical networks then become potentially ideal cases for the study of issue publics. Debates published online are traceable through collections of blog posts, status updates, retweets, and links. Previous studies have examined topical discussions online, focusing on specific cases rather than debates within a wider data set: for example, Bruns's (2007) research into mentions by bloggers of an Australian detainee at Guantanamo Bay, or the analysis of how bloggers responded to Hurricane Katrina by Macias, Hilyard, and Freimuth (2009). Similarly, topical conversations on Twitter have been studied based on individual hashtags or keywords (for example, #ausvotes: Bruns & Burgess, 2011; #wikileaks: Lindgren & Lundström, 2011). Not only are these debates easily searchable and automatically connected through the creation of links for each hashtag, but they can also connect separate discussions around a shared theme—the use of any hashtag is not dependent on following other accounts also posting on this topic. Finally, several studies also track Twitter activity based not on keywords but on a list of user accounts, such as politicians and journalists (Maireder, Ausserhofer, & Kittenberger, 2012); from the collected tweets of these users, keywords can again be used to identify topical discussions within the wider activity captured.

IDENTIFYING AND ANALYZING TOPICAL NETWORKS: CASE STUDY

Topical networks may then be identified within the wider activity on, and across, numerous websites. These networks may be located and analysed from a larger data set of activity, or captured individually as part of an ongoing comparison of online discussions. This process allows researchers to show which subjects attract the widest or most specialist interest among groups of users. Such groups might be genre specific, such as the collection of political blogs studied here, or they might track activity within a local or national user base. For example, Bruns, Burgess, Kirchhoff, and Nicolai

(2012) have mapped how hashtagged discussions representing local and international news stories, sports events, and television programs were distributed across a network of 120,000 Australian Twitter users.

Because of the range of data formats, tools, and methods that might feature within different projects, this article does not seek to list specific, step-by-step processes for the identification and analysis of topical networks. However, this section provides a brief overview of an example from the Australian political blogosphere (Highfield, 2011) to illustrate an approach to topical network analysis. Although aspects of the methods used here might not be appropriate for all studies attempting to track discussions within online communication, the framework guiding the analysis may be applicable to a variety of cases.

The context for the following topical network was a wider research project capturing the published outputs of a sample of Australian and French political bloggers between January and August 2009. During this period, 10,529 posts were archived from 61 Australian political blogs. From each post, data were extracted, such as the date and time posted, links within posts, and the text of each post, ahead of further analysis. The two data sets were then analyzed separately to determine the overall activity represented by the collected posts. This process included identifying the most active sites, most popular sources based on links received, and any peaks or troughs in daily posting activity.

However, these overall patterns cannot show the reactions of bloggers to specific topics. Analyzing just the total posting and linking activity treats these almost as permanent blogging behaviors, where bloggers are active and sources linked to at a constant rate. Within the captured posts, though, myriad topics are discussed, provoking different responses from the bloggers in the sample. Not all bloggers will discuss the same topics, and their own commitments may mean that a blogger does not post for several weeks or months.

These variations can be examined, though, by moving from the wider study of the overall population to the more focused topical networks. This article provides a brief discussion of one such network, formed around the Australian "Utegate" political scandal between June and August 2009. This scandal centered on allegations against the then-prime minister and treasurer of preferential treatment for a Queensland car dealer seeking government assistance in response to the global financial crisis. This case is investigated in further detail, alongside additional political topical networks, by Highfield (2011); for this article, it serves to illustrate the concepts and framework behind topical network analysis.

To locate the topical network, the wider data set was filtered to isolate relevant blog posts. In this case, the data set was filtered at the keyword level (as opposed to limiting the data by a range of dates), in order to track the growth and decline of interest in a topic that had a clear starting point within the collected data. The Utegate topical network was created by searching for posts containing key terms (Utegate, Ozcar) and names specific to the scandal (Godwin Grech). The resulting network drew on data from 52 posts from 17 blogs, published over 8 weeks between June and August 2009.

The filtered data form the basis for the topical network analysis. First, the network was compared to the wider activity during the same period, to evaluate the level of interest in the subject among the bloggers in question. For Utegate, its peak activity on June 21 accounted for more than 10% of the posts published that day. However, within the 2-month period overall, Utegate featured in less than 2% of the total blog posts captured. This suggests that the scandal was not a prominent topic for Australian political bloggers, even though it was a leading story in mainstream media publications at points during the same period.

The topical network analysis then uses different processes to further examine how and why bloggers were discussing the events and issues at hand. Hyperlinks included in each post were extracted to identify which sources were cited during these discussions. Network visualizations aided the process by highlighting the prominence of different sources and bloggers within each topical network. These visualizations were created by representing each link as a directed connection between two sites—*from* the blogger in question *to* the external website. In their coverage of the Utegate scandal, the Australian bloggers contributing to the topical network linked to domestic news sites—as expected, given the local focus of the scandal—and in particular to the websites of News Limited publications.

However, the hyperlink analysis itself is still initially a quantitative process. Here, the context for the links is absent—citing a news article or another blogger is not necessarily endorsement of the views presented, for instance. Textual analysis of the topical network blog posts was then carried out using the Leximancer software to discover the actual subjects featured by the bloggers in the sample. This automated process was supplemented by manually analyzing the posts to evaluate the intentions behind bloggers' choices of links. The Utegate analysis highlights the importance of qualitative methods to topical network analysis. Although News Limited websites were linked to by several Australian bloggers discussing the scandal, these references were not necessarily positive. Instead, bloggers mentioning Utegate commented less on the scandal itself, and more on its disrupting impact on other political issues, or on the way that it was being covered by the mainstream media. In particular, the reporting of Utegate by News Limited publications was criticized by several bloggers, for its content and stance, and also for focusing attention on what the bloggers considered a nonissue. This disapproval was accompanied by links to specific articles that were promoting Utegate instead of political issues that bloggers saw as more worthy or deserving of media attention.

The hyperlink and textual analysis of the topical networks also confirmed patterns from the wider data set; the overall linking patterns between the blogs in the sample suggested that several thematic groups were present within the Australian political blogosphere. These groups were centered on shared topics, including economics and psephology (the study of voting and polling data). While representatives of these groups discussed Utegate, their posts remained within the context of their specialist subjects: for example, analyzing the opinion polls released after the scandal broke, or mentioning Utegate as a contributing factor for rising or falling approval ratings.

These findings further demonstrate the various perspectives and interpretations involved within a single discussion. To illustrate these topical variations within the network itself, composite network visualizations were created. This process drew on both the hyperlink and textual analysis to depict the distribution of different themes through the topical network; Figure 1 shows an example composite visualization, showing the different themes featured, and sources cited, by Australian bloggers commenting on the Utegate scandal.

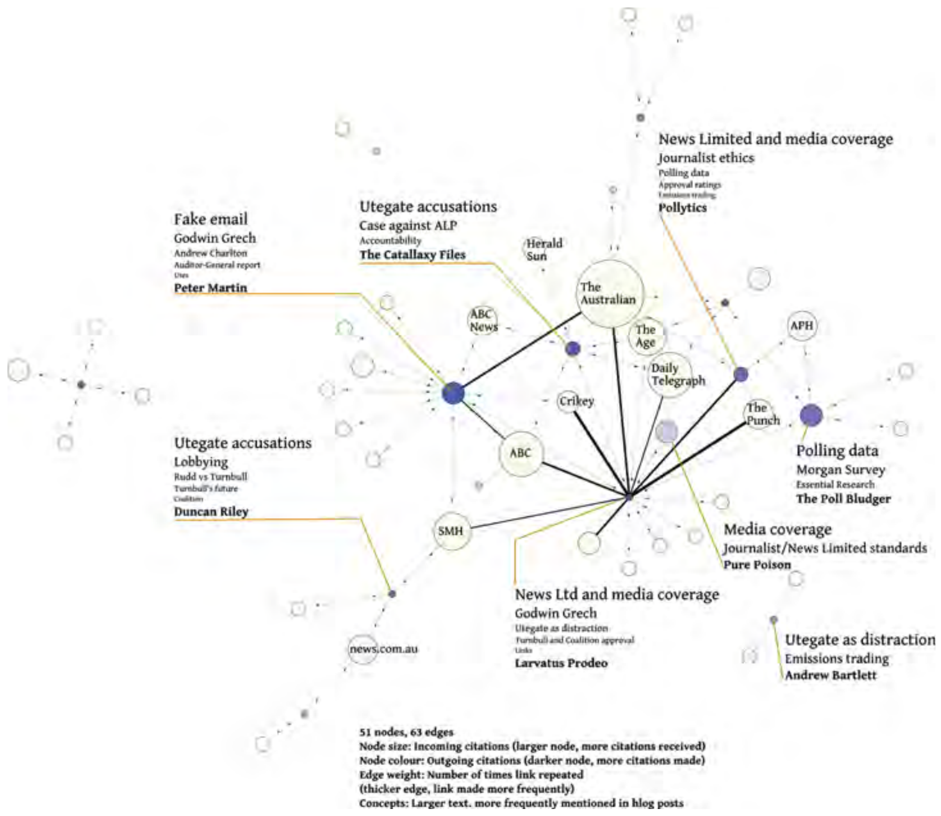


FIGURE 1 Composite Utegate topical network visualization, showing key sources and topics featured by selected bloggers. (Figure available in color online.)

WHY TOPICAL NETWORKS?

As the Utegate case study shows, an advantage of the topical network approach is found in the snapshots extracted from the wider data set. By focusing on temporary groups within a larger population of users or sites, topical networks provide the opportunity to evaluate how public debate takes place online, such as within the blogosphere or social media. For example, different discussions within the political blogosphere might be compared to examine whether bloggers link only to individuals sharing the same political affiliation or ideology, as demonstrated either explicitly on their sites or implicitly through their coverage of issues. Collections of tweets on matters of public interest may be studied to evaluate whether social media users follow the mainstream media in their coverage of issues, or whether they promote alternative interpretations of these themes. However, although topical network analysis method was developed for studying political communication online, the approach has applications beyond this context, as there are many different discussions and uses of social media taking place simultaneously.

The identification and study of topical networks within larger data sets allows for a more nuanced examination of online activity. Patterns and statistics derived from the total data collected provide important contextual information, and serve to introduce the subject of the study—the users or sites tracked. Although the resulting overview of the collected data shows the total activity, though, it does not provide information about the dynamics of discussions within different contexts. The study of topical networks then provides a crucial counterpoint to the analysis of the whole period, composite data set. Instead of viewing the baseline data as the definitive picture of the groups studied, topical networks question the users featured and the connections made between them and other sites. By isolating topical networks within large data sets, the researcher can examine whether the wider patterns are consistent for all contexts, or whether different sites become prominently linked in response to particular themes.

Topical network analysis is still an exploratory method, though, and is not without its limitations. The keyword-oriented method of identifying topical networks does not necessarily locate all relevant material. For research into Twitter activity, for example, more extensive topical networks might be identified around a mixture of keywords and hashtags; in cases where multiple hashtags are used, such as when a central tag has not yet been agreed upon, searching for particular keywords can supplement the filtered data. Posting about a specific subject on Twitter does not also require the relevant hashtags to be included in tweets. Australian political discussion on Twitter often includes the #auspol hashtag, for example, but also encompasses tweets not containing this marker. Similarly, blog posts might include common labels or categories for their posts to note the primary topics featured, but again there is no requirement for this.

As with any study of online activity, especially around Twitter and blogs, it is important to acknowledge the representative limits of the data sets used. While the collected tweets analyzed may number in the thousands or millions, the people involved in the specific discussion on Twitter are not necessarily representative of the total population, nor indeed of everyone using the Internet. Similarly, the presence of links in tweets, blog posts, or on discussion boards does not mean the endorsement of the site linked to, and it certainly does not imply that people seeing the link will follow it. As noted earlier, too, how to define the life span of links and connections between users is a question still to be definitively answered when examining online communication.

The findings from projects tracking a specific group of users or sites, as with the study of French and Australian political bloggers, are also subject to limitations. While the analysis may draw on large data sets, it is highly unlikely that the data will reflect all posts by every political blogger in Australia and France, for example. Although some online communication platforms, such as discussion boards, might provide more closed environments for research, sites such as Wordpress, Blogger, or Twitter, which are not restricted by paywalls or required technical knowledge, have extensive user bases. Instead of trying to track the entire network forming the blogosphere, for example (which, with the presence of locked, and private blogs, is nearly impossible), the research here follows a "partial network" approach (Hogan, 2008). Here, small subsets of the network provide a microcosm of the wider network, with findings and patterns extrapolated upon for more general conclusions about online activity. However, it is still important to note that the research is not studying *all* participants within online discussions. There are also significant ethical questions around collecting online data, and how to use this within research, which have not been definitively answered even for web-based content that is publicly accessible. While it is not the aim of this article to discuss debates of online ethics, these questions will need to be addressed in projects studying Internet-mediated activity.

Limits also apply to the scope of the topical networks themselves; for example, the blogging case study featured here provide an overview of activity within the blogospheres in question. However, it does not take into account any discussions on the same topic published on other websites or social media platforms, or indeed offline. Further topical networks might draw upon multiple websites for their analysis, but this was not the aim of the initial research and is beyond the scope of this article.

Finally, while topical network analysis brings together aspects of quantitative and qualitative methodologies, linking such processes as textual analysis and social network analysis, additional work is required to bring out further detail about who is contributing to the discussion and why. More qualitative work would help to further examine the motivations and rationale behind posting, commenting, or linking. For example, interviewing participants about their uses of online communication and interest in particular topics would provide new, more nuanced information than might be found on the websites in question.

CONCLUSION

The case study outlined here demonstrates how topical network analysis complemented the wider patterns of activity tracked within the total data set used in this project. This example provides an initial account of the use of a method that has important applications for further studies into online communication. As research continues into the dynamics of conversations online, investigating how discussions start and spread, and which topics gain traction where, topical network analysis allows for a consistent approach to identifying, examining, and comparing different discussions within a single data set.

Although it was developed for studying political blogging, the topical network method is transferable across different platforms. By using a mixture of qualitative and quantitative methods, some of which are outlined in this article, research can move beyond the large-scale overviews of analysis into "Big Data," and focus on specific activity within these data sets. This is not to undervalue the insights provided by "Big Data"; topical networks are intended not to replace the analysis of large data sets, but rather to provide additional detail and nuance in examining the online activity tracked in these projects.

There is further scope for developing the topical network method here, particularly by examining multiple platforms concurrently. The discussion of a particular event is not limited to Twitter alone; nor are the participants. Future research may track the dynamics of specific conversations not just within the blogosphere, but across social media in general. As with the single-platform case study outlined here, identifying topical networks within this space will support ongoing research into online activity. By comparing different conversations within a wider data set, the method enables researchers to develop further conclusions than would be possible from looking at a single case study or the baseline data alone. In doing so, topical networks provide more grounded information about how a platform is used, what its users are contributing, and how discussions online may suddenly appear, and just as quickly fade away.

REFERENCES

Adamic, L. A. (2008). The social hyperlink. In J. Turow & L. Tsui (Eds.), *The hyperlinked society: Questioning connections in the digital age* (pp. 227–249). Ann Arbor: University of Michigan Press and University of Michigan Library.

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Bruns, A. (2007). Methodologies for mapping the political blogosphere. *First Monday*, *12*(5). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/235

Bruns, A. (2012). How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Information*, *Communication & Society*, *15*(9), 1323–1351. doi:10.1080/1369118X.2011.635214

Bruns, A., & Burgess, J. (2011). #ausvotes: How Twitter covered the 2010 Australian federal election. *Communication*, *Politics & Culture*, *44*(2), 37–56.

Bruns, A., Burgess, J., Kirchhoff, L., & Nicolai, T. (2012, March). *Mapping the Australian twittersphere*. Paper presented at Digital Humanities Australasia, Canberra.

Dahlgren, P. (2009). *Media and political engagement: Citizens*, *communication*, *and democracy*. New York, NY: Cambridge University Press.

Elmer, G. (2006). Re-tooling the network: Parsing the links and codes of the web world. *Convergence: The International Journal of Research into New Media Technologies*, *12*(1), 9–19. doi:10.1177/1354856506061549

Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere: Politics and dissent online. *New Media & Society*, *12*(8), 1225–1243. doi:10.1177/1461444810385096

Halavais, A. (2008). The hyperlink as organizing principle. In J. Turow & L. Tsui (Eds.), *The hyperlinked society: Questioning connections in the digital age* (pp. 39–55). Ann Arbor: University of Michigan Press and University of Michigan Library.

Highfield, T. (2011). *Mapping intermedia news flows: Topical discussions in the Australian and French political blogospheres*. Doctoral thesis, Queensland University of Technology, Brisbane, Australia. Retrieved from http://eprints.qut.edu.au/48115

Highfield, T., Kirchhoff, L., & Nicolai, T. (2011). Challenges of tracking topical discussion networks online. *Social Science Computer Review*, *29*(3), 340–353. doi:10.1177/0894439310382514

Hogan, B. (2008). Analyzing social networks via the Internet. In N. Fielding, R. M. Lee, & G. Blank (Eds.), *The Sage handbook of online research methods* (pp. 141–160). London, UK: Sage.

Jang, S. M., & Park, Y. J. (2012). The Internet, selective learning, and the rise of issue specialists. *First Monday*, *17*(5). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3888/3206

Kelly, J., & Etling, B. (2008). *Mapping Iran's online public: Politics and culture in the Persian blogosphere*. Berkman Center for Internet & Society: Harvard Law School. Retrieved from http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public

Lindgren, S., & Lundström, R. (2011). Pirate culture and hacktivist mobilization: The cultural and social protocols of #WikiLeaks on Twitter. *New Media & Society*, *13*(6), 999–1018. doi:10.1177/1461444811414833

Macias, W., Hilyard, K., & Freimuth, V. (2009). Blog functions as risk and crisis communication during Hurricane Katrina. *Journal of Computer-Mediated Communication*, *15*(1), 1–31.

Maireder, A., Ausserhofer, J., & Kittenberger, A. (2012). Mapping the Austrian political Twittersphere: How politicians, journalists and political strategists (inter-)act on Twitter. In P. Parycek & N. Edelmann (Eds.), *Proceedings of CeDem12 Conference for E-Democracy and Open Government* (pp. 151–164). Krems, Austria: Danube University. Retrieved from http://phaidra.univie.ac.at/o:154914

Marres, N. (2006). Net-work is format work: Issue networks and the sites of civil society politics. In J. Dean, J. W. Anderson & G. Lovink (Eds.), *Reformatting politics: Information technology and global civil society* (pp. 3–17). Hoboken, NJ: CRC Press.

Rogers, R. (2009). *The end of the virtual: Digital methods*. Amsterdam, The Netherlands: Vossiuspers UvA.

Schneider, S. M., & Foot, K. A. (2005). Web sphere analysis: An approach to studying online action. In C. Hine (Ed.), *Virtual methods: Issues in social research on the Internet* (pp. 157–170). Oxford, UK: Berg.

ABOUT THE AUTHOR

Tim Highfield is a Research Fellow with the ARC Centre of Excellence in Creative Industries and Innovation (CCI), Queensland University of Technology, and a sessional academic with Curtin University. He was awarded his PhD from Queensland University of Technology in 2011. His doctoral research investigated political blogging in Australia and France. Currently his research focuses on social media use around politics, sport, and popular culture.

A Study of 250 million Facebook Users Reveals the Web Isn't As Polarized As We Thought

By Farhad Manjoo | Posted Tuesday, Jan. 17, 2012, at 11:00 AM
| Posted Tuesday, Jan. 17, 2012, at 11:00 AM

Slate.com

ENABLE SOCIAL READING

The End of the Echo Chamber

A study of 250 million Facebook users reveals the Web isn't as polarized as we thought.

Today, Facebook is publishing a study that disproves some hoary conventional wisdom about the Web. According to this new research, the online echo chamber doesn't exist.

Illustration by Alex Eben Meyer.

This is of particular interest to me. In 2008, I wrote True Enough, a book that argued that digital technology is splitting society into discrete, ideologically like-minded tribes that read, watch, or listen only to news that confirms their own beliefs. I'm not the only one who's worried about this. Eli Pariser, the former executive director of MoveOn.org, argued in his recent book The Filter Bubble that Web personalization algorithms like Facebook's News Feed force us to consume a dangerously narrow range of news. The echo chamber was also central to Cass Sunstein's thesis, in his book Republic.com, that the Web may be incompatible with democracy itself. If we're all just echoing our friends' ideas about the world, is society doomed to become ever more polarized and solipsistic?

It turns out we're not doomed. The new Facebook study is one of the largest and most rigorous investigations into how people receive and react to news. It was led by Eytan Bakshy, who began the work in 2010 when he was finishing his Ph.D. in information studies at the University of Michigan. He is now a researcher on Facebook's data team, which conducts academic-type studies into how users behave on the teeming network.

Bakshy's study involves a simple experiment. Normally, when one of your friends shares a link on Facebook, the site uses an algorithm known as EdgeRank to determine whether or not the link is displayed in your feed. In Bakshy's experiment, conducted over seven weeks in the late summer of 2010, a small fraction of such shared links were randomly censored—that is, if a friend shared a link that EdgeRank determined you should see, it was sometimes not displayed in your feed. Randomly blocking links allowed Bakshy to create two different populations on Facebook. In one group, someone would see a link posted by a friend and decide to either share or ignore it. People in the second group would not receive the link—but if they'd seen it somewhere else beyond Facebook, these people might decide to share that same link of their own accord.

By comparing the two groups, Bakshy could answer some important questions about how we navigate news online. Are people more likely to share information because their friends pass it along? And if we are more likely to share stories we see others post, what kinds of friends get us to reshare more often—close friends, or people we don't interact with very often? Finally, the experiment allowed Bakshy to see how "novel information"—that is, information that you wouldn't have shared if you hadn't seen it on Facebook—travels through the network. This is important to our understanding of echo chambers. If an algorithm like EdgeRank favors information that you'd have seen anyway, it would make Facebook an echo chamber of your own beliefs. But if EdgeRank pushes novel information through the network, Facebook becomes a beneficial source of news rather than just a reflection of your own small world.

That's exactly what Bakshy found. His paper is heavy on math and network theory, but here's a short summary of his results. First, he found that the closer you are with a friend on Facebook—the more times you comment on one another's posts, the more times you appear in photos together, etc.—the greater your likelihood of sharing that person's links. At first blush, that sounds like a confirmation of the echo chamber: We're more likely to echo our closest friends.

But here's Bakshy's most crucial finding: Although we're more likely to share information from our close friends, we still share stuff from our weak ties—and the links from those weak ties are the most novel links on the network. Those links from our weak ties, that is, are most likely to point to information that you would not have shared if you hadn't seen it on Facebook. The links from your close ties, meanwhile, more likely contain information you would have seen elsewhere if a friend hadn't posted it. These weak ties "are indispensible" to your network, Bakshy says. "They have access to different websites that you're not necessarily visiting."

The fact that weak ties introduce us to novel information wouldn't matter if we only had a few weak ties on Facebook. But it turns out that most of our relationships on Facebook are pretty weak, according to Bakshy's study. Even if you consider the most lax definition of a "strong tie"—someone from whom you've received a single message or comment—most people still have a lot more weak ties than strong ones. And this means that, when considered in aggregate, our weak ties—with their access to novel information—are the most influential people in our networks. Even though we're more likely to share any one thing posted by a close friend, we have so many more mere acquaintances posting stuff that our close friends are all but drowned out.

In this way, Bakshy's findings complicate the echo chamber theory. If most of the people we encounter online are weak ties rather than close friends, and if they're all feeding us links that we wouldn't have seen elsewhere, this suggests that Facebook (and the Web generally) isn't simply confirming our view of the world. Social networks—even if they're dominated by personalization algorithms like EdgeRank—could be breaking you out of your filter bubble rather than reinforcing it.

Bakshy's work shares some features with previous communications studies on networks, and it confirms some long-held ideas in sociology. (For instance, the idea that weak ties can be important was first floated in a seminal 1973 study by Mark Granovetter.) It also confirms a few other recent studies questioning the echo chamber, including the economists Matthew Gentzkow and Jesse Shapiro's look at online news segregation.

But there are two reasons why Bakshy's research should be considered a landmark.

A study out today by the Facebook data team strikes a blow against the idea that there is an online echo chamber Justin Sullivan/Getty Images.

First, the study is experimental and not merely observational. Bakshy wasn't just watching how people react to news shared by their friends on Facebook. Instead, he was able to actively game the News Feed to create two different worlds in which some people get a certain piece of news and other, statistically identical, people do not get that news. In this way, his study is like a clinical trial: There's a treatment group that's subjected to a certain stimulus and a control group that is not, and Bakshy calculated the differences between the two. This allows him to draw causal relationships between seeing a link and acting on it: If you see a link and reshare it while some other user does not see the link and does not share it, this means that the Facebook feed was responsible for the sharing.

The other crucial thing about this study is that it is almost unthinkably enormous. At the time of the experiment, there were 500 million active users on Facebook. Bakshy's experiment included 253 million of them and more than 75 million shared URLs, meaning that in total, the study observed nearly 1.2 billion instances in which someone was or was not presented with a certain link. This scale is unheard of in academic sociological studies, which usually involve hundreds or, at most, thousands of people communicating in ways that are far less trackable.

At the same time, there's an obvious problem with Bakshy's study: It could only occur with the express consent of Facebook, and in the end it produced a result that is clearly very positive for the social network. The fact that Facebook's P.R. team contacted me about the study and allowed me to interview Bakshy suggests the company is very pleased with the result. If Bakshy's experiment had come to the opposite conclusion—that, say, the News Feed does seem to echo our own ideas—I suspect they wouldn't be publicizing it at all. (Bakshy told me that he has "a good amount of freedom" at the company to research whatever he wants to look into about the social network, and that no one tells him what to investigate and what to leave alone. The study is being submitted to peer-reviewed academic journals.)

Also, so as not to completely tank the ongoing sales of my brilliant book, I'd argue that Bakshy's study doesn't indemnify the modern media against other charges that it's distorting our politics. For one thing, while it shows that our weak ties give us access to stories that we wouldn't otherwise have seen, it doesn't address whether those stories differ ideologically from our own general worldview. If you're a liberal but you don't have time to follow political news very closely, then your weak ties may just be showing you lefty blog links that you agree with—even though, under Bakshy's study, those links would have qualified as novel information. (Bakshy's study covered all links, not just links to news stories; he is currently working on a follow-up that is more narrowly focused on political content.)

What's more, even if social networks aren't pushing us toward news that confirms our beliefs, there's still the question of how we interpret that news. Even if we're all being exposed to a diverse range of stories, we can still decide whose spin we want—and then we go to the Drudge Report or the Huffington Post to get our own views confirmed.

Still, I have to say I’m gratified by Bakshy’s study. The echo chamber is one of many ideas about the Web that we’ve come to accept in the absence of any firm evidence. The troves of data that companies like Facebook are now collecting will help add some empirical backing to our understanding of how we behave online. If some long-held beliefs get overturned in the process, then all the better.

Theory, Culture & Society
http://tcs.sagepub.com/

Reassembling Social Science Methods: The Challenge of Digital Devices

Evelyn Ruppert, John Law and Mike Savage

Theory Culture Society published online 14 May 2013
DOI: 10.1177/0263276413484941

Theory Culture & Society

Theory, Culture & Society
0(0) 1–25

DOI: 10.1177/0263276413484941
tcs.sagepub.com
SAGE

Reassembling Social Science Methods: The Challenge of Digital Devices

Evelyn Ruppert
Goldsmiths College, University of London, UK

John Law
Open University, UK

Mike Savage
London School of Economics, UK

Abstract

The aim of the article is to intervene in debates about the digital and, in particular, framings that imagine the digital in terms of epochal shifts or as redefining life. Instead, drawing on recent developments in digital methods, we explore the lively, productive and performative qualities of the digital by attending to the specificities of digital devices and how they interact, and sometimes compete, with older devices and their capacity to mobilize and materialize social and other relations. In doing so, our aim is to explore the implications of digital devices and data for reassembling *social science methods* or what we call the *social science apparatuses* that assemble digital devices and data to 'know' the social and other relations. Building on recent work at CRESC on the social life of methods, we recommend a genealogical approach that is alive to the ways in which digital devices are simultaneously shaped by social worlds, and can in turn become agents that shape those worlds. This calls for attending to the specificities of digital devices themselves, how they are varied and composed of diverse socio-technical arrangements, and are enrolled in the creation of new knowledge spaces, institutions and actors. Rather than exploring what large-scale changes can be revealed and understood through the digital, we argue for explorations of *how* digital devices themselves are materially implicated in the production and performance of contemporary sociality. To that end we offer the following nine propositions about the implications of digital data and devices and argue that these demand rethinking the theoretical assumptions of social science

Corresponding author:
Evelyn Ruppert, Goldsmiths College, University of London, Lewisham Way, New Cross, London, SE14 6NW.
Email: e.ruppert@gold.ac.uk

methods: transactional actors; heterogeneity; visualization; continuous time; whole populations; granularity; expertise; mobile and mobilizing; and non-coherence.

Keywords
actor network theory, big data, digital devices, genealogy, methodology, performativity, transactional data

> In the second industrial revolution, with its automation of the streams of information, the analysis of discourses has yet to exhaust the forms of knowledge and power. Archaeologies of the present must also take into account data storage, transmission, and calculation in technological media. (Kittler, 1990: 369)

Digital devices and data are becoming ever more pervasive and part of social, commercial, governmental and academic practices. Different digital platforms mobilize and generate ever growing volumes of data on social and other relations: Twitter, Facebook and MySpace organize and produce data on social networks; online purchasing and browsing on Amazon, LastFM and Google facilitate and generate data on usage and transactions; news media such as the BBC or *The Guardian* track and report data on viewing trends and the popularity of articles; apps on mobile phones generate records on user activities and movements; eGovernment sites log digital interactions between governments and citizens, businesses and employees, and government administrative databases register the service activities and movements of people; and eScience and eHumanities projects compile and analyse immense data sets. There are also numerous digital devices produced by software developers for tracing and visualizing data that is circulated on the worldwide web such as Google's PageRank, Technorati's blog post aggregator, or the Lexis Nexis media aggregator (Beer, 2009). Social worlds are thus saturated, being done and materialized by digital devices and what is increasingly being understood as 'big data' of various kinds.[1] Indeed, the other articles in this special issue take up examples such as visualizing devices in the field of proteomics, BBC social media experiments, new forms of digital cultural engagement and self-reported medical data using online medical research platforms.

It is clear that the digital is also the focus of much scholarly analysis. A quick scan of recently published social science books that engage with the digital reveals the myriad themes and concerns of researchers: computer technologies and infrastructures; networked cultures; living and communicating in a computer age; internet and social media activism; interoperability and standardization; Web 2.0, open access and online collaboration; e-social science; cyberwarfare; cyberspace security, privacy, surveillance and censorship; and e-learning technologies. Many of these themes form part of what is called a growing field of 'digital studies'.

But the aim of our article is both broader and narrower than suggested by these developments. It is broader because we seek to unsettle debates about how the proliferation of the digital is implicated in large-scale social change and remaking the governance and organization of contemporary sociality (for instance, Castells' [1996] network society, or the notion of biopolitics [Rose, 2006; Thacker, 2005]). And it is also narrower in that we are concerned with the implications of digital devices and data for reassembling *social science methods* or what we call the *social science apparatus*. Here we build on our interest in elaborating the social life of methods (which is summarized in Savage's introduction to this special issue) through a specific concern with digital devices as increasingly the very stuff of social life in many locations that are reworking, mediating, mobilizing, materializing and intensifying social and other relations. Focusing, in the spirit of Kittler (2006), on issues of ontology, we argue that we need to attend to how these qualities of digital devices demand rethinking the theoretical assumptions of our social science methods and making those assumptions explicit. While digitization is a complex and indeterminate process of intensification whose effects are uncertain, we suggest that it has the potential to reawaken and rework long-established social and political relations (see also Küchler, 2008).

Our objective is thus to pose questions about the consequences of digital devices for social scientific ways of knowing. If digital devices mediate and are in considerable measure the stuff of social, cultural, economic and governmental lives in contemporary northern societies, then what does this mean for our methods for knowing those lives? When we speak of methods here we mean the specific apparatuses that assemble digital devices and data to 'know' the social and other relations. We are saying that digital devices and the data they generate are both the *material* of social lives and form part of many of the apparatuses for *knowing* those lives. So, for instance, devices such as Twitter materialize new forms of sociality and ways for people to interact and know about themselves and others. At the same time Twitter gives rise to various knowledge practices or methods: academic researchers, data journalists and police surveillance units develop combinations – let's call these *apparatuses* – of analytical procedures (algorithms, software), infrastructures (computers, networks) and personnel (analysts, IT experts) to analyse the data that it generates.[2] So our questions are: what are the relations between the elements that make up different apparatuses and how are digital devices reconfiguring those relations?

We argue that any answer to these questions demands a conceptual understanding of the *specificities* of digital devices and the data they generate. It requires the exploration of their qualities, which are likely to be both similar to and different from those of longer-standing social science methods such as survey research. But before attending to those similarities and differences, we first want to step back to reflect on social theory accounts of how the digital is transformative. Our suggestion is that, despite the important issues that such accounts raise, the specificities of digital devices – their materialities, productivities and mediating capacities – are not explored in this literature. We then turn to briefly note how such specificities are being addressed in the development of social science digital methods. Here we note that digital devices are reworking and mediating not only social and other relations, but also the very assumptions of social science methods and how and what we know about those relations. We argue that this calls not simply for reworking methods technically but also addressing their ontological assumptions. In the second part of the article we thus introduce the notion of 'the social life of methods'. Here we emphasize the constitutive role of social science research methods for modern capitalist societies and suggest that this role is changing. But we cast such change *relationally* by exploring how digital devices interact with other kinds of devices, and how they themselves are both varied and composed of diverse sociotechnical arrangements. Then, in the third section, we examine how we can better register the significance of the digital in terms of the capacities it offers for elaborating and mediating transactional (and especially) social relations, and offer a set of propositions for rethinking the assumptions of social science methods.

In sum, we make three major arguments. First, we suggest that the challenge of informationalism can be understood genealogically by tracing how the material and productive effects of the digital are reconfiguring knowledge spaces and the social science apparatus. Second, we explore the limits of 'external critique' and the extent to which standard methods and conceptual tools help us to understand information from the outside. And then, third, in an attempt to handle the challenge of informationalism 'from within', we develop an immanent critique that draws on Foucault's *dispositifs*, the STS (science and technology studies) concern with inscription devices and Bourdieu's field analysis.

A Digital Age

There is much interest in – and much hyperbole about – the digital. But if we strip away the latter, the capacity of social scientists and cultural theorists to understand the significance of the digital challenge seems less certain. As Mackenzie (2005: 72) wrote:

> Although there has been wide acknowledgement of the mobility, dynamism and operationality associated with information

> networks, understanding the cultural specificity of software or code objects remains difficult.

Towards addressing the cultural specificity of digital code, Mackenzie and Vurdubakis (2011: 4) recently assembled a special issue of this journal that seeks to 'go beyond the restricted (and often restricting) understanding of code as the language of machines'. Instead, they explore codes 'not only in terms of software but also in terms of cultural, moral, ethical and legal codes of conduct' and what they 'tell us about the ways in which the "will to power" and the "will to knowledge" tend to be enacted in the contemporary world'. Importantly, rather than a 'general theory of code', the special issue attends to the specificities of code in domains from social networks to highway engineering.

But Mackenzie and Vurdubakis's approach and contribution stand apart from how the digital is predominantly approached in social theory. One influential approach imparts intrinsic properties to the digital, which is imagined to grow and unfold so that its qualities become more widely disseminated. The suggestion that the digital marks a profound, epochal, rupture in social change is familiar. We are surrounded by claims about the distinctive characteristics of 'knowing capitalism' (Thrift, 2005), 'the information age' (Lash, 2002; Poster, 2001; Webster, 1995), and 'the network society' (Castells, 1996). However, a re-reading of many of these seminal texts a decade later suggests that they treat information technologies and the digital in a derivative way. Rather than offering novel arguments about its revolutionary capacities, reflections on the innovatory character of the digital tend to reflect concerns with epochal change originally developed in the context of other kinds of claims.[3] So, for instance, in writing about the digital, both Castells and Lash rework familiar arguments about globalization, postmodernism, and reflexivity.

Castells' (1996, 1997) seminal work on the 'network society' remains a key reference. This work was responsible for introducing digital technologies fully into the debates about post-industrial social change that had been raging for two decades since Daniel Bell's (1976) *The Coming of Post-Industrial Society*. To this extent, Castells' intervention is the latest in a long line of debates preoccupied with the role of automation – for instance that represented by commentators such as Toffler (1980) and Bell (1976), and even the Marxist analysis of the labour process (Braverman, 1974).[4] Yet at the same time Castells broke new ground by emphasizing the networked character of digital communication. He argued that information can be divided into 'packets' and thus distributed in a non-linear and distributed fashion, an operation essential to contemporary capitalism (see e.g. Castells, 1996: 351–2). In this way he provided a distinctive twist to the familiar claims of Harvey (1989), Giddens (1991) and Beck (1992) about the power of globalization, the break-up of social collectivities, and the creation of new kinds of fluid and mobile identities.

Yet, for all the emphasis on 'the culture of real virtuality', the technological underpinnings of Castells' treatise are relatively underdeveloped. In the way he treats it, the 'information technology' paradigm has five characteristics: (1) 'technologies act on information'; (2) there is 'pervasiveness of effects of new technology'; (3) there is a networking logic; (4) 'flexibility'; and (5) 'convergence of specific technologies into a highly integrated system' (Castells, 1996: 61–2). But there are various problems here. For instance, until the formation of the worldwide web and networked computing, information technology did not obviously have a 'networking logic'. Again, and more specifically, digital devices and their specific modes of operation do not feature in the list. Our suggestion is that Castells is claiming the digital to be of profound social importance, but his work is more easily understood as a restatement of more conventional, pre-digital themes.

It is perhaps Lash (2002) who has placed informationalism on a more elaborated conceptual basis. Following in the spirit of Castells, he sees it as an ushering in of epochal change, and argues that talk of

> [i]nformation society is . . . preferable to postmodernism in that the former says what society's principle is rather than saying merely what it comes after. . . . Second, postmodernism deals largely with disorder, fragmentation, irrationality, whilst the notion of information accounts for both . . . order and disorder . . . . Information is preferable and more powerful as a notion because it operates from a unified principle. (Lash, 2002: 1–2)

This is appealing, yet, in practice, this unified principle is difficult to tease out. Lash reworks Wittgenstein to invoke an idea of 'technological forms of life' (see e.g. Gane, 2004). His aim is to think through the immanent properties of information in order to find a basis for critique that is not external or transcendent to that which it criticizes. Yet it is unclear how successful this is. Like Bauman, he tends to treat the digital as if its deficiencies are its defining features. For instance, it is 'non-linear' and discontinuous:

> technological forms of life are *really* stretched out. They are too long, stretched out too far for linearity. They are so stretched out that they tear asunder. Spatial link and social bond break. (Lash, 2002: 20)

How well this argument works is uncertain (see also Simondon, 1989). Thus it presupposes a linearity that is no longer at work. It poses the question as to how stretched out forms of life have to be before they break. The relational qualities of information are relatively underplayed (it is whatever is transmitted to others; Yoshimi, 2006). And the argument also fits uncertainly with substantial empirical research that shows that the digital is profoundly associated with the making of what might be termed 'local' social relations. For instance, Woolgar's (2002) 'five rules of virtuality', which are derived from a series of detailed case studies of virtual social relations, lead us away from Lash's thesis. Thus Woolgar tells us that 'virtual technologies supplement rather than substitute for real activities', 'the more virtual the more real', and 'the more global, the more local'. Similarly, as Strathern (2000) argues, rather than being decontextualized, the digital actualizes relations and connections that are otherwise beyond perception and thus inherent to the very imagining of social relations. They are materializations of what Latour (1998) has called a traceable social that is being rendered visible. And finally, as Knox et al. (2007) show, the use of digital communication in large corporations is associated with intensive local negotiation. Rather than occupying a 'space of flows' or a virtual informationalized world, digital data is itself a materiality that is 'alive', embodied and mobile. Our point here is that to yoke the digital to epochalist accounts of social change is to treat it as a reflection of familiar theoretical arguments, and tends to direct attention away from the materiality and productivity of digital devices.

Finally, another set of literature relates the digital to emergent forms of biopolitics (Agamben, 2005; Thacker, 2005). Here, and in part drawing on Foucault, the interest is in the productive capacities of the digital to generate new kinds of emergent relations, and most particularly with new conceptions of 'life itself':

> The molecular knowledge of life that has taken shape since the 1960s has been linked to all sorts of highly sophisticated techniques of experimentation that have intervened upon life at this molecular level . . . the laboratory has become a kind of factory for the creation of new forms of molecular life. And in doing so, it has fabricated a new way of understanding life itself. (Rose, 2006: 13)

Thacker (2005: 28) explicitly links these new forms of life to informational politics:

> Information in biopolitics is precisely that which can account for the material and embodied and, furthermore, that which can produce the material, the embodied, the biological, the living – 'life itself'.

Thacker's argument about how digital technology erases boundaries between the natural and the social is also related to claims about globalization and so carries epochalist overtones. For just as theorists of postmodernism made much of the flattening of affect and the dominance of self-referential simulacra, now life itself is seen as complicit with informational and representational processes. This body of literature has also cross-fertilized with recent work on vitalism (e.g. Barry, 2005; Fraser et al., 2006). Here, then, the digital is seen as a way of reconfiguring life 'to conceive life as not confined to living organisms, but as movement, a radical becoming' (Fraser et al., 2006: 3).

But if our interest is with digital devices, it may or may not be productive to focus on 'life itself'. Here the legacy of Foucault's concern with the production of the human subject in disciplinary and governing devices comes through. Yet, as several contributors to this special issue note, rather than privileging the life sciences it becomes important to attend to the more mundane uses of digital devices in information systems, marketing processes and cultural systems, all of which offer different vantage points. Here it isn't the redefinition of life that is important. Instead it is the 'liveliness of data' and the making of transformational agents that come into focus. So while these accounts are provocative and raise vital issues they do not place the digital, in its ubiquity, its routinization and its mundanity, at centre-stage. How then have social scientists engaged with these specificities? One way has been through the development of digital methods.

Digital Methods

Several social science research centres and initiatives have taken up the challenge of digital data and methods. In Europe this includes, for example, the National Centre for e-Social Science (NCeSS, UK), now the Manchester eResearch Centre (MeRC); the Digital Methods Initiative (DMI, Amsterdam); the Oxford e-Research Centre and Oxford Internet Institute (UK); the Bartlett Centre for Advanced Spatial Analysis (UK); the Centre for Research on Socio-cultural Change (UK); the médialab (Sciences Po, France); and the eHumanities Group at the Royal Netherlands Academy of the Arts and Social Sciences (KNAW). To a varying extent these initiatives seek to understand digital devices while, at the same time, developing conceptual framings and innovative methods for analysing their effects.

In the academy, researchers have also adapted social science methods to forms of the digital such as virtual ethnography (Hine, 2000), virtual methods (Hine, 2005, 2006), and digital methods such as the IssueCrawler (Marres and Rogers, 2005; McNally, 2005). The growing availability of digital traces is also promoting a new form of computational social science that relies on the computer-aided manipulation of huge quantities of data (Lazer et al., 2009). Manovich's (2009) work on 'cultural analytics' has shown that disciplines like cultural studies can be transformed by the capacity to compile and analyse unprecedented volumes of digital records. Text mining is also being taken up in many other areas, such as scientometrics (Börner, 2010), computer assisted qualitative data analysis[5] and controversy mapping, whereby text analysis enables tracking and visualizing the alignments and oppositions in actor discourses (Venturini, 2010).

These examples could be extended, but are sufficient to make our point that a number of initiatives are under way to develop social science methods for compiling and analysing digital data. Though varied, some tend to emphasize technical issues – how can we adapt social science methods (e.g. virtual ethnography) or develop new digital methods (e.g. cultural analytics) to know social worlds in new ways? At the same time, by attending to specificities some are identifying and suggesting that the digital is challenging theoretical assumptions of social science methods. This is because emerging methods rely upon and mobilize digital data and devices, which are mostly generated outside the academy in social, commercial and governmental sites. For others, digital methods are giving rise to a new ontology of the social (e.g. Latour, 2010; Rogers, 2009b). It is this direction that we find most promising and which we explore below. Our focus is on how digital data and devices are reconfiguring social science methods and the very assumptions about what we know about social and other relations. To think about this well we do not simply need to rework methods technically, but also to rethink their ontological assumptions including, for instance, their often humanist underpinnings.

Digital Devices and the Social Life of Methods

We need a better analytical grasp of the challenge of the digital than is offered in social theory and technical accounts of method. But how? No doubt there are many possibilities, but our approach is to explore how the social is materialized in and saturated with devices – or what Featherstone (2009) calls 'ubiquitous media' – that are also part of the apparatuses for knowing social lives. So the question is: how do those devices and data get assembled into specific apparatuses to 'know' social and other relations? We use the term 'apparatus' to suggest that methods are purposeful assemblages, just as Foucault used the notion of *dispositif* and Latour that of the inscription device. Foucault (1980: 194) maintained that a *dispositif* is:

> a thoroughly heterogeneous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions – in short, the said as much as the unsaid. Such are the elements of the apparatus.

For Foucault an *episteme* is discursive. It sets limits to what can and cannot be said in a field. However, a *dispositif* (often translated into English as ‘apparatus’) includes an array of material, institutional and behavioural elements. For example, in relation to sexuality, it consists of a heterogeneous ensemble that includes ‘the body, the sexual organs, pleasures, kinship relations, interpersonal relations and so on’ (Foucault, 1980: 210). But a similar argument works for the digital: it is composed of many different kinds of elements, ranging from computer networks, scanners, algorithms, software and applications to different actors, institutions, regulations and controversies. Devices generate digital data (versions of what Latour [1990] calls inscriptions) in the context of sets of social and technical practices and relations. And those devices and data are assembled together to analyse and visualize Castells’ ‘informationalization’. It is through such cascades of inscriptions – for instance from reams of data to indices – that simpler and more mobile digital inscriptions are often generated. And if some of those inscriptions have become more or less stable, difficult to undo or immutable, then this is because of the scale of investment (literal and metaphorical) that has gone into making them up. It has become too ‘expensive’ to undo them. Latour (1990: 15–16, italics in original) warns us, therefore, that:

> the precise focus should be carefully set, because it is not the inscription by itself that should carry the burden of explaining the power of science; it is the inscription as *the fine edge* and *the final stage* of a whole process of mobilisation.... So, the phenomenon we are tackling is *not* inscription per se, but the *cascade* of ever simplified inscriptions that allow harder facts to be produced at greater cost.

Latour is talking about natural science, but offers a valuable provocation for our concerns here. The suggestion is that we need to be attentive not only to the digital in general terms, but to the more specific mobilizations which allow the digital to be rendered visible and hence effective in particular locations. In this way, we can see the extensive history of ‘failed’ digital projects as entirely germane. Our first suggestion then is that social science methods can themselves be treated as situated cascades of *dispostifs* and inscriptions. For example, Rogers (2009a) distinguishes the ‘natively digital’ – that is, data generated by online devices – and the digitization of ‘traditional’ data-gathering devices such as surveys. Each enrols different devices, arrangements and relations. And this leads us to our second suggestion that such cascades are simultaneously embedded in and shaped by social worlds, and can in turn become agents that act in and shape those worlds. In a nutshell, this is one meaning of ‘the social life of methods’, which is elaborated in the introductory essay to this special issue. But if we are to understand this in the context of the digital, then we need to attend to the lives and specificities of devices and data themselves: where and how they happen, who and what they are attached to and the relations they forge, how they get assembled, where they travel, their multiple arrangements and mobilizations, and, of course, their instabilities, durabilities and how they sometimes get disaggregated too.

This approach draws in part on STS and more specifically from actor network theory's concern with the agency of objects. Thus much STS literature argues that scientific and technical objects are socially efficacious. Early STS work tended to explore natural science (Latour and Woolgar, 1986) and technologies (Callon, 1986; Law, 2002). However, more recently, there has been increasing STS work on social science techniques and methods, with Callon (1998, 2007) and MacKenzie's (2008) work on the performativity of economics, for example. There have been important studies of how social scientific censuses, mapping and survey techniques are associated with the generation of powerful social entities such as the 'national economy' (Mitchell, 2002), caste groups (Dirks, 2001), social aggregates such as classes (Savage, 2010) and populations (Ruppert, 2009, 2011).

The conclusion is that in relation to digital devices, then, we need to get our hands dirty and explore their affordances: how it is that they collect, store and transmit numerical, textual, aural or visual signals; how they work with respect to standard social science techniques such as sampling and comprehensiveness; and how they relate to social and political institutions. To tease out these specificities and qualities it is useful to consider, in a historical register, how digital devices compare with other, older, socio-technical devices, and consider the different affordances that they offer in a nuanced manner. This is an approach also taken up by 'media archaeologists' who challenge accounts of the 'newness' of various forms of digital media by examining how they often rework 19th-century technologies (Huhtamo and Parikka, 2011: 1).

The available work reveals that the digital has not displaced sensuous human interaction, but has instead reworked sophisticated sets of devices that pre-existed it. These include the technologies of surveillance and control dissected by Foucault (1976) together with the arts of government, but also and perhaps more critically, involve a battery of social science devices that proliferated in the second half of the 20th century. So, for instance, the period from 1950 to 2000 saw a dramatic intensification of social research methods, notably the sample survey (first conducted on a large national scale in the UK in the 1930s) and interview methods (see Savage, 2010). These methods were championed as mechanisms to elicit everyday, ordinary, and mundane accounts, and were not only embodied in research agencies but also in popular media and corporate customer services departments. They departed from previous research repertoires based on observational technologies, which depended on the implicit authority of the 'knowing' observer, who was deemed able to delineate a moralized account of social relations. Again, as Thrift (2005) argues, new research methods became fully enmeshed in the circuits of ‘knowing capitalism’, in which the systematic gathering of information about customers, clients, employees and competitors became routine to corporate strategy.

Our suggestion is that it is the dominance of the ‘social science apparatus’ and its methods that is being called into question by the digital. Three features are important for our argument here. First, the devices that make up social science methods differ from many in the natural sciences by being physically unspectacular. They are not embedded in laboratories or huge pieces of machinery. Instead methods rely on chains of interconnected and cascading devices, and consist of largely statistical procedures, with relatively large corps of skilled ‘administrators’ (interviewers, surveyors, enumerators, etc.), and simple devices such as clip boards, sheets of paper and, more recently, laptop computers to record social evidence. In short, they have entered the mundane circuits of social relations with no consecrated ‘laboratories’. But this mundanity is being challenged in part by the digital.

Second, these social science methods and their devices are deeply implicated in the formation of human subjects. The census and the survey both presuppose, yet also enact, the knowing, self-aware individual, who is able to account for him or herself. Ruppert (2007, 2011) analyses how censuses produce and engage subjects in identifying with classification schemes that principally measure biographical characteristics such as gender, income, occupation and ethnicity, self-elicited identifications that focus on social categories. Whether individuals or enumerators complete census forms, subjects require particular reflexive capacities and agencies for the device to operate, including the ability to categorize and creatively make themselves legible. Similarly, Osborne and Rose (1999) describe how the production of ‘opinioned or opinionated people’ was part and parcel of the creation of the technology of public opinion research in the early 20th century. They argue that genealogies of devices can be paralleled with genealogies of persons: in the case of public opinion polls, people ‘learned’ to have opinions, became opinioned or opinionated, which means that opinion polls ‘made up’ people.

In another example, Savage (2010) examines the way that the sample survey abstracts lone individuals from their household arrangements (which had been the traditional focus of community studies) and allows the very concept of the non-sexed individual to come to the fore. (Within earlier traditions of community research, sexed and household characteristics were seen as given, primordial.) If this is right, then social science research devices were critically implicated in the formation of the self-organizing and self-accounting individual. Those devices, together with the recent, largely post-Second World War, ‘social science apparatus', which were based on the primacy of enumerating and sampling individual accounts (through censuses, interviews and surveys), helped champion a biopolitics of the 'human individual', detached from his or her environment. But all of this is being challenged and indeed undermined with the development of digital devices (Savage, 2010).

Third, the social science apparatus was dependent on a specific infrastructure of humans and devices to generate appropriate 'social data'. Without teams of interviewers, survey instruments, census enumerators and the like, such an apparatus would not have existed. This kind of knowledge is not a by-product of *other kinds* of data-generating devices and processes. Rather, this apparatus operates in a similar way to the skilled physician, standing outside the social body, and intervening in it with various devices to collect, array, analyse and codify samples of social tissue. These procedures are in keeping with how Rose (1991) defines liberal expertise, which is dependent on the knowing expert, and with Bauman's (1987) invocation of the 'intellectual as legislator'.

But what does it mean if we argue that social science methods are becoming dependent on digital devices not of their making? One answer is that the digital is bound up with processes of re-territorialization, and the creation of new knowledge spaces, institutions, actors, devices and apparatuses. But *specificity* is needed if we are to make this argument. We need to be wary of large claims. It is, for instance, likely (we'll argue this below) that these apparatuses draw from, or resonate with, older technologies of surveillance. Rather than a large-scale and external emphasis on flows and mobilities, or epochal change, we are suggesting that it is important to attend to the emerging stabilizations and fixities being performed in cascades of (partly social science) devices in particular locations. And rather than simply exploring what can be revealed and understood through such devices, it becomes important to explore *how* digital devices themselves are materially implicated in the production, performance and knowledge of contemporary sociality. So how to think about this?

The Challenge of Digital Devices: Nine Propositions for Reassembling Social Science Methods

In line with what we have been saying about apparatuses, inscription devices and their agential capacities, we want to argue that *digital devices observe and follow activities and 'doings'* – often, but not always or exclusively, those of people. Such 'doings' might include physical movements, but have more to do with *actions* (transactions, choices, statements, interactions) and their *traceability*. From loyalty cards, online purchasing, blogs, mobile phones, websites, wikis and social networking sites to government administrative databases, patents, reports and scientific and newspaper articles there are, as we have argued, heterogeneous and multiple cascades of devices. Included in such cascades are numerous applications and software for simplifying, summarizing, visualizing and analysing digital data. Within these cascades a device can make, compile and transmit digital data and/or remake, analyse and translate data into information and interventions. But, this is the crucial point, all of these digital devices are modes of observation that trace and track doings. In the context of people, instead of tracking a subject that is reflexive and self-eliciting, they track the *doing subject*.[6]

So how, then, do social relations emerge and how are they linked to the apparatuses of social science? On the one hand, we want to suggest, controversially, that we are seeing a partial return to an older, observational kind of knowledge economy, based on the political power of the visualization and mapping of administratively derived data about whole populations. On the other hand, as a genealogical approach demands, we need to attend to the differential problems, concerns and devices through which observation is being performed by the digital and its material and productive effects, including the reconfiguration of knowledge spaces and social science expertise. However, we cannot attempt such detailed genealogies here. Instead, we offer nine propositions that arise from social science analyses of digital data and devices and argue that these demand rethinking the theoretical assumptions of social science methods.

(1) *Transactional actors*. Whereas interview-based social science methods elicit individual accounts and make these the centrepiece of social research, digital devices record data switches (exchanges), as two (or more) parties (including people and things) do business, exchange and interact. They are thus not derived from conscious intervention by the knowing researcher, but are the by-product of switches and what Rogers (2009a) calls the natively digital (e.g. data generated from online purchasing). These switches can be multiple, complex and minute. For example, a graphic illustration of mobile phone transactions demonstrates the structure of communication flows between members of a network. It is a form of social network analysis, with no data at all on specific individuals, but instead a mapping of specific transactions between parties. It thus has affinities with the field analysis of Kurt Lewin's sociometric social psychology, the poverty studies of Charles Booth, and the interwar Chicago School. Here, the focus of inquiry is not on the individual factors that affect behaviour, but on the spatial flows of behaviours and contacts: contagion, pollution, influence, etc. Similarly, data generated by digital devices allow non-individualist and non-humanist accounts of the social, where the play of fluid and dynamic transactions is the focus of attention.

(2) *Heterogeneity*. Building on this first point, the extent to which digital data sources relate to people – or indeed to populations of people – is limited. The fact that some of those transactions are then

pinned to people who are said to engage in doing is important, but it is not given in the logics of transaction. This thought can be extended in several directions. First, there are many transactions – consider the movement of items through logistics networks – that don't directly have to do with people at all. Entities quite other than people make up these networks and the patterns that they reveal. Second, even if people are involved – as often they are – they are being disassembled into sets of specific transactions or interactions. It may or may not happen that they are reassembled into 'people'. In some sense, then, transactional 'doers' may be people, but in and of itself this has no special significance. Indeed, to say as we just did that people 'are being disassembled into transactions or interactions' is already to risk missing the point. People aren't disassembled. Rather, and perhaps exceptionally, they are sometimes assembled. Third, then, and more generally, it needs to be said that the move to the digital is *a move to heterogeneity*. Perhaps, following Tarde and Latour, we need to say that the social is about *heterogeneous association* rather than societies and people. It is about factors, impulses, risk profiles, and circuits and the post-demographic, as Rogers (2009b) has suggested. To this extent, humanist conceptions of society are being eclipsed.

(3) *Visualization*. The re-emergence of visualization as key to social analysis is striking. This stands in stark contrast to the hegemonic use of numerical and textual devices within the social science apparatus (in this respect, the social sciences parted company from the natural sciences, where visualizations have always enjoyed more legitimacy). In the social science apparatus, the marked differentiation between numbers and text takes historical form, since the two have not always been defined in opposition (Kittler, 2006). But in the move to the digital visualization now becomes a means of showing how 'excessive' information can be reduced to a form in which it can be meaningfully, if partially, rendered for interpretation. In this way, as Amoore (2009) shows, aesthetic criteria can be re-introduced into the use of digital data sources. Rather than statistical analyses (through modelling procedures), visualization becomes a summarizing inscription device for stabilizing and representing patterns so that they can be interpreted. Although different in construction to (for instance) Booth's 19th-century poverty maps, they nonetheless share a common concern with observing patterns, circulation, flows, and boundary maintenance and leakage.

(4) *Continuous, rather than bundled time*. Both interviews and surveys can detect change, not by comparing disparate sources but through internal inspection of unitary data or linked datasets. In the qualitative interview, narratives disclose temporal sequencing through story devices. Surveys permit temporal analysis through comparison of age groups (quasi cohort analysis), or, in the case of panel studies, by tracking the same individual at different time points. Both thereby allow trends to be

discerned through internal analysis, rather than through the messy amalgamation of different sources, as practised by historians. These procedures involved the eclipse of landscaped and territorial approaches to the social, which were grounded in earlier generations of observational social research, due to the way that they depend on abstracting sampled individuals from their environment, increasingly by using the national boundary as the unit in which societies were deemed to operate. In these analyses, time is treated as linear, as a set of standardized points (e.g. years) between which comparisons can take place. Censuses take fixed 'snapshots' of populations every five or ten years and then compare quantities of social categories between intervals to reveal change. By contrast, new data sources such as social network platforms and digitized government administrative data deploy continuous time and constitute on-going and dynamic measurements of the movements and transactions of populations (Ruppert, 2010). For example, eBorders databases focus on the identification of factors that shape 'unknown futures' (Amoore, 2009). Such a perspective offers a shifting platform on which to view change as risk factors are modified. However, some digital data is not routinely archived and, because it is not focused on the individual, it has no identifying unit that can allow for comparison over time.[7] In many cases it thus elicits flat, pliable registers of populations.

(5) *Whole populations*. Social science methods depend on sampling, and hence social knowledge is generated on the basis of data derived from only a small selection of points, which are then generalized into accounts of social aggregates through statistical procedures. New digital data sources work on the basis of entire systems of records, so that the aggregate is not as important as the individual profile. Through these means, there is a return to a problematic of 'whole populations', in which it is not enough to know aggregate properties of the social world, but to know how everyone and every transaction can be scanned, monitored, and subject to analysis and intervention. Every individual who uses a Tesco clubcard has a unique 'DNA' profile which records their spending patterns, and those who analyse such data insist on its value in allowing a granular knowledge that surpasses knowledge of aggregated social groups. (Instead, aggregated social groups are derived inductively as discussed below.) This concern with whole populations also elicits a descriptive mode of analysis, which clusters and classifies to produce social maps that are simultaneously moralized and normative. Good examples of these are the extensive geodemographic profiles widely used within marketing. It is instructive to note the similarities between the 'lifestyle' maps produced by these systems and the maps generated by Booth and Rowntree a hundred years earlier.

(6) *Granularity*. 'The devil lies in the detail' of new data sources. There is a suspicion of aggregated properties that are derived deductively. Instead, the focus is on particularistic identifiers. In credit scoring,

security services, social welfare or criminal targeting, and commercial marketing, it is particular suspect, risky or at-risk populations that are sought out and identified. Databases such as Experian, for example, classify unique postcodes. In such processes aggregates may also be derived (as clusters of granular cases), but these are inductively created and not 'imposed' onto data sources. Similarly, government administrative databases record multiple cross-agency transactions that reveal detailed and unique identifications of populations when they are joined up. This focus on granularity drives forward a concern with the microscopic, the way that amalgamations of databases can allow ever more granular, unique, specification.[8] This is part of a desire for wholeness, an embrace of the total and comprehensive which is never-ending but which generates a politics of mash-ups, compilation and data assemblage. Perhaps this helps to explain the attraction of Deleuzian perspectives, where the empirical is held not to be outside the concept but in interaction with it.[9] The subject is materialized by digital devices in new ways and may be understood as a monad, a conceptualization that Latour and others have advanced in relation to digital methods such as controversy mapping.[10]

(7) *Expertise*. Survey and interview methods demand intervention from the expert social scientist. The idea that these experts can actually intervene and generate empirical data is one that was largely new in the post-war years, and eclipsed their older, gentlemanly role in which they used by-product data generated by inspectors, social workers and the like. The idea that experts had to intervene in the social world to gather appropriate data that would otherwise be absent and would limit social science was absolutely central to the emergence of critical social science. However, new digital sources create data as a by-product. One does not have to conduct special questionnaire or interview research on Amazon customers to identify which other books customers are likely to buy. Such data is routinely gathered through normal transactional processes and allows customers to be bombarded with information about what people like themselves have bought. This is comparable to the way that social knowledge in the 19th and 20th centuries was generated from routine administrative practices of social workers, school inspectors and the like. This is now the source of population knowledge to which governments are 'returning'. Some governments, for instance, have replaced, or are planning to replace, traditional questionnaire-based censuses with administrative records, which at one time were the mainstay of population knowledge (Ruppert, 2010). Data generated as a by-product of everyday transactions with governments (registration, taxation, benefits) are recordings of exchange processes and do not rely on experts to intervene to elicit knowledge of populations. Whether in commercial or governmental domains, different experts, such as computing engineers and software designers or the emerging profession of 'data scientist', are becoming more prominent mediators.

(8) *Mobile and mobilizing*. Digital data sources, and especially Web 2.0 technologies, also allow various publics to be enrolled and enacted in the digital in active ways (Ruppert and Savage, 2012). There is a range of freely available online data, ‘apps’, software visualization devices and so on. For Stiegler, these produce ‘an associated milieu in the sense that all members belonging to the milieu participate in it and are functions of the milieu’ (Venn et al., 2007: 335). We once again need to remind ourselves that, rather than being new, this is in many regards a return to the tradition of Mass Observation and the various field research activities of the middle 20th century, all of which emphasized how publics could research themselves through writing and observing. This current persisted well into the 1960s, perhaps most notably in the Consumers Association journal *Which* that relied on letters from the public to judge the quality of products. By contrast, the social science repertoires of the post-war years sought to construct respondents in more passive forms so that their accounts could be rendered comparable and equivalent to each other. Be that as it may, what is different is both the location and relation of publics to the numerous devices that make up the digital. Publics are now enacted and enabled to intervene actively by making up their own devices as well as by contributing to the dominance of particular devices through their mass take-up. Here we need to account for the *mobility of the digital itself*, and the capacity for the circulation, sharing and take-up of devices and data across numerous sites that increasingly transcend institutional boundaries.[11]

(9) *Non-coherence*. The proliferation of devices for tracking, tracing and visualizing relations has a further consequence. It is at least in some measure *distributed*. In an era of WikiLeaks it is important not to get caught up by hype. Nevertheless, it is nonetheless the case that much transactional data is widely, and in some cases generally, available for those with access to the internet. It is also the case that there are very large numbers of ‘apps’ available in the public domain for mining and visualizing that data. The consequence is that there are many distributed locations of socially relevant digitally derived knowledge. There are various ways of thinking about this. Some would claim that this represents a ‘democratization’ of knowledge, though we would be wary of such a large claim. At the other end of the spectrum, others would argue that this represents the erosion of properly validated knowledge of and expertise about the social.[12] We would be equally cautious about making this argument. What we would suggest, however, is that since both the *distribution* of digital devices and inscriptions is widespread, and that cascading devices work in different ways to produce different effects in different locations and circumstances, it is more readily apparent that knowledges do not cohere to generate a single authoritative representation of the social. In short, we want to suggest that social knowledge is more visibly non-coherent than it was in the recent past (though we

would need to emphasize that this does not mean that it is necessarily incoherent, which is a different and normative claim).[13]

Conclusion

We have suggested the need for a heterogeneous understanding of the digital, one that does not seek to ascribe fixed characteristics to it, but which emphasizes the contingencies by which it can be mobilized and deployed. But we also want to emphasize that digital devices and data imply a significant challenge to the social science apparatus. Where, then, in such cascades are social science methods located? What is their relative location and role within the productive, material and performative work of the digital?

We suggest that an analogy with Bourdieu's concept of field analysis will help. In this, agents are not seen to possess intrinsic qualities and capacities in and of themselves, but only with respect to other agents who are also struggling for position of advantage in a competitive field. Applied to digital devices, this suggests that they do not carry innate meanings in and of themselves, but are championed as competitors and (if we may extend the metaphor) are complementary to other devices. Overall, it is their comparative relationships with one another that define their efficacy or indispensability. Thus, for Latour (1990), it is investments in inscriptions and their mobilizations that are the sources of dominance. Rather than competition between ideas, it is competition between material devices where those that assemble and summarize can become 'centres of calculation'. But crucial to this is their mobility, transmission and circulation, and the similar movement of inscriptions. There is no room for epochs here. Instead we need to explore *fields of devices* as relational spaces where some devices survive and dominate in particular locations while others are eclipsed, at least for the moment.

In thinking about this, we have tried to argue that it does not help to imagine the digital in terms of epochal shifts or redefinitions of life. The lively and productive changes brought by the digital are no doubt large, but they need to be explored carefully, with due attention to their specificities. And, as a part of this, we have also argued that they often turn out to instantiate and reconstitute older practices, forms of stabilization and control. There are many productive devices in the representational landscape – and those that are new interact and sometimes compete with those that are older. Rather than assuming a simple teleology in which the former simply displace the latter, we have recommended a genealogical approach that is alive both to the ways in which digital devices reconfigure expertise and institutional circuits, and the ways that social agents of various kinds contest their value and efficacy. At the same time, we have argued that it is important to attend to their distinctive qualities as 'automated' devices in which data are by-products that do not require the awareness or intervention of transacting individuals or academic experts. If we are to do this well we will need to vary the magnification as we explore the chains of relations and practices enrolled in the social science apparatus.

Acknowledgements

We are grateful for the comments and suggestions of anonymous reviewers and the journal's editors.

Notes

1. While variously defined, 'big data' refers to large volumes of digital content that is generated either online or offline in social, commercial, scientific and governmental databases. But the term does not simply signify an increase in the volume but also the velocity of data collection and the increasing variety of data sources and formats. These qualities make it difficult to analyse data using traditional data management and processing applications. Thus, an additional defining characteristic is the innovation of data structures, computational capacities, and processing tools and analytics to capture, curate, store, search, trace, link, share, visualize and analyse big datasets. See, for example, discussions in boyd and Crawford (2012), Manovich (2011) and Schroeder and Meyer (2012).
2. This was well illustrated in the 2011 'riots' in England. Twitter data was analysed by the police, researchers and journalists to generate knowledge about the disturbances: for example, Manchester eResearch Centre's Twitter analysis project with *The Guardian* examined how the news of the riots spread (http://bit.ly/w3IHS6), and the Metropolitan Police trawled through Twitter and other social networking sites to gather evidence of people inciting rioting. Christakis (2012) has made a similar argument in relation to the internet: it is changing social science methods as well as its objects and subjects of analysis in different domains.
3. Though we should state that Lyotard (1979) did note the role of technoscientific transformations in cybernetics, communication theory, data storage and transmission as elements in his account of the postmodern condition.
4. See, for instance, the comments of Yoshimi (2006: 276):

> it is generally assumed that information technology alone can fundamentally alter society. The exact nature of the technology cited as the explanatory variable has changed with the times. At one time it was television; later it was the main-frame computer; then it was the computer network, and most recently, mobile media.

And similarly:

> Clearly, there is nothing 'post-' modern about information society theory. It is no more than a faithful reproduction of the principles of 'modern' industrialism adjusted to fit the 'new' conditions of information technology.

5. See: caqdas.soc.surrey.ac.uk.
6. In relation to social network analysis, Watts (2007: 489) has argued that new computational analytics of millions of network data enables the tracing of ties and social behaviour that does not rely on self-reports from participants which are full of 'cognitive biases, errors of perception and framing ambiguities'.
7. The detailed accounts of transactions collected as part of the Tesco loyalty card system, for instance, are not preserved for more than two years. However, other forms of digital data, such as certain archives and government databases, have longer durations.
8. Watts (2007: 489), for example, has argued that internet-based communication has now enabled the analysis of the 'real-time interactions of millions of people at a resolution that is sensitive to effects at the level of the individual'.
9. See, for instance, Rose: 'when I talk about empiricism a la Deleuze . . . I mean . . . an attempt to set up a constant dynamic engagement between thought and its object, and thus a concern with engaging the specificities of situations, cases and elements' (in Gane, 2004: 176).
10. The resurgence of monadology within contemporary social science has been marked by recent work on Tarde. See, for instance, Candea (2010). However, as we have just implied, it is also embedded in Deleuze's influential writing (see in particular Deleuze, 1993). Though it is often treated otherwise, actor network theory is also a form of monadology (see Latour, 1988).
11. Though there is still much boundary making, especially in government and commercial data and applications.
12. See, for example, the discussion in Savage and Burrows (2007).
13. We have phrased this carefully. Knowledges have always been different. It is the *visibility* of difference that has changed.

References

Agamben, G. (2005) *State of Exception*. Chicago: University of Chicago Press.

Amoore, L. (2009) 'Lines of sight: on the visualization of unknown futures', *Citizenship Studies* 13(1): 17–30.

Barry, A. (2005) 'Pharmaceutical matters: the invention of informed materials', *Theory, Culture & Society* 22(1): 51–69.

Bauman, Z. (1987) *Legislators and Interpreters: On Modernity, Post-modernity and Intellectuals*. London: Polity.

Beck, U. (1992) *Risk Society: Towards a New Modernity*. London: Sage.

Beer, D. (2009) 'Power through the algorithm? Participatory web cultures and the technological unconscious', *New Media & Society* 11(6): 985–1002.

Bell, D. (1976) *The Coming of Post-Industrial Society: A Venture in Social Forecasting*. New York: Basic Books.

Börner, K. (2010) *Atlas of Science: Visualizing What We Know*. Cambridge, MA: MIT Press.

boyd, d. and Crawford, K. (2012) 'Critical questions for big data', *Information, Communication & Society* 15(5): 662–679.

Braverman, H. (1974) *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York: Monthly Review Press.

Callon, M. (1986) Elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: Law J (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* London: Routledge.

Callon, M. (1998) *The Laws of the Market*. Oxford: Blackwell.

Callon, M. (2007) 'An essay on the growing contribution of economic markets to the proliferation of the social', *Theory, Culture & Society* 24(7–8): 139–163.

Candea, M. (ed.) (2010) *The Social after Gabriel Tarde: Debates and Assessments*. London: Routledge.

Castells, M. (1996) *The Rise of the Network Society: The Information Age – Economy, Society and Culture*, Vol. 1. Oxford: Blackwell.

Castells, M. (1997) *The Power of Identity: The Information Age – Economy, Society and Culture*, Vol. 2. Oxford: Blackwell.

Christakis, N.A. (2012) A new kind of social science for the 21st century. *Edge* 8 August.

Deleuze, G. (1993) *The Fold: Leibniz and the Baroque*. London: Athlone Press.

Dirks, N. (2001) *Castes of Mind: Colonialism and the Making of Modern India*. Princeton, NJ: Princeton University Press.

Featherstone, M. (2009) 'Ubiquitous media: an introduction', *Theory, Culture & Society* 26(3): 1–22.

Foucault, M. (1976) *Discipline and Punish*. London: Penguin.

Foucault, M. (1980) The confession of the flesh (1977 interview). In: Gordon C (ed.) *Power/Knowledge Selected Interviews and Other Writings*. New York: Pantheon.

Fraser, M., Kember, S. and Lury, C. (eds) (2006) *Inventive Life: Approaches to the New Vitalism*. London: Sage.

Gane, N. (2004) *The Future of Social Theory*. London: Continuum.

Giddens, A. (1991) *The Consequences of Modernity*. Stanford, CA: Stanford University Press.

Harvey, D. (1989) *The Condition of Post-modernity*. Oxford: Blackwell.

Hine, C. (2000) *Virtual Ethnography*. London: Sage.

Hine, C. (ed.) (2005) *Virtual Methods: Issues in Social Research on the Internet*. Oxford: Berg.

Hine, C. (ed.) (2006) *New Infrastructures for Knowledge Production: Understanding E-Science*. London: Idea Group Inc.

Huhtamo, E. and Parikka, J. (2011) *Media Archaeology: Approaches, Applications, and Implications*. Berkeley: University of California Press.

Kittler, F.A. (1990) *Discourse Networks 1800/1900*. Stanford, CA: Stanford University Press.

Kittler, F.A. (2006) 'Number and numeral', *Theory, Culture & Society* 23(7–8): 51–61.

Knox, H., O'Doherty, D., Vurdubakis, T. and Westrup, C. (2007) 'Transformative capacity, information technology, and the making of business "experts"', *Sociological Review* 55(1): 22–41.

Küchler, S. (2008) 'Technological materiality: beyond the dualist paradigm', *Theory, Culture & Society* 25(1): 101–120.

Lash, S. (2002) *Critique of Information*. London: Sage.

Latour, B. (1988) *Irréductions* [published with *The Pasteurisation of France*]. Cambridge, MA: Harvard University Press.

Latour, B. (1990) Drawing things together. In: Lynch M and Woolgar S (eds) *Representation in Scientific Practice*. Cambridge, MA: MIT Press.

Latour, B. (1998) Thought experiments in social science: from the social contract to virtual society. 1st Virtual Society? Annual Public Lecture, Brunel University.

Latour, B. (2010) Tarde's idea of quantification. In: Candea M (ed.) *The Social after Gabriel Tarde: Debates and Assessments*. London: Routledge.

Latour, B. and Woolgar, S. (1986) *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Law, J. (2002) *Aircraft Stories: Decentering the Object in Technoscience*. Durham, NC: Duke University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., et al. (2009) 'Computational social science', *Science* 323(5915): 721–723.

Lyotard, J.-F. (1979) *The Postmodern Condition: A Report on Knowledge*. Minnesota: University of Minnesota Press.

Mackenzie, A. (2005) 'The performativity of the code: software and cultures of circulation', *Theory, Culture & Society* 22(1): 71–92.

Mackenzie, A. and Vurdubakis, T. (2011) 'Codes and codings in crisis: signification, performativity and excess', *Theory, Culture & Society* 28(6): 3–23.

MacKenzie, D. (2008) *An Engine, Not a Camera: How Financial Models Shape Markets*. London: MIT Press.

Manovich, L. (2009) 'Cultural analytics: visualising cultural patterns in the era of "more media"', *Domus* (spring).

Manovich, L. (2011) Trending: the promises and the challenges of big social data. Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (accessed April 2013).

Marres, N. and Rogers, R. (2005) Recipe for tracing the fate of issues and their publics on the web. In: Latour B (ed.) *Making Things Public: Atmospheres of Democracy*. Cambridge, MA: MIT Press.

McNally, R.M. (2005) 'Sociomics! Using the IssueCrawler to map, monitor and engage with the global proteomics research network', *Proteomics* 5(12): 3010–3016.

Mitchell, T. (2002) *Rule of Experts: Egypt, Techno-politics, Modernity*. Berkeley: University of California Press.

Osborne, T. and Rose, N. (1999) 'Do the social sciences create phenomena? The example of public opinion research', *British Journal of Sociology* 50(3): 367–396.

Poster, M. (2001) The Information Subject. London: Routledge.

Rogers, R. (2009a) *The End of the Virtual: Digital Methods*. Amsterdam: Amsterdam University Press.

Rogers, R. (2009b) Post-demographic machines. In: Dekker A and Wolfsberger A (eds) *Walled Garden*. Amsterdam: Virtueel Platform.

Rose, N. (1991) 'Governing by numbers: figuring out democracy', *Accounting, Organizations and Society* 16(7): 673–692.

Rose, N. (2006) *The Politics of Life Itself*. Princeton, NJ: Princeton University Press.

Ruppert, E. (2007) *Producing Population*. CRESC Working Paper Series, No. 37. Available at: http://oro.open.ac.uk/9217/ (accessed April 2013).

Ruppert, E. (2009) ‘Becoming peoples: “counting heads in northern wilds”’, *Journal of Cultural Economy* 2(1–2): 11–31.
Ruppert, E. (2010) ‘Making populations: from censuses to metrics’, *Special issue of Leviathan (Berliner Zeitschrift für Sozialwissenschaft)* 35: 157–173.
Ruppert, E. (2011) ‘Population objects: interpassive subjects’, *Sociology* 45(2): 218–233.
Ruppert, E. and Savage, M. (2012) Transactional politics. In: Adkins L and Lury C (eds) *Measure and Value*. Sociological Review Monograph Series. London: Wiley-Blackwell, pp. 73–92.
Savage, M. (2010) *Identities and Social Change in Britain since 1940: The Politics of Method*. Oxford: Oxford University Press.
Savage, M. and Burrows, R. (2007) ‘The coming crisis of empirical sociology’, *Sociology* 45(5): 885–889.
Schroeder, R. and Meyer, E. (2012) Big data: what’s new? Paper presented at ‘Internet, Politics, Policy 2012: Big Data, Big Challenges?’ Oxford: Oxford Internet Institute.
Simondon, G. (1989) *L’Individuation psychique et collective. A la lumière des notions de forme, information, potentiel et metastabilité*. Paris: Editions Aubier.
Strathern, M. (2000) Abstraction and decontextualisation: an anthropological comment, Or: E for ethnography. Paper presented at the ‘Virtual Society? Get Real!’ conference, University of Cambridge.
Thacker, E. (2005) *The Global Genome: Biotechnology, Politics and Culture*. Cambridge, MA: MIT Press.
Thrift, N. (2005) *Knowing Capitalism*. London: Sage.
Toffler, A. (1980) *The Third Wave*. New York: Bantam Books.
Venn, C., Boyne, R., Phillips, J. and Bishop, R. (2007) ‘“Technics, media, teleology”: interview with Bernard Stiegler’, *Theory, Culture & Society* 24(7–8): 334–341.
Venturini, T. (2010) ‘Building on faults: how to represent controversies with digital methods’, *Public Understanding of Science* 21(7): 796–812.
Watts, D.J. (2007) ‘A twenty-first century science’, *Nature* 445: 489.
Webster, F. (1995) *Theories of the Information Society*. London: Routledge, Chapman and Hall.
Woolgar, S. (2002) *Virtual Society? Technology, Cyberbole, Reality*. Oxford: Oxford University Press.
Yoshimi, S. (2006) ‘Information’, *Theory, Culture & Society* 23(2–3): 271–288.

Author Biographies

Evelyn Ruppert is a Senior Lecturer in the Department of Sociology, Goldsmiths College, University of London.

John Law is Co-Director of the Centre for Research on Socio-cultural Change (CRESC) and Professor in the Department of Sociology at the Open University.

Mike Savage is a Professor in the Department of Sociology, London School of Economics.

Ruppert, Law and Savage have co-led at different times CRESC's integrative theme, the Social Life of Methods.

Reading Salon #4: Known Knowns (More or Less)

Moderators: Sabine Niederer and Lonneke van der Velden

Borra, Erik, and Ingmar Weber. 2012. “Political Insights: Exploring Partisanship in Web Search Queries.” First Monday 17 (7) (June 23). 9 months of Yahoo!’s US web search query logs.

boyd, D.M. & Ellison, N.B., 2008. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), pp. 210–230.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” Proceedings of the National Academy of Sciences (March 11).

Rieder, B., 2013. Studying Facebook via Data Extraction: The Netvizz Application. Proceedings of ACM Web Science 2013, Paris, May 2-4.

Rogers, Richard. 2009. “Post-demographic Machines.”

Zimmer, M., 2010. “But the data is already public”: on the ethics of research in Facebook. *Ethics and information technology*, 12(4), pp.313–325.

First Monday, Volume 17, Number 7 - 2 July 2012

f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Political Insights: Exploring partisanship in Web search queries

by Erik Borra and Ingmar Weber

Abstract

We developed Political Insights, an online searchable database of politically charged queries, which allows you to obtain topical insights into partisan concern. In this paper we demonstrate how you can discover such political queries and how to lay bare which issues are most salient to political audiences. We employ anonymized search engine queries resulting in a click on U.S. political blogs to calculate the probability that a query will land on blogs of a particular leaning. We are thus able to 'charge' queries politically and to group them along opposing partisan lines. Finally, by comparing the zip codes of users submitting these queries with election results, we find that the leaning of blogs people read correlates well with their likely voting behavior.

Contents

Introduction

Big corporations and start–ups alike have long since recognized the potential value of the data derived from monitoring and capturing online interactions for marketing and advertising purposes. Recently, scholars have called for an investment in fields such as digital humanities and computational social science, by using the kind of data available in 'big data companies' (Lazer, *et al.*, 2009; Borgman, 2009; Manovich, 2012). This paper takes up these calls, and demonstrates Political Insights, a tool for research into political partisanship, based on nine months of anonymized U.S.–based Yahoo! query logs, which have been found representative of the U.S. population (Weber and Castillo, 2010) [1].

Various studies have heralded query logs as viable alternatives or additions to traditional social science methods for gathering data, such as polls and surveys. In what follows we summarily discuss the pros and cons of employing search engine query logs to gather social and cultural data. Particularly important is the claim that query logs have limited depth because intent is hard to infer (Grimes, *et al.*, 2007). Can query logs then only provide superficial descriptions of social and cultural preferences? We briefly discuss the kind of data encountered in query logs, how they have typically been studied to infer intent, as well as their usefulness for social and cultural research. A short review is provided on how they have already been employed to measure collective preferences; this ranges from grouping queries by the users' geographical location and user demographics to measuring public attention by correlating search queries with patterns of real–world activity. We argue that as big data in general, and Web data in particular, are 'fundamentally networked' (boyd and Crawford, 2011) the consideration of additional variables and data can complement the short queries with a more elaborate description. Marres (in press) more strongly argues that loading (data–)objects with issues can turn them into placeholder objects, where matters of concern and action resonate. Could we then use query logs in combination with carefully chosen additional data in order to 'charge' queries with matters of concern?

In this paper we are specifically interested in how to locate political issues and partisan polarization. To this end, we look at all queries landing on 155 top U.S. political blogs annotated with a political leaning and subsequently assign a leaning to each query, proportional to the number of times it landed on such a blog. We review previous work on political blogs and argue why they are good proxies to infer partisan concern from the queries landing on them. We demonstrate that the resulting tool consisting of politically charged queries made by hundreds of thousands of users, allows for detailed insights into topical partisan concerns. Consider for instance queries containing [obamacare], which turned out to result much more likely in a click on right–leaning blogs, while queries containing [healthcare bill] much more likely resulted in a click on left–leaning blogs [2]. In order to ground our methodology, we validated our data set with voting polls of the 2010 U.S. mid–term elections: people visiting blogs of a particular leaning are more likely to have a zip code with a higher proportion of voters of that leaning. We conclude the paper with a short summary and provide a number of directions for future research.

Query logs as a source of data for social and cultural research

Traditionally social and cultural data are collected via field studies, user panels, focus groups, interviews, questionnaires and surveys (small or huge like the decennial U.S. census) [3]. Increasingly, the social and cultural interactions passing through or taking place on the Web are considered as valuable sources of data for social and cultural research (Lazer, *et al.*, 2009; Venturini, 2010; Rogers, in press). In this article we focus on the queries submitted to search engines, which have been among the main entry points to the Web (Dodge, 2007).

As a user who submits a search query to a search engine is necessarily motivated by some degree of interest in a particular issue, and has the willingness to invest time in it, recent literature suggests that observations based on analyzing large–scale query logs are viable alternatives, or at least additions, to some of the more traditional methods (Grimes, *et al.*, 2007; Richardson, 2008; Granka, 2009, 2010; Gruszczynski, 2011; Mohebbi, *et al.*, 2011; Ripberger, 2011; Scharkow and Vogelgesang, 2011; Scheitle, 2011; Weber and Jaimes, 2011). Query logs are recommended for their coverage (*e.g.*, the entire U.S.), ease of collection, scope (any entry into a search engine), cost (analyzing text is cheap), and up–to–datedness (almost in real–time). Additionally, studying queries is less prone to the observer effects present in other types of social data collection (Webb, *et al.*, 1972), nor are particular response categories imposed. Query logs seem particularly attractive when alternative data sources are very expensive or not electronically available at all.

These surveys point out that query logs as sources of social and cultural data present difficulties, too. The data are often considered 'noisy' or 'messy' (*e.g.*, because of misspellings, spammers, or a small set of highly biased users), they need to be anonymized (with the risk of tainting, according to some authors), and they are not freely available (query logs generally require commercially negotiated access). Moreover, data must be validated or grounded so that the claims based on Web data in general, and query log data specifically, can be trusted (Thelwall, *et al.*, 2005; Rogers, in press). The biggest difficulty, however, seems to be that as intent is hard to infer, query logs have limited depth.

Advancing such work, this paper introduces the use of query logs to provide insight into partisan concern. We discuss seminal examples of research with query logs, focusing in particular on those which grouped and combined query logs with other data in order to infer collective preference and opinion. Subsequently, charging queries (politically) is proposed as a specific methodology to infer political partisanship. First, we address how search engines have sought to understand the user, and what information can be found in a query log.

Inferring (collective) preference from query logs

In order to learn from what the user does and wants, search engines will typically keep track of what their users search for and the result they click on. The kind of information collected prompted Battelle (2006) to depict search engines as 'databases of intentions' as they store massive amounts of 'desires, needs, wants, and preferences' [4]. However, queries are typically short (two or three terms) (Jansen and Spink, 2006) and often ambiguous: if a query reads [washington] it is not clear whether the city, newspaper, president or actor is meant. In order to provide the user with the best results, search engines use a variety of techniques to infer user intent. In this respect, information science literature usually distinguishes between three types of queries: navigational (to reach a specific Web site), informational (to find more information about a subject) and transactional (to perform some Web mediated activity) (Brenes, *et al.*, 2009). Increasingly, the user will be offered localized and personalized results too. For example, when a user searches for [restaurant], most likely one close by will be preferred. As location can be, approximately, inferred on the basis of a user's IP address or profile information, search engines offer local domain versions in order to more precisely determine the scope of results returned (Goldsmith and Wu, 2006) [5]. Except for location, a user's past queries turn out to be important in determining a query's intent (Grimes, *et al.*, 2007), hence the push to personalize search results. Additionally your friends' preferences, as expressed through their respective search histories, can be taken into account for the personalization of results (Feuz, *et al.*, 2011).

Except to improve a user's result rankings, query logs have also been used to study the distribution of specific cultural and social preferences, by employing three variables: query terms (including volume), where the queries were made (location), and the queries' date stamps (Spink, *et al.*, 2009; Rogers, in press). By introducing the searcher's profile information (age, gender, zip code) combined with U.S. census data, one can not only track changes in the distribution of queries along geographic regions but also along demographic dimensions (Weber and Castillo, 2010; Weber and Jaimes, 2010). If we consider that search queries are valid indicators which can be employed for social and cultural research, search engine query histories might thus be used to provide the time and place, as well as the intensity of social and cultural preferences.

Most recently, Seth, *et al.* (2011) examined the full log of one month of queries submitted to the U.S. version of Google's search engine. They grouped the queries on a city–level, based on the user's IP address, and calculated an excess score to focus on those queries which occur either more or less than expected (*e.g.*, in each city [facebook] will be a very frequent query, but this is probably not the most interesting feature to characterize a city by). Based on disparities in query volume they calculated a city–similarity and compared it with a ground truth of city similarity based on census data. They found that 'query logs can be a good representation of the interests of the city's inhabitants and a useful characterization of the city itself' [6]. Weber and Jaimes (2011) came to similar findings based on Yahoo!'s query logs, which they exemplified by noting that the fraction of searches related to actors is about three times higher in the L.A. area, which includes Hollywood, than in any other region considered. Similarly, the fraction of queries related to gambling is highest in Las Vegas and lowest in Salt Lake City. Linking users' zip codes to U.S. census data, Weber and Castillo (2010) found that the Yahoo! query logs provide a good demographic description of the U.S. population and that different segments of the population differ in the topics they search for as well as in their search behavior (see also Weber and Jaimes, 2011).

Various authors recently have sought to appropriate trends of query volume as measures of public attention. Such research found that the search volume of specific (politics and issue) queries often correlates with fluctuations in news coverage (Weeks and Southwell, 2010; Granka 2009, 2010; Ripberger, 2011) and to a certain extent also with polls and surveys (Granka, 2009; Scheitle, 2011).

In a similar vein, other projects sought to match queries, considered as expressions of public interest and concern, to external data. One of the better–known projects, Google Flu Trends, asked whether the frequency of specific search queries (out of the 50 million most recurring U.S. queries) could be used as an indicator of regionally specific seasonal outbreaks of influenza. The project tried to match specific queries to Google with the U.S. Centers for Disease Control and Prevention's historical data on influenza outbreaks. The project concluded that very specific and frequent influenza–related queries can provide topical and geographically precise indicators of such an outbreak (Ginsberg, *et al.*, 2008).

Other work demonstrated that the trends in volume of specific search queries correlates surprisingly well with consumer activities as expressed in economic indicators like retail, automotive, and home sales, travel statistics, and unemployment indicators (Choi and Varian, 2009; Varian and Choi, 2009). Although search terms were found to provide valuable indicators of off–line phenomena, their predictive value often does not exceed simple baseline models (Goel, *et al.*, 2010). Similarly, while query volume of candidate names may reflect topical popularity, query volume is less likely to predict who wins the next election (Lui, *et al.*, 2011).

In many of these projects prior knowledge often influences the choice of queries, so as to match an external baseline. In May 2011 Google released a tool reversing this methodology. Instead of matching specific query trends to external data, on the basis of a pattern of some real world activity submitted, the tool automatically 'surfaces queries which correspond with [that] particular pattern of activity' [7].

In this research, we similarly regard queries as expressions of public interest. However, we do not attempt to match queries to some pattern of off–line activity but take inspiration from research pursued on the high–traffic recipe site allrecipes.com (http://allrecipes.com/). The researchers analyzed the queries and locations of over 750,000 users which searched for a recipe on the site, prior to Thanksgiving 2009 (the U.S. holiday feast), and found that 'regional differences [in taste] and the precise time [of a user's interest] could be pinpointed as never before' (Severson, 2009). The U.S. east coast, for instance, was more interested in recipes on 'sweet potato casserole' and the South and Middle more in 'pecan pie.'

Enriching query logs with other data has made them more useful for social and cultural research. Here, at first, we are not interested in correlating queries with data such as location and demographics. We take advantage of the relations within query logs and look at the queries leading to a click on a specific group of sites: political blogs. Just as allrecipes.com was used as a proxy to measure differences in taste, we have used political blogs as proxies of political intent to charge queries politically.

Political Insights

This study attempts to detect political issues and concerns by looking into the collective search histories of users querying the U.S. version of the Yahoo! Web search engine, a data set extracted from nine months of anonymized Yahoo! search query logs, from May 2010 to January 2011 [8].

Blogs as proxies of political concern

Political blogs were chosen as our proxy for gathering queries with political intent as in U.S. politics they are an important source of political commentary. Amongst others, political blogs were studied in terms of link structure and content (Adamic and Glance, 2005; Hargittai, *et al.*, 2008; Kelly, 2010), author demographics and reachability (Hindman, 2008), technological adoption and use (Benkler and Shaw, 2010), as well as readership (Lawrence, *et al.*, 2010). Although initially political blogs where hoped to be vehicles to increase political deliberation, all these studies found these blogs to be polarized along opposing partisan lines [9]. Consequently, we hypothesize that in grouping the queries by the leaning of the blogs on which they land, a meaningful description of partisan concern is provided.

The 155 political blogs for which we gathered the queries landing on them, were listed by Benkler and Shaw (2010) who triangulated seven lists of top blogs, manually coding them as leaning towards the political spectrum's left, center, or right [10]. We considered other sites like those of election candidates, but queries to these sites turned out to be mostly navigational, showing interest in the candidate but not in the candidate's issues. Nor did we use the U.S. House of Representatives' or Congress' sites, not wanting to restrict politics to (official) government information. News sites were rejected as well, as they cover much more than politics alone.

Filtering the query log and retaining only those queries resulting in a click on a predefined set of political blogs' URLs, is based on the assumption that if a specific URL for a particular query is clicked, there is a relation between the two. This relation depends on three elements: the user submitting the query, the URL with content relevant to the query, and the search engine providing a ranked list of results relevant to the query (generally the results contain all the query's words). A search engine will typically return a variety of different (types of) sites, ranging from Wikipedia articles, videos, or news related to the query, to sites run by businesses, NGOs, individuals and authorities. By clicking a certain URL for the query submitted, the user thus not only shows interest in a specific URL but also in a particular type of site [11]. Although in our research the user thus reinforces the political relevance of a query, the focus is not on users but on how queries can be enriched by considering the types of site clicked.

Leveraging search engine results to enrich queries is similar to Goel, *et al.* (2010) who categorized queries as movie or game–related if such a site's URL appeared on the first page of search results. In this study we consider a query to be political if after submitting the query a political blog was clicked. As noted above, we drew inspiration from the work on allrecipes.com, which showed that queries to a specific type of site can provide (regional) characteristics of taste. In this article, however, at first we are not interested in the regional characteristics of queries but whether topical political sites can charge queries politically.

To our knowledge, nobody has yet studied queries landing on political blogs. Mishne and de Rijke (2006) investigated general characteristics of queries submitted to blog search engines. Hindman (2008) compares closest to our study by looking at queries landing on political sites. While Hindman only considered the top twenty queries of one month, we considered all queries landing on political blogs over nine months.

Politically charged queries

We filtered the query log retaining only those queries resulting in a click on the URLs of a predefined set of political blogs [12]. As these blogs were attributed a political leaning too, we could not only politically charge a query, but also determine its partisanship by attributing the query with a value for each leaning, proportional to the number of times the query landed on a blog of that leaning.

Several additional steps ensured that the queries are indeed politically relevant. After aggregating all queries landing on the described political blogs we removed all queries containing personally identifiable information such as credit card numbers, infrequent personal names, social security numbers, or street addresses. In the resulting set, many of the queries landing on political blogs turned out to be navigational (and thus hardly indicative of partisan concern). To filter out these navigational queries we used two complementary techniques. First we looked at the click entropy for each query to find out whether a diverse set of sites was clicked for a particular query. Queries with more than two occurrences but landing mostly on the same site (with an entropy not larger than 1.0), were considered navigational. Additionally, through the use of simple heuristics we tested whether there was a close match between the query and the clicked domain. We first tokenized queries and URLs (based on dots and spaces), stemmed plurals, and alphabetized the words. Subsequently, a query–URL pair is considered navigational if it contains a domain component such as 'www' or '.com,' the domain of the URL is contained in the query (or vice versa), or when the edit distance between queries and the domain is smaller than 2 (for queries with more than four characters). For example, [drudge], [drudge report], and [drudgerport] landing on http://www.drudgereport.com are all considered navigational queries.

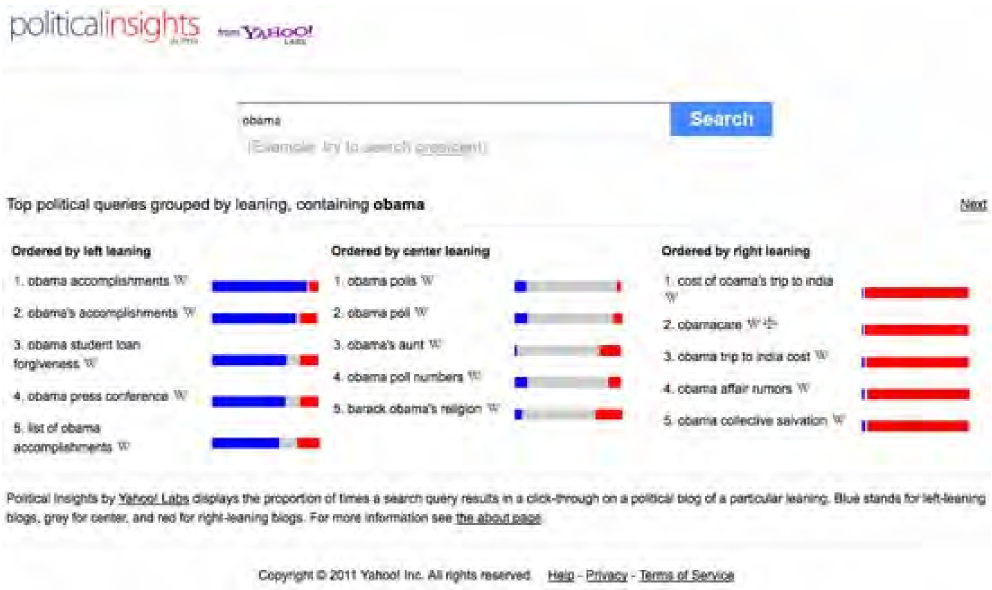
To ascertain that our queries had a minimum shared uptake and relevance, we filtered the data to only retain queries resulting in a click–through to at least three political blogs. To prevent one blog from setting the agenda we also removed queries with a very high query volume but resulting in click–throughs to very few political blogs.

Not all queries are equally frequent, and some might lead to, say, three clicks corresponding to a particular leaning. To address the corresponding sparsity issue and to avoid prematurely marking this query as 'strongly partisan,' we applied Bayesian smoothing which, in practice, means that we evenly distributed a small number of artificial clicks over all leanings, before accounting for the actually incurred clicks. Moreover, certain blogs in our list, *e.g.*, *Huffington Post*, attracted far more traffic than others; in turn making the left attract considerably more click volume than the right. As this potentially tainted the analysis and created a systematic bias towards the left, we normalized each leaning's total click counts by attributing the same total weight to the left, center and right. This might, however, be overly enthusiastic as the Web, for example, might overall be more left–leaning.

Political Insights: A gauge of partisanship

Previously we ascertained that the queries in our dataset are politically relevant; we politically 'charged' each query and assigned partisanship by the fraction of times it landed on a blog with a particular political leaning. We provide a searchable database of such politically charged queries at http://politicalinsights.sandbox.yahoo.com, ranking queries according to their assigned proportion of a particular leaning, *i.e.*, left, center, and right side of the political spectrum. The landing page of Political Insights displays a global ranking based on nine months' data. In fact we built a *political partisanship machine* by sifting out those queries most strongly linked to a particular political ideology.

Our system also allows the user to search for specific queries containing a particular word or phrase. In case of a match, all queries containing the search will be shown and ranked. See Figure 1: when searching for [obama] you will see queries like [obama accomplishments] to be more on the left side, and [obamacare] to be more on the right side of the political spectrum [13].



Top political queries grouped by leaning, containing **obama**

Ordered by left leaning

1. obama accomplishments W
2. obama's accomplishments W
3. obama student loan forgiveness W
4. obama press conference W
5. list of obama accomplishments W

Ordered by center leaning

1. obama polls W
2. obama poll W
3. obama's aunt W
4. obama poll numbers W
5. barack obama's religion W

Ordered by right leaning

1. cost of obama's trip to india W
2. obamacare W ⚖
3. obama trip to india cost W
4. obama affair rumors W
5. obama collective salvation W

Political Insights by Yahoo! Labs displays the proportion of times a search query results in a click-through on a political blog of a particular leaning. Blue stands for left-leaning blogs, grey for center, and red for right-leaning blogs. For more information see the about page.

Copyright © 2011 Yahoo! Inc. All rights reserved. Help - Privacy - Terms of Service

Figure 1: Screenshot of Political Insights. Top results for searches containing [obama]. Available online at http://politicalinsights.sandbox.yahoo.com/index.php?q=obama, accessed 1 July 2012.

In order for the user of our system to get an illustration of the relationship between a query and its politics, we provide the following. Clicking the query itself opens a new window containing its current search results restricted to the blogs of a particular leaning. Furthermore, we mapped all queries to the most relevant Wikipedia articles; by clicking the 'W' next to the query this article is shown, together with the article's categories. Finally, as highly partisan queries might be the result of an effort to introduce slant or spin, we tried to link queries to external 'fact–checking' sites, *i.e.*, http://factcheck.org, http://politifact.com, and http://snopes.com. For instance, when searching for [obama] and clicking on the scales symbol next to the query [obamacare], three results from politifact.com will be shown. For example, Mitt Romney, candidate for the 2012 Republican Party presidential nomination, is found to make a false allegation stating that 'Repealing the health care law would save $95 billion in 2016' [14].

Table 1 shows the top result rankings per leaning for exemplary queries containing particular politicians, issues and stances. As can be seen, in general the results are intuitively correctly aligned. By not only politically charging queries but also charting them along oppositional partisan lines, we actually shifted the notion from *actor partisanship* to *query partisanship*, arguably opening up new ways of research into framing (Entman, 1993). While we made sure to remove navigational queries (*e.g.*, the names of individual blogs), highly partisan queries might be termed navigational as well, as they predominantly lead to blogs of one particular leaning.

Table 1: Examples of queries ranked by leaning. For clarity's sake, we only included queries which were attributed more than 50 percent of one respective leaning. All examples are available on http://politicalinsights.sandbox.yahoo.com (accessed 11 January 2012).
Note: * This query is included because the system allows partial matches.

Query	Left	Right
[obama] (politician)	[obama accomplishments] [obama student loan forgiveness] [obama press conference]	[cost of obama's trip to india] [obamacare] [obama affair rumours]
[bush] (politician)	[bush deficit] [george w. bush] [president george w bush costs of trip to crawford texas]	[jobs created under bush] [bush vs obama vacation days] [obama extending bush tax cuts]
[lies] (stance)	[glenn beck lies] [fox news lies] [list of republican lies]	[inconvenient truth lies] [lies about obama] [racist signs at tea party rallies]*
[violence] (issue)	[tea party violence] [mexico violence] [right wing violence]	[left wing violence] [liberal violence]
[immigration] (issue)	[immigration] [immigration reform 2010] [arizona immigration news]	[hanson moral implications of illegal immigration] [mexico immigration laws] [az immigrations news]
[gun] (issue)	[gun control] [ergun caner]* [chicago gun ban]	[is the government trying to take away our guns] [eric holder guns] [mexico false claims of us guns causing crime]
[job] (issue)	[jobs bill] [take our jobs]	[jobs created under bush] [1961 bill to send jobs overseas] [epa regulations to cost nearly a million jobs]

In what follows we look at whether our approach using blogs as proxies of political intent resulted in data which can be grounded in voter demographics and to what extent the data are representative of 'off–line' political preference.

Grounding the data

While the differences are insightful, the question remains whether charging queries politically has any relation with the 'off–line.' To what extent do users submitting political queries represent the U.S. (voting) population? We look at both voter demographics and voting preferences.

(Voter) demographics

By combining the user–provided zip code with U.S. census information we investigated whether gender, age, race and educational level were representative for the U.S. population [15]. Using the U.S. 2000 census data on, where appropriate, a per–zip code level we found that (i) our users were predominantly male (54.7 percent vs. 49.1 percent in the census), (ii) older (median age of 45 vs. 35 in the census), (iii) more white (78.4 percent vs. 75.1 percent in the census) and (iv) more highly educated (27.8 percent vs. 24.4 percent in the census — fraction of population of 25 years and older with at least a B.A. degree) [16].

Comparing the same census data with the 2010 voting records for registered voters we observed that (i) the gender bias was even more pronounced (54.7 percent vs. 46.6 percent), (ii) our users were slightly younger (median age of 45 vs. approximately 47), (iii) not white enough (78.4 percent vs. 83.4 percent) and (iv) less educated (27.8 percent vs. 32.1 percent with at least a B.A. degree) [17]. However, some caution is appropriate. First, the available census data date back from 2000, the voting records from 2010. So the U.S. population has aged since then, probably eliminating our observed age gap, has become less white, further increasing our observed racial gap, and more highly educated, reducing the educational gap. Concerning the latter, the voting records include people of age 18 and above, while the census definition for educational attainment in this category considers only ages 25 and higher. This would even increase the actual gap. However, people holding a B.A. degree or higher, regardless of the characteristics of their zip code, are more than 10 percent more likely to register to vote than an average citizen. This most likely explains the observed educational difference and we do *not* believe that the users in our sample have a lower educational attainment than the average voter.

2010 U.S. midterm elections

To ascertain whether the online notion of 'left (or right) leaning blog' is linked to the off–line notion of 'voting Democrat (or Republican)' we used per–zip results for the 2010 U.S. House of Representatives elections. We computed the probability that a person who clicked on left blogs ('left–clicking') voted Democrat in the 2010 U.S. elections. To estimate this probability we used the election results for the zip code from the user's profile and assumed that the user was drawn uniformly at random from the voting population. Averaging these over all left–clicking users and all zip codes led to the following equation.

Equation 1: $$\frac{\Sigma_z c_z^l \cdot v_z^D}{\Sigma_z c_z^l}$$

Here c_z^l is the count of left–clicking users in zip code *z*. v_z^D is the fraction of voters voting Democrat in zip code *z*. In a similar manner we can define c_z^r for right–clicking users and v_z^R for Republican voting fractions.

We can thus estimate the probability that a left–clicking user voted Democrat or that a right–clicking user voted Republican. The value above was multiplied by 100 so that the probability estimate lies between 0 percent and 100 percent.

If each zip code voted either 100 percent Democrat or 100 percent Republican, the estimate could theoretically attain 100 percent. However, if each zip code was split 50–50 the maximum would be 50 percent. In fact, across the entire U.S. the Democrat–Republican split was 44.8–51.4 and for zip codes with users clicking on the considered blogs this split was 45.1–49.3, where zip codes were weighted by the number of users.

Given this roughly equal fraction of Democrat and Republican votes, we computed a more realistic bound for our probability estimates. We replaced the number of left–clicking users in a given zip code in Equation 1 by the total number of users in the zip code multiplied by the fraction voting Democrat. This bound corresponds to the case where our assumptions are fully correct and 'left–clicking' equals 'voting Democrat.' This leads to the following equation.

Equation 2: $$\frac{\Sigma_z c_z^l v_z^D \cdot v_z^D}{\Sigma_z c_z^l v_z^D}$$

Again, similar bounds were obtained for Republicans.

If all users clicking at least one political blog are taken into account for this estimate we get Table 2. The upper bounds for this case are 53.1 for left–clicking and Democrat and 56.6 for right–clicking and Republican. Although Table 2 indicates that the trends go into the right direction, *i.e.*, left–clicking users are more likely to vote Democrat, the difference with the upper bound is still considerable. In an attempt to reduce this gap, we experimented with two ideas. First, we hypothesized a temporal dimension, implying that the match between clicking and voting behavior improved closer to the actual election date. However, overlapping intervals of three months did not reveal any temporal dynamics. Second we hypothesized that users clicking more frequently on a blog of a particular leaning are better indicators for the voting behavior in the corresponding zip code. To test this, we used the top 1,000 users in terms of numbers of clicks for each of the three leanings considered [18]. The results are presented in Table 3. For this set of users the upper bound for left–clicking and Democrat was 54.2 and for right–clicking and Republican 56.2. All the pair–wise differences in Table 3 were found to be significant at a level of 1 percent, using a t–test where each user and the voting estimate of the zip code corresponded to one data point.

Table 2: Estimated voting probability of all users clicking right, center, or left leaning blogs.

Clicked	**Estimated voting probability**	
	Democrat	Republican
Right	43.3	50.8
Center	44.9	49.6
Left	45.5	49.0

Table 3: Estimated voting probability of the top 1,000 users for each leaning.

Clicked	**Estimated voting probability**	
	Democrat	Republican
Right	43.2	51.2
Center	45.9	48.3
Left	49.1	46.4

Overall, our observations indicate that the leaning of the blogs a person clicks on in response to Web search queries correlates with the voting behavior of the area where the person resides. This correlation is stronger for users who repeatedly click on blogs of a particular leaning. The fact that we did not quite attain the bounds for a perfect fit of our model (49.1 < 54.2 and 51.2 < 56.2) can be explained in a number of ways. Not all blogs focus purely on politics; they contain different content as well. This is particularly valid for the *Huffington Post*, which also covers celebrity news; it demonstrates how important good source (proxy) selection is. Another explanation is technical in nature. The election results were retrieved from *USA Today*, which displayed the results per district ID instead of zip code, requiring a not always unambiguous conversion: zip codes, through redistricting, may belong to different legislative districts at different times, in function of population changes [19]. Our information came from the ZCTA to district mapping from the 110th Congress (applicable from 2007 to 2009) [20]. Thus, matching ZCTA to zip codes and using the 110th instead of the 112th legislative district mapping might have introduced errors.

Other explanations include the possibility that the voting and blog–clicking populations are not identical. This could hold on a nation–wide level where, say, older people are less likely to use the Internet or on a per–zip level where voters in a clearly defined state or region do not consult blogs to shape their voting choice.

Summarizing, we find that compared to the average voter our users have about the right age, are predominantly male, not white enough, and have about the right educational background. In addition, we verified that people clicking blogs of a specific leaning are more likely to live in a zip code with a higher proportion of voters with that leaning. This finding is in line with the survey about political blog readership by Lawrence, *et al.* who find that "blog readers gravitate towards blogs that accord with their political beliefs" [21]. It might be argued that politically charged queries disclose partisan concern, as the likely voting behavior of users submitting those queries correlates with the leaning of the blogs they click.

Conclusion and future work

Query logs have been heralded as an addition, or even alternative, to traditional social science data because they are unprecedented in scope, scale and detail, and the queries are obtained from within their natural environment. However, queries being short they are often hard to interpret; they seem to have no depth and little associated context. In this paper the careful selection of topical sites, clicked in response to a query, is presented as a proxy with which queries with shared concern can be discovered. The recognition that partisan political blogs can be used as such a proxy to detect political concern, allowed us to charge queries politically and attribute partisanship. This in turn allows us to sift out highly partisan queries to provide detailed insights into political concerns. Subsequently, we found that the leaning of the blogs people read correlate with their likely voting behavior.

The Political Insights tool is based on static data dating from around the 2010 U.S. midterm elections. Ranking the queries according to partisanship made the tool into a *gauge* of query partisanship. In other work we have extended upon this core methodology by using more fresh data and tracking changes over time, permitting us to consider trending queries which can then be ranked by partisanship (Weber, *et al.*, 2012). The resulting *barometer* of political partisanship allows answering questions such as: 'what is trending among the political left?' Additionally, we countered noise and misspellings by grouping similar queries on their stemmed and normalized form. A more extensive use of fact–checking sites, linking the truth–value of queries back to leanings, allowed us to investigate for instance which leaning has the highest query volume in relation to false allegations. We are also considering various other extensions of our application. The 'search the left' or 'search the right' functionality, as in the section describing our application, could be an interesting service in itself, juxtaposing the two leanings' results and queries extracted thereof. Instead of providing a national outlook we could also zoom in on a smaller geographical level. Last but not least, we intend to test our hypothesis that charging queries with shared matters of concern and additionally ranking them by opposing partisanship can be employed to show partisanship in other domains too, such as climate change alarmists versus skeptics.

After Weber and Castillo's work (2010), Yahoo! Clues was made public, a search analysis service allowing 'you to instantly discover what's popular to a select group of searchers — by age or gender [or location] — over the past day, week or even over the past year' (Theodore, 2011). Similarly, we released Political Insights hoping that it will be useful for (re–)searchers. We believe that such online tools offer researchers new horizons to sociological research by simultaneously allowing access to the individual component (queries), as well as the aggregated structure (demographic breakdown and political partisanship respectively).

Whilst our tool cannot predict who will win the next U.S. Presidential elections, it describes which issues resonate most with different sides of the political spectrum and provides insight into political partisanship by placing side by side competing claims. The increasing polarization of (political) discourse, combined with online recommendation cultures suggesting information based on what like–minded have done before, led to warnings for echo chamber effects (Sunstein, 2006) and filter bubbles (Pariser, 2011). We believe that it is beneficial to make these effects insightful. As Lippmann so eloquently stated:

> The individual not directly concerned may still choose to join the self–interested group and support its cause. But at least he will know that he has made himself a partisan, and thus perhaps he may be somewhat less likely to mistake a party's purpose for the aim of mankind. [22]

About the authors

Erik Borra is a Ph.D. candidate and lecturer at the University of Amsterdam's Media Studies department. He is also lead developer for the Digital Methods Initiative, the Ph.D. research program in New Media at the University of Amsterdam.
E–mail: borra [at] uva [dot] nl

Ingmar Weber is a researcher at Yahoo! Research Barcelona. He works on query log analysis, often with a demographic angle, and on large–scale information extraction from Web content.
E–mail: ingmar [at] yahoo-inc [dot] com

Acknowledgements

We would like to thank Venkata Rama Kiran Garimella for the implementation of the fact–checking functionality, and Richard Rogers and Esther Weltevrede for their insightful comments.

Notes

[1] In our study, all queries were anonymized by removing personally identifiable information such as telephone numbers, street addresses, social security numbers or infrequent personal names. Yahoo! user names pertaining to queries were replaced by random numbers. All of our data analysis is done in aggregate, without tracking individual users. As of 1 July 2012 Political Insights is still available at http://politicalinsights.sandbox.yahoo.com. However, as the development of this tool has quickly progressed, a more advanced version using the exact same methodology but including search trends is readily available at http://politicalsearchtrends.sandbox.yahoo.com (accessed 1 July 2012). This paper introduces the core methodology on which both tools are built.

[2] Whenever we write about a specific query we encapsulate it in brackets to delineate it.

[3] Webb, *et al.*, 1972, pp. 1–2; Ferguson, 2000, pp. 21–30.

[4] Batelle, 2006, p. 6.

[5] Search engines often provide additional services like e–mail for which a user is obliged to fill in a profile. If a user is logged into such a service, logs can be complemented with information, such as a zip code, contained in the profile.

[6] Seth, *et al.*, 2011, p. 1.

[7] Mohebbi, *et al.*, 2011, p. 2.

[8] Note that this period includes the November 2010 U.S. midterm elections.

[9] In this study we do not look for the causes or effects of partisanship but take it as a given. For an insightful discussion on the political effects of media choice, see Prior (2007).

[10] A 2005 poll of 2,209 U.S. citizens indicates that the labels left and right are generally associated with liberal and conservative respectively (PR Newswire, 2005). We, and our colleagues very familiar with U.S. politics, verified the importance and coding of the blogs.

[11] In the studied period results were not personalized; all users were shown the same results.

[12] The procedural subsection is also available at http://erikborra.net/blog/2012/04/methods-for-exploring-partisan-search-queries (23 April 2012), accessed 20 May 2012.

[13] Note that similar queries might be displayed in the ranked list of one leaning, as in the current version of the application queries are not grouped, but left untouched.

[14] http://www.politifact.com/truth-o-meter/statements/2011/nov/04/mitt-romney/mitt-romney-said-repealing-obamacare-would-save-95/, accessed 15 January 2012.

[15] To discover relevant queries per leaning we used all available query–click pairs, whereas here we look at those with an associated user profile.

[16] http://factfinder.census.gov/home/saff/main.html?_lang=en, accessed 15 December 2011.

[17] Page 4 on http://www.census.gov/prod/2010pubs/p20-562.pdf lists voter demographics. The exact median age is not reported and must be inferred from results reported for age buckets, accessed 10 December 2011.

[18] For all other parts of our analysis many more than 1,000 users contributed.

[19] The election results were scraped from http://projects.usatoday.com/news/politics/2010/elections/, accessed 15 December 2011.

[20] http://www.census.gov/geo/www/cd110th/natl_code/zcta_cd110_natl.txt, accessed 15 December 2011.

[21] Lawrence, *et al.*, 2010, p. 141.

[22] Lippmann, 1993, p. 104.

References

L.A. Adamic and N. Glance, 2005. "The political blogosphere and the 2004 U.S. election: Divided they blog," *LinkKDD '05: Proceedings of the Third International Workshop on Link Discovery*, pp. 36–43.

J. Battelle, 2006. *The search: How Google and its rivals rewrote the rules of business and transformed our culture*. New York: Portfolio.

Y. Benkler and A. Shaw, 2010. "A tale of two blogospheres: Discursive practices on the left and right," Harvard University Berkman Center for Internet and Society, Research Publication, number 2010–6 and Harvard Public Law Working Paper, number 10–33, at http://cyber.law.harvard.edu/publications/2010/Tale_Two_Blogospheres_Discursive_Practices_Left_Right, accessed 23 June 2012.

C.L. Borgman, 2009. "The digital future is now: A call to action for the humanities," *Digital Humanities Quarterly*, volume 3, number 4, at http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html, accessed 6 February 2010.

d. boyd and K. Crawford, 2011. "Six provocations for big data," *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society* (September), at http://ssrn.com/abstract=1926431, accessed 14 January 2012.

D.J. Brenes, D. Gayo–Avello and K. Pérez–González, 2009. "Survey and evaluation of query intent detection methods," *WSCD '09: Proceedings of the 2009 Workshop on Web Search Click Data*, pp. 1–7.

H. Choi and H. Varian, 2009. "Predicting initial claims for unemployment benefits," at http://research.google.com/archive/papers/initialclaimsUS.pdf, accessed 13 December 2011.

D. Dodge, 2007. "Search engines are the Start page for the Internet," *Don Dodge on The Next Big Thing* (13 December), at http://dondodge.typepad.com/the_next_big_thing/2007/12/search-engines.html, accessed 14 January 2012.

R.M. Entman, 1993. "Framing: Toward clarification of a fractured paradigm," *Journal of Communication*, volume 43, number 4, pp. 51–58.

S.D. Ferguson, 2000. *Researching the public opinion environment: Theories and methods*. London: Sage.

M. Feuz, M. Fuller and F. Stalder, 2011. "Personal Web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalization," *First Monday*, volume 16, number 2, at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3344/2766, accessed 23 June 2012.

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski and L. Brilliant, 2008. "Detecting influenza epidemics using search engine query data," *Nature*, volume 457, number 7232, pp. 1,012–1,014.

S. Goel, J.M. Hofman, S. Lahaie, D.M. Pennock and D.J. Watts, 2010. "Predicting consumer behavior with Web search," *Proceedings of the National Academy of Sciences of the United States of America*, volume 107, number 41, pp. 17,486-17,490.

J.L. Goldsmith and T. Wu, 2006. *Who controls the Internet? Illusions of a borderless world*. New York: Oxford University Press.

L.A. Granka, 2010. "Measuring agenda setting with online search traffic: Influences of online and traditional media," *Annual Meeting of the American Political Science Association*, at http://ssrn.com/abstract=1658172, accessed 23 June 2012.

L.A. Granka, 2009. "Inferring the public agenda from implicit query data," *SIGIR '09: Understanding the User–Logging and Interpreting User Interactions in Information Search and Retrieval*, at http://laura.granka.com/publications/granka_SIGIR09paper.pdf, accessed 23 June 2012.

C. Grimes, D. Tang and D.M. Russell, 2007. "Query logs alone are not enough," *WWW 2007: Workshop on Query Log Analysis*, at http://www2007.org/workshop-W6.php, accessed 23 June 2012.

M. Gruszczynski, 2011. "Examining the role of affective language in predicting the agenda–setting effect," *APSA 2011 Annual Meeting Paper*, at http://ssrn.com/abstract=1902270, accessed 23 June 2012.

E. Hargittai, J. Gallo and M. Kane, 2008. "Cross–ideological discussions among conservative and liberal bloggers," *Public Choice*, volume 134, numbers 1–2, pp. 67–86.

M. Hindman, 2008. *The myth of digital democracy*. Princeton, N.J.: Princeton University Press.

B.J. Jansen and A. Spink, 2006. "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing & Management*, volume 42, number 1, pp. 248–263.

J. Kelly, 2010. "Parsing the online ecosystem: Journalism, media, and the blogosphere," In: G. Einav (editor). *Transitioned media: A turning point into the digital realm*. New York: Springer, pp. 93–108, at http://www.springerlink.com/content/q836x0472284076k/, accessed 4 January 2012.

E. Lawrence, J. Sides and H. Farrell, 2010. "Self–segregation or deliberation? Blog readership, participation, and polarization in American politics," *Perspectives on Politics*, volume 8, number 1, pp. 141–157.

D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.–L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy and M. Van Alstyne, 2009. "Life in the network: The coming age of computational social science," *Science*, volume 323, number 5915, pp. 721–723.

W. Lippmann, 1993. *The phantom public*. With a new introduction by W.M. McClay. London: Transaction Publishers.

C. Lui, P.T. Metaxas and E. Mustafaraj, 2011. "On the predictability of the U.S. elections through search volume activity," *Proceedings of the IADIS International Conference on e–Society*, at http://cs.wellesley.edu/~pmetaxas/e-Society-2011-GTrends-Predictions.pdf, accessed 23 June 2012.

L. Manovich, 2012. "Trending: The promises and the challenges of big social data," In M.K. Gold (editor). *Debates in the digital humanities*. Minneapolis: University of Minnesota Press, pp. 460–475.

N. Marres, in press. "The environmental teapot and other loaded household objects: Re–connecting the politics of technology, issues and things," In: P. Harvey, E. Casella, G. Evans, H. Knox, C. McLean, E. Silva, N. Thoburn and K. Woodward (editors). *Objects and materials: A Routledge companion*. London: Routledge.

G. Mishne and M. de Rijke, 2006. "A study of blog search," In: M. Lalmas, A. MacFarlane, S.M. Rüger, A. Tombros, T. Tsikrika and A. Yavlinsky (editors). *Advances in information retrieval, 28th European Conference on IR Research, ECIR 2006, London, U.K., April 10–12, 2006, Proceedings. Lecture Notes in Computer Science*, number 3936, pp. 289–301.

M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi and S. Kumar, 2011. "Google Correlate Whitepaper" (9 June), at https://www.google.com/trends/correlate/whitepaper.pdf, accessed 13 December 2011.

E. Pariser, 2011. *The filter bubble: What the Internet is hiding from you*. London: Penguin Press.

M. Prior, 2007. *Post–broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. New York: Cambridge University Press.

PR Newswire, 2005. "Political labels: Majorities of U.S. adults have a sense of what conservative, liberal, right wing or left wing means, but many do not" (9 February), at http://www.prnewswire.com/news-releases/political-labels-majorities-of-us-adults-have-a-sense-of-what-conservative-liberal-right-wing-or-left-wing-means-but-many-do-not-54020207.html, accessed 15 January 2012.

M. Richardson, 2008. "Learning about the world through long–term query logs," *ACM Transactions on the Web*, volume 2, number 4, article number 21.

J.T. Ripberger, 2011. "Capturing curiosity: Using Internet search trends to measure public attentiveness," *Policy Studies Journal*, volume 39, number 2, pp. 239–259.

R. Rogers, in press. *Digital methods*. Cambridge, Mass.: MIT Press.

M. Scharkow and J. Vogelgesang, 2011. "Measuring the public agenda using search engine queries," *International Journal of Public Opinion Research*, volume 23, number 1, pp. 104–113.

C.P. Scheitle, 2011. "Google's insights for search: A note evaluating the use of search engine data in social research," *Social Science Quarterly*, volume 92, number 1, pp. 285–295.

R. Seth, M. Covell, D. Ravichandran, D. Sivakumar and S. Baluja, 2011. "A tale of two (similar) cities: Inferring city similarity through geo–spatial query log analysis," In: J. Filipe and A.L.N. Fred (editors). *KDIR 2011: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pp. 179–189.

K. Severson, 2009. "Butterballs or cheese balls, an online barometer," *New York Times* (25 November), at http://www.nytimes.com/2009/11/26/dining/26search.html, accessed 5 June 2011.

A. Spink, B. Jansen and I. Taksa, 2009. "Web log analysis: Diversity of research methodologies," In: B.J. Jansen, A. Spink and I. Taksa (editors). *Handbook of research on Web log analysis*. Hershey, Pa.: Information Science Reference, pp. 506–522.

C.R. Sunstein, 2006. *Infotopia: How many minds produce knowledge*. New York: Oxford University Press.

M. Thelwall, L. Vaughan and L. Björneborn, 2005. "Webometrics," *Annual Review of Information Science and Technology*, volume 39, number 1, pp. 81–135.

B. Theodore, 2011. "New Yahoo! Clues launches," *Yahoo! Search Blog* (29 June), at http://www.ysearchblog.com/2011/06/29/new-yahoo-clues-launches/, accessed 30 June 2011.

H.R. Varian and H. Choi, 2009. "Predicting the present with Google Trends," *Google Research Blog* (2 April), at http://ssrn.com/abstract=1659302, accessed 5 December 2011.

T. Venturini, 2010. "Diving in magma: How to explore controversies with actor-network theory," *Public Understanding of Science*, volume 19, number 3, pp. 258–273.

E.J. Webb, D.T. Campbell, R.D. Schwartz and L. Sechrest, 1972. *Unobtrusive measures:*

Nonreactive measures in the social sciences. Chicago: Rand McNally.

I. Weber and A. Jaimes, 2011. “Who uses Web search for what? And how?” *WSDM ’11: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 15–24.

I. Weber and C. Castillo, 2010. “The demographics of Web search,” *SIGIR ’10: Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 523–530.

I. Weber and A. Jaimes, 2010. “Demographic information flows,” *CIKM ’10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 1,521–1,524.

I. Weber, V. R. K. Garimella and E.K. Borra, 2012. “Mining Web query logs to analyze political issues,” *ACM Web Science 2012: Conference Proceedings*, pp. 479–488.

B. Weeks and B. Southwell, 2010. “The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and Presidential politics,“ *Mass Communication and Society*, volume 13, number 4, pp. 341–360.

Editorial history

Received 20 May 2012; accepted 23 June 2012.

Political Insights: Exploring partisanship in Web search queries
by Erik Borra and Ingmar Weber
First Monday, Volume 17, Number 7 - 2 July 2012
http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/printerFriendly/4070/3272

Social Network Sites: Definition, History, and Scholarship

danah m. boyd

School of Information
University of California-Berkeley

Nicole B. Ellison

Department of Telecommunication, Information Studies, and Media
Michigan State University

Social network sites (SNSs) are increasingly attracting the attention of academic and industry researchers intrigued by their affordances and reach. This special theme section of the Journal of Computer-Mediated Communication *brings together scholarship on these emergent phenomena. In this introductory article, we describe features of SNSs and propose a comprehensive definition. We then present one perspective on the history of such sites, discussing key changes and developments. After briefly summarizing existing scholarship concerning SNSs, we discuss the articles in this special section and conclude with considerations for future research.*

doi:10.1111/j.1083-6101.2007.00393.x

Introduction

Since their introduction, social network sites (SNSs) such as MySpace, Facebook, Cyworld, and Bebo have attracted millions of users, many of whom have integrated these sites into their daily practices. As of this writing, there are hundreds of SNSs, with various technological affordances, supporting a wide range of interests and practices. While their key technological features are fairly consistent, the cultures that emerge around SNSs are varied. Most sites support the maintenance of pre-existing social networks, but others help strangers connect based on shared interests, political views, or activities. Some sites cater to diverse audiences, while others attract people based on common language or shared racial, sexual, religious, or nationality-based identities. Sites also vary in the extent to which they incorporate new information and communication tools, such as mobile connectivity, blogging, and photo/video-sharing.

Scholars from disparate fields have examined SNSs in order to understand the practices, implications, culture, and meaning of the sites, as well as users' engagement with them. This special theme section of the *Journal of Computer-Mediated Communication* brings together a unique collection of articles that analyze a wide spectrum of social network sites using various methodological techniques, theoretical traditions, and analytic approaches. By collecting these articles in this issue, our goal is to showcase some of the interdisciplinary scholarship around these sites.

The purpose of this introduction is to provide a conceptual, historical, and scholarly context for the articles in this collection. We begin by defining what constitutes a social network site and then present one perspective on the historical development of SNSs, drawing from personal interviews and public accounts of sites and their changes over time. Following this, we review recent scholarship on SNSs and attempt to contextualize and highlight key works. We conclude with a description of the articles included in this special section and suggestions for future research.

Social Network Sites: A Definition

We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site.

While we use the term "social network site" to describe this phenomenon, the term "social networking sites" also appears in public discourse, and the two terms are often used interchangeably. We chose not to employ the term "networking" for two reasons: emphasis and scope. "Networking" emphasizes relationship initiation, often between strangers. While networking is possible on these sites, it is not the primary practice on many of them, nor is it what differentiates them from other forms of computer-mediated communication (CMC).

What makes social network sites unique is not that they allow individuals to meet strangers, but rather that they enable users to articulate and make visible their social networks. This can result in connections between individuals that would not otherwise be made, but that is often not the goal, and these meetings are frequently between "latent ties" (Haythornthwaite, 2005) who share some offline connection. On many of the large SNSs, participants are not necessarily "networking" or looking to meet new people; instead, they are primarily communicating with people who are already a part of their extended social network. To emphasize this articulated social network as a critical organizing feature of these sites, we label them "social network sites."

While SNSs have implemented a wide variety of technical features, their backbone consists of visible profiles that display an articulated list of Friends[1] who are also users of the system. Profiles are unique pages where one can "type oneself into being" (Sundén, 2003, p. 3). After joining an SNS, an individual is asked to fill out

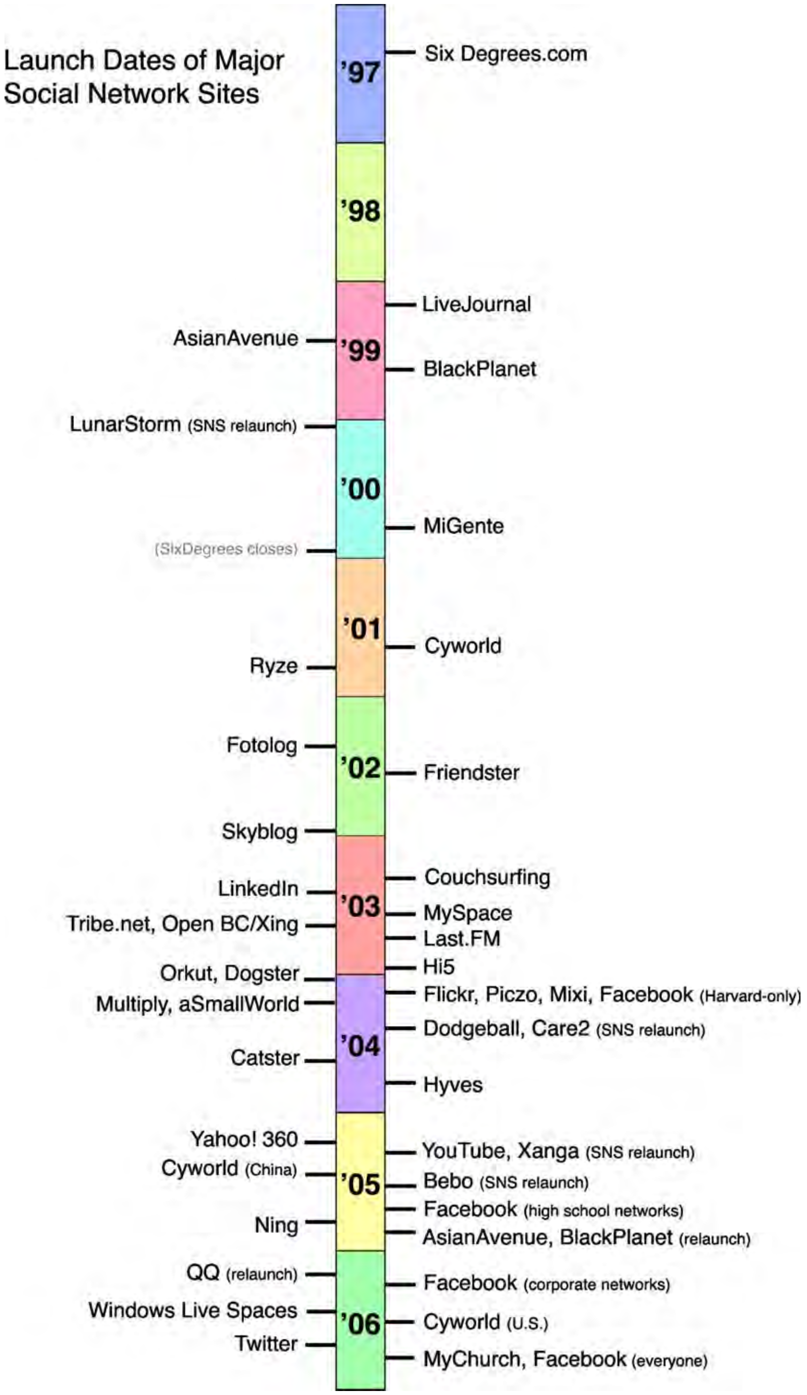


Figure 1 Timeline of the launch dates of many major SNSs and dates when community sites re-launched with SNS features

forms containing a series of questions. The profile is generated using the answers to these questions, which typically include descriptors such as age, location, interests, and an "about me" section. Most sites also encourage users to upload a profile photo. Some sites allow users to enhance their profiles by adding multimedia content or modifying their profile's look and feel. Others, such as Facebook, allow users to add modules ("Applications") that enhance their profile.

The visibility of a profile varies by site and according to user discretion. By default, profiles on Friendster and Tribe.net are crawled by search engines, making them visible to anyone, regardless of whether or not the viewer has an account. Alternatively, LinkedIn controls what a viewer may see based on whether she or he has a paid account. Sites like MySpace allow users to choose whether they want their profile to be public or "Friends only." Facebook takes a different approach—by default, users who are part of the same "network" can view each other's profiles, unless a profile owner has decided to deny permission to those in their network. Structural variations around visibility and access are one of the primary ways that SNSs differentiate themselves from each other.

After joining a social network site, users are prompted to identify others in the system with whom they have a relationship. The label for these relationships differs depending on the site—popular terms include "Friends," "Contacts," and "Fans." Most SNSs require bi-directional confirmation for Friendship, but some do not. These one-directional ties are sometimes labeled as "Fans" or "Followers," but many sites call these Friends as well. The term "Friends" can be misleading, because the connection does not necessarily mean friendship in the everyday vernacular sense, and the reasons people connect are varied (boyd, 2006a).

The public display of connections is a crucial component of SNSs. The Friends list contains links to each Friend's profile, enabling viewers to traverse the network graph by clicking through the Friends lists. On most sites, the list of Friends is visible to anyone who is permitted to view the profile, although there are exceptions. For instance, some MySpace users have hacked their profiles to hide the Friends display, and LinkedIn allows users to opt out of displaying their network.

Most SNSs also provide a mechanism for users to leave messages on their Friends' profiles. This feature typically involves leaving "comments," although sites employ various labels for this feature. In addition, SNSs often have a private messaging feature similar to webmail. While both private messages and comments are popular on most of the major SNSs, they are not universally available.

Not all social network sites began as such. QQ started as a Chinese instant messaging service, LunarStorm as a community site, Cyworld as a Korean discussion forum tool, and Skyrock (formerly Skyblog) was a French blogging service before adding SNS features. Classmates.com, a directory of school affiliates launched in 1995, began supporting articulated lists of Friends after SNSs became popular. AsianAvenue, MiGente, and BlackPlanet were early popular ethnic community sites with limited Friends functionality before re-launching in 2005–2006 with SNS features and structure.

Beyond profiles, Friends, comments, and private messaging, SNSs vary greatly in their features and user base. Some have photo-sharing or video-sharing capabilities; others have built-in blogging and instant messaging technology. There are mobile-specific SNSs (e.g., Dodgeball), but some web-based SNSs also support limited mobile interactions (e.g., Facebook, MySpace, and Cyworld). Many SNSs target people from specific geographical regions or linguistic groups, although this does not always determine the site's constituency. Orkut, for example, was launched in the United States with an English-only interface, but Portuguese-speaking Brazilians quickly became the dominant user group (Kopytoff, 2004). Some sites are designed with specific ethnic, religious, sexual orientation, political, or other identity-driven categories in mind. There are even SNSs for dogs (Dogster) and cats (Catster), although their owners must manage their profiles.

While SNSs are often designed to be widely accessible, many attract homogeneous populations initially, so it is not uncommon to find groups using sites to segregate themselves by nationality, age, educational level, or other factors that typically segment society (Hargittai, this issue), even if that was not the intention of the designers.

A History of Social Network Sites

The Early Years

According to the definition above, the first recognizable social network site launched in 1997. SixDegrees.com allowed users to create profiles, list their Friends and, beginning in 1998, surf the Friends lists. Each of these features existed in some form before SixDegrees, of course. Profiles existed on most major dating sites and many community sites. AIM and ICQ buddy lists supported lists of Friends, although those Friends were not visible to others. Classmates.com allowed people to affiliate with their high school or college and surf the network for others who were also affiliated, but users could not create profiles or list Friends until years later. SixDegrees was the first to combine these features.

SixDegrees promoted itself as a tool to help people connect with and send messages to others. While SixDegrees attracted millions of users, it failed to become a sustainable business and, in 2000, the service closed. Looking back, its founder believes that SixDegrees was simply ahead of its time (A. Weinreich, personal communication, July 11, 2007). While people were already flocking to the Internet, most did not have extended networks of friends who were online. Early adopters complained that there was little to do after accepting Friend requests, and most users were not interested in meeting strangers.

From 1997 to 2001, a number of community tools began supporting various combinations of profiles and publicly articulated Friends. AsianAvenue, BlackPlanet, and MiGente allowed users to create personal, professional, and dating profiles—users could identify Friends on their personal profiles without seeking approval for those connections (O. Wasow, personal communication, August 16, 2007). Likewise, shortly after its launch in 1999, LiveJournal listed one-directional connections on user pages. LiveJournal's creator suspects that he fashioned these Friends after instant messaging buddy lists (B. Fitzpatrick, personal communication, June 15, 2007)—on LiveJournal, people mark others as Friends to follow their journals and manage privacy settings. The Korean virtual worlds site Cyworld was started in 1999 and added SNS features in 2001, independent of these other sites (see Kim & Yun, this issue). Likewise, when the Swedish web community LunarStorm refashioned itself as an SNS in 2000, it contained Friends lists, guestbooks, and diary pages (D. Skog, personal communication, September 24, 2007).

The next wave of SNSs began when Ryze.com was launched in 2001 to help people leverage their business networks. Ryze's founder reports that he first introduced the site to his friends—primarily members of the San Francisco business and technology community, including the entrepreneurs and investors behind many future SNSs (A. Scott, personal communication, June 14, 2007). In particular, the people behind Ryze, Tribe.net, LinkedIn, and Friendster were tightly entwined personally and professionally. They believed that they could support each other without competing (Festa, 2003). In the end, Ryze never acquired mass popularity, Tribe.net grew to attract a passionate niche user base, LinkedIn became a powerful business service, and Friendster became the most significant, if only as "one of the biggest disappointments in Internet history" (Chafkin, 2007, p. 1).

Like any brief history of a major phenomenon, ours is necessarily incomplete. In the following section we discuss Friendster, MySpace, and Facebook, three key SNSs that shaped the business, cultural, and research landscape.

The Rise (and Fall) of Friendster

Friendster launched in 2002 as a social complement to Ryze. It was designed to compete with Match.com, a profitable online dating site (Cohen, 2003). While most dating sites focused on introducing people to strangers with similar interests, Friendster was designed to help friends-of-friends meet, based on the assumption that friends-of-friends would make better romantic partners than would strangers (J. Abrams, personal communication, March 27, 2003). Friendster gained traction among three groups of early adopters who shaped the site—bloggers, attendees of the Burning Man arts festival, and gay men (boyd, 2004)—and grew to 300,000 users through word of mouth before traditional press coverage began in May 2003 (O'Shea, 2003).

As Friendster's popularity surged, the site encountered technical and social difficulties (boyd, 2006b). Friendster's servers and databases were ill-equipped to handle its rapid growth, and the site faltered regularly, frustrating users who replaced email with Friendster. Because organic growth had been critical to creating a coherent community, the onslaught of new users who learned about the site from media coverage upset the cultural balance. Furthermore, exponential growth meant a collapse in social contexts: Users had to face their bosses and former classmates alongside their close friends. To complicate matters, Friendster began restricting the activities of its most passionate users.

The initial design of Friendster restricted users from viewing profiles of people who were more than four degrees away (friends-of-friends-of-friends-of-friends). In order to view additional profiles, users began adding acquaintances and interesting-looking strangers to expand their reach. Some began massively collecting Friends, an activity that was implicitly encouraged through a "most popular" feature. The ultimate collectors were fake profiles representing iconic fictional characters: celebrities, concepts, and other such entities. These "Fakesters" outraged the company, who banished fake profiles and eliminated the "most popular" feature (boyd, in press-b). While few people actually created Fakesters, many more enjoyed surfing Fakesters for entertainment or using functional Fakesters (e.g., "Brown University") to find people they knew.

The active deletion of Fakesters (and genuine users who chose non-realistic photos) signaled to some that the company did not share users' interests. Many early adopters left because of the combination of technical difficulties, social collisions, and a rupture of trust between users and the site (boyd, 2006b). However, at the same time that it was fading in the U.S., its popularity skyrocketed in the Philippines, Singapore, Malaysia, and Indonesia (Goldberg, 2007).

SNSs Hit the Mainstream

From 2003 onward, many new SNSs were launched, prompting social software analyst Clay Shirky (2003) to coin the term YASNS: "Yet Another Social Networking Service." Most took the form of profile-centric sites, trying to replicate the early success of Friendster or target specific demographics. While socially-organized SNSs solicit broad audiences, professional sites such as LinkedIn, Visible Path, and Xing (formerly openBC) focus on business people. "Passion-centric" SNSs like Dogster (T. Rheingold, personal communication, August 2, 2007) help strangers connect based on shared interests. Care2 helps activists meet, Couchsurfing connects travelers to people with couches, and MyChurch joins Christian churches and their members. Furthermore, as the social media and user-generated content phenomena grew, websites focused on media sharing began implementing SNS features and becoming SNSs themselves. Examples include Flickr (photo sharing), Last.FM (music listening habits), and YouTube (video sharing).

With the plethora of venture-backed startups launching in Silicon Valley, few people paid attention to SNSs that gained popularity elsewhere, even those built by major corporations. For example, Google's Orkut failed to build a sustainable U.S. user base, but a "Brazilian invasion" (Fragoso, 2006) made Orkut the national SNS of Brazil. Microsoft's Windows Live Spaces (a.k.a. MSN Spaces) also launched to lukewarm U.S. reception but became extremely popular elsewhere.

Few analysts or journalists noticed when MySpace launched in Santa Monica, California, hundreds of miles from Silicon Valley. MySpace was begun in 2003 to compete with sites like Friendster, Xanga, and AsianAvenue, according to co-founder Tom Anderson (personal communication, August 2, 2007); the founders wanted to attract estranged Friendster users (T. Anderson, personal communication, February 2, 2006). After rumors emerged that Friendster would adopt a fee-based system, users posted Friendster messages encouraging people to join alternate SNSs, including Tribe.net and MySpace (T. Anderson, personal communication, August 2, 2007). Because of this, MySpace was able to grow rapidly by capitalizing on Friendster's alienation of its early adopters. One particularly notable group that encouraged others to switch were indie-rock bands who were expelled from Friendster for failing to comply with profile regulations.

While MySpace was not launched with bands in mind, they were welcomed. Indie-rock bands from the Los Angeles region began creating profiles, and local promoters used MySpace to advertise VIP passes for popular clubs. Intrigued, MySpace contacted local musicians to see how they could support them (T. Anderson, personal communication, September 28, 2006). Bands were not the sole source of MySpace growth, but the symbiotic relationship between bands and fans helped MySpace expand beyond former Friendster users. The bands-and-fans dynamic was mutually beneficial: Bands wanted to be able to contact fans, while fans desired attention from their favorite bands and used Friend connections to signal identity and affiliation.

Futhermore, MySpace differentiated itself by regularly adding features based on user demand (boyd, 2006b) and by allowing users to personalize their pages. This "feature" emerged because MySpace did not restrict users from adding HTML into the forms that framed their profiles; a copy/paste code culture emerged on the web to support users in generating unique MySpace backgrounds and layouts (Perkel, in press).

Teenagers began joining MySpace *en masse* in 2004. Unlike older users, most teens were never on Friendster—some joined because they wanted to connect with their favorite bands; others were introduced to the site through older family members. As teens began signing up, they encouraged their friends to join. Rather than rejecting underage users, MySpace changed its user policy to allow minors. As the site grew, three distinct populations began to form: musicians/artists, teenagers, and the post-college urban social crowd. By and large, the latter two groups did not interact with one another except through bands. Because of the lack of mainstream press coverage during 2004, few others noticed the site's growing popularity.

Then, in July 2005, News Corporation purchased MySpace for $580 million (BBC, 2005), attracting massive media attention. Afterwards, safety issues plagued MySpace. The site was implicated in a series of sexual interactions between adults and minors, prompting legal action (Consumer Affairs, 2006). A moral panic concerning sexual predators quickly spread (Bahney, 2006), although research suggests that the concerns were exaggerated.[2]

A Global Phenomenon

While MySpace attracted the majority of media attention in the U.S. and abroad, SNSs were proliferating and growing in popularity worldwide. Friendster gained traction in the Pacific Islands, Orkut became the premier SNS in Brazil before growing rapidly in India (Madhavan, 2007), Mixi attained widespread adoption in Japan, LunarStorm took off in Sweden, Dutch users embraced Hyves, Grono captured Poland, Hi5 was adopted in smaller countries in Latin America, South America, and Europe, and Bebo became very popular in the United Kingdom, New Zealand, and Australia. Additionally, previously popular communication and community services began implementing SNS features. The Chinese QQ instant messaging service instantly became the largest SNS worldwide when it added profiles and made friends visible (McLeod, 2006), while the forum tool Cyworld cornered the Korean market by introducing homepages and buddies (Ewers, 2006).

Blogging services with complete SNS features also became popular. In the U.S., blogging tools with SNS features, such as Xanga, LiveJournal, and Vox, attracted broad audiences. Skyrock reigns in France, and Windows Live Spaces dominates numerous markets worldwide, including in Mexico, Italy, and Spain. Although SNSs like QQ, Orkut, and Live Spaces are just as large as, if not larger than, MySpace, they receive little coverage in U.S. and English-speaking media, making it difficult to track their trajectories.

Expanding Niche Communities

Alongside these open services, other SNSs launched to support niche demographics before expanding to a broader audience. Unlike previous SNSs, Facebook was designed to support distinct college networks only. Facebook began in early 2004 as a Harvard-only SNS (Cassidy, 2006). To join, a user had to have a harvard.edu email address. As Facebook began supporting other schools, those users were also required to have university email addresses associated with those institutions, a requirement that kept the site relatively closed and contributed to users' perceptions of the site as an intimate, private community.

Beginning in September 2005, Facebook expanded to include high school students, professionals inside corporate networks, and, eventually, everyone. The change to open signup did not mean that new users could easily access users in closed networks—gaining access to corporate networks still required the appropriate .com address, while gaining access to high school networks required administrator approval. (As of this writing, only membership in regional networks requires no permission.) Unlike other SNSs, Facebook users are unable to make their full profiles public to all users. Another feature that differentiates Facebook is the ability for outside developers to build "Applications" which allow users to personalize their profiles and perform other tasks, such as compare movie preferences and chart travel histories.

While most SNSs focus on growing broadly and exponentially, others explicitly seek narrower audiences. Some, like aSmallWorld and BeautifulPeople, intentionally restrict access to appear selective and elite. Others—activity-centered sites like Couchsurfing, identity-driven sites like BlackPlanet, and affiliation-focused sites like MyChurch—are limited by their target demographic and thus tend to be smaller. Finally, anyone who wishes to create a niche social network site can do so on Ning, a platform and hosting service that encourages users to create their own SNSs.

Currently, there are no reliable data regarding how many people use SNSs, although marketing research indicates that SNSs are growing in popularity worldwide (comScore, 2007). This growth has prompted many corporations to invest time and money in creating, purchasing, promoting, and advertising SNSs. At the same time, other companies are blocking their employees from accessing the sites. Additionally, the U.S. military banned soldiers from accessing MySpace (Frosch, 2007) and the Canadian government prohibited employees from Facebook (Benzie, 2007), while the U.S. Congress has proposed legislation to ban youth from accessing SNSs in schools and libraries (H.R. 5319, 2006; S. 49, 2007).

The rise of SNSs indicates a shift in the organization of online communities. While websites dedicated to communities of interest still exist and prosper, SNSs are primarily organized around people, not interests. Early public online communities such as Usenet and public discussion forums were structured by topics or according to topical hierarchies, but social network sites are structured as personal (or "egocentric") networks, with the individual at the center of their own community. This more accurately mirrors unmediated social structures, where "the world is composed of networks, not groups" (Wellman, 1988, p. 37). The introduction of SNS features has introduced a new organizational framework for online communities, and with it, a vibrant new research context.

Previous Scholarship

Scholarship concerning SNSs is emerging from diverse disciplinary and methodological traditions, addresses a range of topics, and builds on a large body of CMC research. The goal of this section is to survey research that is directly concerned with social network sites, and in so doing, to set the stage for the articles in this special issue. To date, the bulk of SNS research has focused on impression management and friendship performance, networks and network structure, online/offline connections, and privacy issues.

Impression Management and Friendship Performance

Like other online contexts in which individuals are consciously able to construct an online representation of self—such as online dating profiles and MUDS—SNSs constitute an important research context for scholars investigating processes of impression management, self-presentation, and friendship performance. In one of the earliest academic articles on SNSs, boyd (2004) examined Friendster as a locus of publicly articulated social networks that allowed users to negotiate presentations of self and connect with others. Donath and boyd (2004) extended this to suggest that "public displays of connection" serve as important identity signals that help people navigate the networked social world, in that an extended network may serve to validate identity information presented in profiles.

While most sites encourage users to construct accurate representations of themselves, participants do this to varying degrees. Marwick (2005) found that users on three different SNSs had complex strategies for negotiating the rigidity of a prescribed "authentic" profile, while boyd (in press-b) examined the phenomenon of "Fakesters" and argued that profiles could never be "real." The extent to which portraits are authentic or playful varies across sites; both social and technological forces shape user practices. Skog (2005) found that the status feature on LunarStorm strongly influenced how people behaved and what they choose to reveal—profiles there indicate one's status as measured by activity (e.g., sending messages) and indicators of authenticity (e.g., using a "real" photo instead of a drawing).

Another aspect of self-presentation is the articulation of friendship links, which serve as identity markers for the profile owner. Impression management is one of the reasons given by Friendster users for choosing particular friends (Donath & boyd, 2004). Recognizing this, Zinman and Donath (2007) noted that MySpace spammers leverage people's willingness to connect to interesting people to find targets for their spam.

In their examination of LiveJournal "friendship," Fono and Raynes-Goldie (2006) described users' understandings regarding public displays of connections and how the Friending function can operate as a catalyst for social drama. In listing user motivations for Friending, boyd (2006a) points out that "Friends" on SNSs are not the same as "friends" in the everyday sense; instead, Friends provide context by offering users an imagined audience to guide behavioral norms. Other work in this area has examined the use of Friendster Testimonials as self-presentational devices (boyd & Heer, 2006) and the extent to which the attractiveness of one's Friends (as indicated by Facebook's "Wall" feature) impacts impression formation (Walther, Van Der Heide, Kim, & Westerman, in press).

Networks and Network Structure

Social network sites also provide rich sources of naturalistic behavioral data. Profile and linkage data from SNSs can be gathered either through the use of automated collection techniques or through datasets provided directly from the company, enabling network analysis researchers to explore large-scale patterns of friending, usage, and other visible indicators (Hogan, in press), and continuing an analysis trend that started with examinations of blogs and other websites. For instance, Golder, Wilkinson, and Huberman (2007) examined an anonymized dataset consisting of 362 million messages exchanged by over four million Facebook users for insight into Friending and messaging activities. Lampe, Ellison, and Steinfield (2007) explored the relationship between profile elements and number of Facebook friends, finding that profile fields that reduce transaction costs and are harder to falsify are most likely to be associated with larger number of friendship links. These kinds of data also lend themselves well to analysis through network visualization (Adamic, Buyukkokten, & Adar, 2003; Heer & boyd, 2005; Paolillo & Wright, 2005).

SNS researchers have also studied the network structure of Friendship. Analyzing the roles people played in the growth of Flickr and Yahoo! 360's networks, Kumar, Novak, and Tomkins (2006) argued that there are passive members, inviters, and linkers “who fully participate in the social evolution of the network” (p. 1). Scholarship concerning LiveJournal’s network has included a Friendship classification scheme (Hsu, Lancaster, Paradesi, & Weniger, 2007), an analysis of the role of language in the topology of Friendship (Herring et al., 2007), research into the importance of geography in Friending (Liben-Nowell, Novak, Kumar, Raghavan, and Tomkins, 2005), and studies on what motivates people to join particular communities (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006). Based on Orkut data, Spertus, Sahami, and Buyukkokten (2005) identified a topology of users through their membership in certain communities; they suggest that sites can use this to recommend additional communities of interest to users. Finally, Liu, Maes, and Davenport (2006) argued that Friend connections are not the only network structure worth investigating. They examined the ways in which the performance of tastes (favorite music, books, film, etc.) constitutes an alternate network structure, which they call a “taste fabric.”

Bridging Online and Offline Social Networks

Although exceptions exist, the available research suggests that most SNSs primarily support pre-existing social relations. Ellison, Steinfield, and Lampe (2007) suggest that Facebook is used to maintain existing offline relationships or solidify offline connections, as opposed to meeting new people. These relationships may be weak ties, but typically there is some common offline element among individuals who friend one another, such as a shared class at school. This is one of the chief dimensions that differentiate SNSs from earlier forms of public CMC such as newsgroups (Ellison et al., 2007). Research in this vein has investigated how online interactions interface with offline ones. For instance, Lampe, Ellison, and Steinfield (2006) found that Facebook users engage in “searching” for people with whom they have an offline connection more than they “browse” for complete strangers to meet. Likewise, Pew research found that 91% of U.S. teens who use SNSs do so to connect with friends (Lenhart & Madden, 2007).

Given that SNSs enable individuals to connect with one another, it is not surprising that they have become deeply embedded in user’s lives. In Korea, Cyworld has become an integral part of everyday life—Choi (2006) found that 85% of that study’s respondents “listed the maintenance and reinforcement of pre-existing social networks as their main motive for Cyworld use” (p. 181). Likewise, boyd (2008) argues that MySpace and Facebook enable U.S. youth to socialize with their friends even when they are unable to gather in unmediated situations; she argues that SNSs are “networked publics” that support sociability, just as unmediated public spaces do.

Privacy

Popular press coverage of SNSs has emphasized potential privacy concerns, primarily concerning the safety of younger users (George, 2006; Kornblum & Marklein, 2006). Researchers have investigated the potential threats to privacy associated with SNSs.

In one of the first academic studies of privacy and SNSs, Gross and Acquisti (2005) analyzed 4,000 Carnegie Mellon University Facebook profiles and outlined the potential threats to privacy contained in the personal information included on the site by students, such as the potential ability to reconstruct users' social security numbers using information often found in profiles, such as hometown and date of birth.

Acquisti and Gross (2006) argue that there is often a disconnect between students' desire to protect privacy and their behaviors, a theme that is also explored in Stutzman's (2006) survey of Facebook users and Barnes's (2006) description of the "privacy paradox" that occurs when teens are not aware of the public nature of the Internet. In analyzing trust on social network sites, Dwyer, Hiltz, and Passerini (2007) argued that trust and usage goals may affect what people are willing to share—Facebook users expressed greater trust in Facebook than MySpace users did in MySpace and thus were more willing to share information on the site.

In another study examining security issues and SNSs, Jagatic, Johnson, Jakobsson, and Menczer (2007) used freely accessible profile data from SNSs to craft a "phishing" scheme that appeared to originate from a friend on the network; their targets were much more likely to give away information to this "friend" than to a perceived stranger. Survey data offer a more optimistic perspective on the issue, suggesting that teens are aware of potential privacy threats online and that many are proactive about taking steps to minimize certain potential risks. Pew found that 55% of online teens have profiles, 66% of whom report that their profile is not visible to all Internet users (Lenhart & Madden, 2007). Of the teens with completely open profiles, 46% reported including at least some false information.

Privacy is also implicated in users' ability to control impressions and manage social contexts. Boyd (in press-a) asserted that Facebook's introduction of the "News Feed" feature disrupted students' sense of control, even though data exposed through the feed were previously accessible. Preibusch, Hoser, Gürses, and Berendt (2007) argued that the privacy options offered by SNSs do not provide users with the flexibility they need to handle conflicts with Friends who have different conceptions of privacy; they suggest a framework for privacy in SNSs that they believe would help resolve these conflicts.

SNSs are also challenging legal conceptions of privacy. Hodge (2006) argued that the fourth amendment to the U.S. Constitution and legal decisions concerning privacy are not equipped to address social network sites. For example, do police officers have the right to access content posted to Facebook without a warrant? The legality of this hinges on users' expectation of privacy and whether or not Facebook profiles are considered public or private.

Other Research

In addition to the themes identified above, a growing body of scholarship addresses other aspects of SNSs, their users, and the practices they enable. For example, scholarship on the ways in which race and ethnicity (Byrne, in press; Gajjala, 2007), religion (Nyland & Near, 2007), gender (Geidner, Flook, & Bell, 2007; Hjorth & Kim, 2005), and sexuality connect to, are affected by, and are enacted in social network sites raise interesting questions about how identity is shaped within these sites. Fragoso (2006) examined the role of national identity in SNS use through an investigation into the "Brazilian invasion" of Orkut and the resulting culture clash between Brazilians and Americans on the site. Other scholars are beginning to do cross-cultural comparisons of SNS use—Hjorth and Yuji (in press) compare Japanese usage of Mixi and Korean usage of Cyworld, while Herring et al. (2007) examine the practices of users who bridge different languages on LiveJournal—but more work in this area is needed.

Scholars are documenting the implications of SNS use with respect to schools, universities, and libraries. For example, scholarship has examined how students feel about having professors on Facebook (Hewitt & Forte, 2006) and how faculty participation affects student-professor relations (Mazer, Murphy, & Simonds, 2007). Charnigo and Barnett-Ellis (2007) found that librarians are overwhelmingly aware of Facebook and are against proposed U.S. legislation that would ban minors from accessing SNSs at libraries, but that most see SNSs as outside the purview of librarianship. Finally, challenging the view that there is nothing educational about SNSs, Perkel (in press) analyzed copy/paste practices on MySpace as a form of literacy involving social and technical skills.

This overview is not comprehensive due to space limitations and because much work on SNSs is still in the process of being published. Additionally, we have not included literature in languages other than English (e.g., Recuero, 2005 on social capital and Orkut), due to our own linguistic limitations.

Overview of This Special Theme Section

The articles in this section address a variety of social network sites—BlackPlanet, Cyworld, Dodgeball, Facebook, MySpace, and YouTube—from multiple theoretical and methodological angles, building on previous studies of SNSs and broader theoretical traditions within CMC research, including relationship maintenance and issues of identity, performance, privacy, self-presentation, and civic engagement.

These pieces collectively provide insight into some of the ways in which online and offline experiences are deeply entwined. Using a relational dialectics approach, **Kyung-Hee Kim** and **Haejin Yun** analyze how Cyworld supports both interpersonal relations and self-relation for Korean users. They trace the subtle ways in which deeply engrained cultural beliefs and activities are integrated into online communication and behaviors on Cyworld—the online context reinforces certain aspects of users' cultural expectations about relationship maintenance (e.g., the concept of reciprocity), while the unique affordances of Cyworld enable participants to overcome offline constraints. **Dara Byrne** uses content analysis to examine civic engagement in forums on BlackPlanet and finds that online discussions are still plagued with the problems offline activists have long encountered. Drawing on interview and observation data, **Lee Humphreys** investigates early adopters' practices involving Dodgeball, a mobile social network service. She looks at the ways in which networked communication is reshaping offline social geography.

Other articles in this collection illustrate how innovative research methods can elucidate patterns of behavior that would be indistinguishable otherwise. For instance, **Hugo Liu** examines participants' performance of tastes and interests by analyzing and modeling the preferences listed on over 127,000 MySpace profiles, resulting in unique "taste maps." Likewise, through survey data collected at a college with diverse students in the U.S., **Eszter Hargittai** illuminates usage patterns that would otherwise be masked. She finds that adoption of particular services correlates with individuals' race and parental education level.

Existing theory is deployed, challenged, and extended by the approaches adopted in the articles in this section. **Judith Donath** extends signaling theory to explain different tactics SNS users adopt to reduce social costs while managing trust and identity. She argues that the construction and maintenance of relations on SNSs is akin to "social grooming." **Patricia Lange** complicates traditional dichotomies between "public" and "private" by analyzing how YouTube participants blur these lines in their video-sharing practices.

The articles in this collection highlight the significance of social network sites in the lives of users and as a topic of research. Collectively, they show how networked practices mirror, support, and alter known everyday practices, especially with respect to how people present (and hide) aspects of themselves and connect with others. The fact that participation on social network sites leaves online traces offers unprecedented opportunities for researchers. The scholarship in this special theme section takes advantage of this affordance, resulting in work that helps explain practices online and offline, as well as those that blend the two environments.

Future Research

The work described above and included in this special theme section contributes to an on-going dialogue about the importance of social network sites, both for practitioners and researchers. Vast, uncharted waters still remain to be explored. Methodologically, SNS researchers' ability to make causal claims is limited by a lack of experimental or longitudinal studies. Although the situation is rapidly changing, scholars still have a limited understanding of who is and who is not using these sites, why, and for what purposes, especially outside the U.S. Such questions will require large-scale quantitative and qualitative research. Richer, ethnographic research on populations more difficult to access (including non-users) would further aid scholars' ability to understand the long-term implications of these tools. We hope that the work described here and included in this collection will help build a foundation for future investigations of these and other important issues surrounding social network sites.

Acknowledgments

We are grateful to the external reviewers who volunteered their time and expertise to review papers and contribute valuable feedback and to those practitioners and analysts who provided information to help shape the history section. Thank you also to Susan Herring, whose patience and support appeared infinite.

Notes

[1] To differentiate the articulated list of Friends on SNSs from the colloquial term "friends," we capitalize the former.

[2] Although one out of seven teenagers received unwanted sexual solicitations online, only 9% came from people over the age of 25 (Wolak, Mitchell, & Finkelhor, 2006). Research suggests that popular narratives around sexual predators on SNSs are misleading—cases of unsuspecting teens being lured by sexual predators are rare (Finkelhor, Ybarra, Lenhart, boyd, & Lordan, 2007). Furthermore, only .08% of students surveyed by the National School Boards Association (2007) met someone in person from an online encounter without permission from a parent.

References

Acquisti, A., & Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the Facebook. In P. Golle & G. Danezis (Eds.), *Proceedings of 6th Workshop on Privacy Enhancing Technologies* (pp. 36–58). Cambridge, UK: Robinson College.

Adamic, L. A., Büyükkökten, O., & Adar, E. (2003). A social network caught in the Web. *First Monday*, **8**(6). Retrieved July 30, 2007 from http://www.firstmonday.org/issues/issue8_6/adamic/index.html

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. *Proceedings of 12th International Conference on Knowledge Discovery in Data Mining* (pp. 44–54). New York: ACM Press.

Bahney, A. (2006, March 9). Don't talk to invisible strangers. *New York Times*. Retrieved July 21, 2007 from http://www.nytimes.com/2006/03/09/fashion/thursdaystyles/09parents.html

Barnes, S. (2006). A privacy paradox: Social networking in the United States. *First Monday*, **11**(9). Retrieved September 8, 2007 from http://www.firstmonday.org/issues/issue11_9/barnes/index.html

BBC. (2005, July 19). *News Corp in $580m Internet buy*. Retrieved July 21, 2007 from http://news.bbc.co.uk/2/hi/business/4695495.stm

Benzie, R. (2007, May 3). Facebook banned for Ontario staffers. *The Star*. Retrieved July 21, 2007 from http://www.thestar.com/News/article/210014

boyd, d. (2004). Friendster and publicly articulated social networks. *Proceedings of ACM Conference on Human Factors in Computing Systems* (pp. 1279–1282). New York: ACM Press.

boyd, d. (2006a). Friends, Friendsters, and MySpace Top 8: Writing community into being on social network sites. *First Monday*, **11**(12). Retrieved July 21, 2007 from http://www.firstmonday.org/issues/issue11_12/boyd/

boyd, d. (2006b, March 21). Friendster lost steam. Is MySpace just a fad? *Apophenia Blog*. Retrieved July 21, 2007 from http://www.danah.org/papers/FriendsterMySpaceEssay.html

boyd, d. (in press-a). Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence*, **14**(1).

boyd, d. (in press-b). None of this is real. In J. Karaganis (Ed.), *Structures of Participation*. New York: Social Science Research Council.

boyd, d. (2008). Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), *Youth, Identity, and Digital Media* (pp. 119–142). Cambridge, MA: MIT Press.

boyd, d., & Heer, J. (2006). Profiles as conversation: Networked identity performance on Friendster. *Proceedings of Thirty-Ninth Hawai'i International Conference on System Sciences*. Los Alamitos, CA: IEEE Press.

Byrne, D. (in press). The future of (the) 'race': Identity, discourse and the rise of computer-mediated public spheres. In A. Everett (Ed.), *MacArthur Foundation Book Series on Digital Learning: Race and Ethnicity Volume* (pp. 15–38). Cambridge, MA: MIT Press.

Cassidy, J. (2006, May 15). Me media: How hanging out on the Internet became big business. *The New Yorker*, **82**(13), 50.

Chafkin, M. (2007, June). How to kill a great idea! *Inc. Magazine*. Retrieved August 27, 2007 from http://www.inc.com/magazine/20070601/features-how-to-kill-a-great-idea.html

Charnigo, L., & Barnett-Ellis, P. (2007). Checking out Facebook.com: The impact of a digital trend on academic libraries. *Information Technology and Libraries*, **26**(1), 23.

Choi, J. H. (2006). Living in *Cyworld*: Contextualising Cy-Ties in South Korea. In A. Bruns & J. Jacobs (Eds.), *Use of Blogs (Digital Formations)* (pp. 173–186). New York: Peter Lang.

Cohen, R. (2003, July 5). Livewire: Web sites try to make internet dating less creepy. *Reuters*. Retrieved July 5, 2003 from http://asia.reuters.com/newsArticle.jhtml?type=internetNews&storyID=3041934

comScore. (2007). Social networking goes global. Reston, VA. Retrieved September 9, 2007 from http://www.comscore.com/press/release.asp?press=1555

Consumer Affairs. (2006, February 5). Connecticut opens MySpace.com probe. *Consumer Affairs*. Retrieved July 21, 2007 from http://www.consumeraffairs.com/news04/2006/02/myspace.html

Donath, J., & boyd, d. (2004). Public displays of connection. *BT Technology Journal*, **22**(4), 71–82.

Dwyer, C., Hiltz, S. R., & Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. *Proceedings of AMCIS 2007*, Keystone, CO. Retrieved September 21, 2007 from http://csis.pace.edu/~dwyer/research/DwyerAMCIS2007.pdf

Ellison, N., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends": Exploring the relationship between college students' use of online social networks and social capital. *Journal of Computer-Mediated Communication*, **12**(3), article 1. Retrieved July 30, 2007 from http://jcmc.indiana.edu/vol12/issue4/ellison.html

Ewers, J. (2006, November 9). Cyworld: Bigger than YouTube? *U.S. News & World Report*. Retrieved July 30, 2007 from *LexisNexis*.

Festa, P. (2003, November 11). Investors snub Friendster in patent grab. *CNet News*. Retrieved August 26, 2007 from http://news.com.com/2100-1032_3-5106136.html

Finkelhor, D., Ybarra, M., Lenhart, A., boyd, d., & Lordan, T. (2007, May 3). Just the facts about online youth victimization: Researchers present the facts and debunk myths. *Internet Caucus Advisory Committee Event*. Retrieved July 21, 2007 from http://www.netcaucus.org/events/2007/youth/20070503transcript.pdf

Fono, D., & Raynes-Goldie, K. (2006). Hyperfriendship and beyond: Friends and social norms on LiveJournal. In M. Consalvo & C. Haythornthwaite (Eds.), *Internet Research Annual Volume 4: Selected Papers from the AOIR Conference* (pp. 91–103). New York: Peter Lang.

Fragoso, S. (2006). WTF a crazy Brazilian invasion. In F. Sudweeks & H. Hrachovec (Eds.), *Proceedings of CATaC 2006* (pp. 255–274). Murdoch, Australia: Murdoch University.

Frosch, D. (2007, May 15). Pentagon blocks 13 web sites from military computers. *New York Times*. Retrieved July 21, 2007 from http://www.nytimes.com/2007/05/15/washington/15block.html

Gajjala, R. (2007). Shifting frames: Race, ethnicity, and intercultural communication in online social networking and virtual work. In M. B. Hinner (Ed.), *The Role of Communication in Business Transactions and Relationships* (pp. 257–276). New York: Peter Lang.

Geidner, N. W., Flook, C. A., & Bell, M. W. (2007, April). *Masculinity and online social networks: Male self-identification on Facebook.com*. Paper presented at Eastern Communication Association 98th Annual Meeting, Providence, RI.

George, A. (2006, September 18). Living online: The end of privacy? *New Scientist*, 2569. Retrieved August 29, 2007 from http://www.newscientist.com/channel/tech/mg19125691.700-living-online-the-end-of-privacy.html

Goldberg, S. (2007, May 13). Analysis: Friendster is doing just fine. *Digital Media Wire*. Retrieved July 30, 2007 from http://www.dmwmedia.com/news/2007/05/14/analysis-friendster-is-doing-just-fine

Golder, S. A., Wilkinson, D., & Huberman, B. A. (2007, June). Rhythms of social interaction: Messaging within a massive online network. In C. Steinfield, B. Pentland, M. Ackerman, & N. Contractor (Eds.), *Proceedings of Third International Conference on Communities and Technologies* (pp. 41–66). London: Springer.

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. *Proceedings of WPES'05* (pp. 71–80). Alexandria, VA: ACM.

Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. *Information, Communication, & Society*, **8**(2), 125–147.

Heer, J., & boyd, d. (2005). Vizster: Visualizing online social networks. *Proceedings of Symposium on Information Visualization* (pp. 33–40). Minneapolis, MN: IEEE Press.

Herring, S. C., Paolillo, J. C., Ramos Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L. A., & Clark, B. (2007). Language networks on LiveJournal. *Proceedings of the Fortieth Hawai'i International Conference on System Sciences*. Los Alamitos, CA: IEEE Press.

Hewitt, A., & Forte, A. (2006, November). *Crossing boundaries: Identity management and student/faculty relationships on the Facebook*. Poster presented at CSCW, Banff, Alberta.

Hjorth, L., & Kim, H. (2005). Being there and being here: Gendered customising of mobile 3G practices through a case study in Seoul. *Convergence*, **11**(2), 49–55.

Hjorth, L., & Yuji, M. (in press). Logging on locality: A cross-cultural case study of virtual communities Mixi (Japan) and Mini-hompy (Korea). In B. Smaill (Ed.), *Youth and Media in the Asia Pacific*. Cambridge, UK: Cambridge University Press.

Hodge, M. J. (2006). The Fourth Amendment and privacy issues on the "new" Internet: Facebook.com and MySpace.com. *Southern Illinois University Law Journal*, **31**, 95–122.

Hogan, B. (in press). Analyzing social networks via the Internet. In N. Fielding, R. Lee, & G. Blank (Eds.), *Sage Handbook of Online Research Methods*. Thousand Oaks, CA: Sage.

H. R. 5319. (2006, May 9). *Deleting Online Predators Act of 2006*. H.R. 5319, 109th Congress. Retrieved July 21, 2007 from http://www.govtrack.us/congress/billtext.xpd?bill=h109-5319

Hsu, W. H., Lancaster, J., Paradesi, M. S. R., & Weninger, T. (2007). Structural link analysis from user profiles and friends networks: A feature construction approach. *Proceedings of ICWSM-2007* (pp. 75–80). Boulder, CO.

Jagatic, T., Johnson, N., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, **5**(10), 94–100.

Kopytoff, V. (2004, November 29). Google's orkut puzzles experts. San Francisco Chronicle. Retrieved July 30, 2007 from http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2004/11/29/BUGU9A0BH441.DTL

Kornblum, J., & Marklein, M. B. (2006, March 8). What you say online could haunt you. *USA Today*. Retrieved August 29, 2007 from http://www.usatoday.com/tech/news/internetprivacy/2006-03-08-facebook-myspace_x.htm

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. *Proceedings of 12th International Conference on Knowledge Discovery in Data Mining* (pp. 611–617). New York: ACM Press.

Lampe, C., Ellison, N., & Steinfield, C. (2006). A Face(book) in the crowd: Social searching vs. social browsing. *Proceedings of CSCW-2006* (pp. 167–170). New York: ACM Press.

Lampe, C., Ellison, N., & Steinfeld, C. (2007). A familiar Face(book): Profile elements as signals in an online social network. *Proceedings of Conference on Human Factors in Computing Systems* (pp. 435–444). New York: ACM Press.

Lenhart, A., & Madden, M. (2007, April 18). Teens, privacy, & online social networks. *Pew Internet and American Life Project Report*. Retrieved July 30, 2007 from http://www.pewinternet.org/pdfs/PIP_Teens_Privacy_SNS_Report_Final.pdf

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005) Geographic routing in social networks. *Proceedings of National Academy of Sciences*, **102**(33) 11,623–11,628.

Liu, H., Maes, P., & Davenport, G. (2006). Unraveling the taste fabric of social networks. *International Journal on Semantic Web and Information Systems*, **2**(1), 42–71.

Madhavan, N. (2007, July 6). India gets more Net Cool. *Hindustan Times*. Retrieved July 30, 2007 from http://www.hindustantimes.com/StoryPage/StoryPage.aspx?id=f2565bb8-663e-48c1-94ee-d99567577bdd

Marwick, A. (2005, October). *"I'm a lot more interesting than a Friendster profile:" Identity presentation, authenticity, and power in social networking services*. Paper presented at Internet Research 6.0, Chicago, IL.

Mazer, J. P., Murphy, R. E., & Simonds, C. J. (2007). I'll see you on "Facebook:" The effects of computer-mediated teacher self-disclosure on student motivation, affective learning, and classroom climate. *Communication Education*, **56**(1), 1–17.

McLeod, D. (2006, October 6). QQ Attracting eyeballs. *Financial Mail (South Africa)*, p. 36. Retrieved July 30, 2007 from *LexisNexis*.

National School Boards Association. (2007, July). *Creating and connecting: Research and guidelines on online social—and educational—networking*. Alexandria, VA. Retrieved September 23, 2007 from http://www.nsba.org/site/docs/41400/41340.pdf

Nyland, R., & Near, C. (2007, February). *Jesus is my friend: Religiosity as a mediating factor in Internet social networking use*. Paper presented at AEJMC Midwinter Conference, Reno, NV.

O'Shea, W. (2003, July 4-10). Six Degrees of sexual frustration: Connecting the dates with Friendster.com. *Village Voice*. Retrieved July 21, 2007 from http://www.villagevoice.com/news/0323,oshea, 44576, 1.html

Paolillo, J. C., & Wright, E. (2005). Social network analysis on the semantic web: Techniques and challenges for visualizing FOAF. In V. Geroimenko & C. Chen (Eds.), *Visualizing the Semantic Web* (pp. 229–242). Berlin: Springer.

Perkel, D. (in press). Copy and paste literacy? Literacy practices in the production of a MySpace profile. In K. Drotner, H. S. Jensen, & K. Schroeder (Eds.), *Informal Learning and Digital Media: Constructions, Contexts, Consequences*. Newcastle, UK: Cambridge Scholars Press.

Preibusch, S., Hoser, B., Gürses, S., & Berendt, B. (2007, June). Ubiquitous social networks—opportunities and challenges for privacy-aware user modelling. *Proceedings of Workshop on Data Mining for User Modeling. Corfu, Greece*. Retrieved October 20, 2007 from http://vasarely.wiwi.hu-berlin.de/DM.UM07/Proceedings/05-Preibusch.pdf

Recuero, R. (2005). O capital social em redes sociais na Internet. *Revista FAMECOS*, **28**, 88–106. Retrieved September 13, 2007 from http://www.pucrs.br/famecos/pos/revfamecos/28/raquelrecuero.pdf

S. 49. (2007, January 4). *Protecting Children in the 21st Century Act*. S. 49, 110th Congress. Retrieved July 30, 2007 from http://thomas.loc.gov/cgi-bin/query/F?c110:1:./temp/~c110dJQpcy:e445:

Shirky, C. (2003, May 13). People on page: YASNS… *Corante's Many-to-Many*. Retrieved July 21, 2007 from http://many.corante.com/archives/2003/05/12/people_on_page_yasns.php

Skog, D. (2005). Social interaction in virtual communities: The significance of technology. *International Journal of Web Based Communities*, **1**(4), 464–474.

Spertus, E., Sahami, M., & Buyukkokten, O. (2005). Evaluating similarity measures: A large-scale study in the orkut social network. *Proceedings of 11th International Conference on Knowledge Discovery in Data Mining* (pp. 678–684). New York: ACM Press.

Stutzman, F. (2006). An evaluation of identity-sharing behavior in social network communities. *Journal of the International Digital Media and Arts Association*, **3**(1), 10–18.

Sundén, J. (2003). *Material Virtualities*. New York: Peter Lang.

Walther, J. B., Van Der Heide, B., Kim, S. Y., & Westerman, D. (in press). The role of friends' appearance and behavior on evaluations of individuals on Facebook: Are we known by the company we keep? *Human Communication Research*.

Wellman, B. (1988). Structural analysis: From method and metaphor to theory and substance. In B. Wellman & S. D. Berkowitz (Eds.), *Social Structures: A Network Approach* (pp. 19–61). Cambridge, UK: Cambridge University Press.

Wolak, J., Mitchell, K., & Finkelhor, D. (2006). Online victimization of youth: Five years later. *Report from Crimes Against Children Research Center, University of New Hampshire*. Retrieved July 21, 2007 from http://www.unh.edu/ccrc/pdf/CV138.pdf

Zinman, A., & Donath, J. (2007, August). *Is Britney Spears spam?* Paper presented at the Fourth Conference on Email and Anti-Spam, Mountain View, CA.

About the Authors

danah m. boyd is a Ph.D. candidate in the School of Information at the University of California-Berkeley and a Fellow at the Harvard University Berkman Center for Internet and Society. Her research focuses on how people negotiate mediated contexts like social network sites for sociable purposes.
Address: 102 South Hall, Berkeley, CA 94720–4600, USA

Nicole B. Ellison is an assistant professor in the Department of Telecommunication, Information Studies, and Media at Michigan State University. Her research explores issues of self-presentation, relationship development, and identity in online environments such as weblogs, online dating sites, and social network sites.
Address: 403 Communication Arts and Sciences, East Lansing, MI 48824, USA

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait "Openness," prediction accuracy is close to the test–retest accuracy of a standard personality test. We give examples of associations between attributes and Likes and discuss implications for online personalization and privacy.

social networks | computational social science | machine learning | big data | data mining | psychological assessment

A growing proportion of human activities, such as social interactions, entertainment, shopping, and gathering information, are now mediated by digital services and devices. Such digitally mediated behaviors can easily be recorded and analyzed, fueling the emergence of computational social science (1) and new services such as personalized search engines, recommender systems (2), and targeted online marketing (3). However, the widespread availability of extensive records of individual behavior, together with the desire to learn more about customers and citizens, presents serious challenges related to privacy and data ownership (4, 5).

We distinguish between data that are actually recorded and information that can be statistically predicted from such records. People may choose not to reveal certain pieces of information about their lives, such as their sexual orientation or age, and yet this information might be predicted in a statistical sense from other aspects of their lives that they do reveal. For example, a major US retail network used customer shopping records to predict pregnancies of its female customers and send them well-timed and well-targeted offers (6). In some contexts, an unexpected flood of vouchers for prenatal vitamins and maternity clothing may be welcome, but it could also lead to a tragic outcome, e.g., by revealing (or incorrectly suggesting) a pregnancy of an unmarried woman to her family in a culture where this is unacceptable (7). As this example shows, predicting personal information to improve products, services, and targeting can also lead to dangerous invasions of privacy.

Predicting individual traits and attributes based on various cues, such as samples of written text (8), answers to a psychometric test (9), or the appearance of spaces people inhabit (10), has a long history. Human migration to digital environment renders it possible to base such predictions on digital records of human behavior. It has been shown that age, gender, occupation, education level, and even personality can be predicted from people's Web site browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or "Like") online content, such as photos, friends' status updates, Facebook pages of products, sports, musicians, books, restaurants, or popular Web sites. Likes represent a very generic class of digital records, similar to Web search queries, Web browsing histories, and credit card purchases. For example, observing users' Likes related to music provides similar information to observing records of songs listened to online, songs and artists searched for using a Web search engine, or subscriptions to related Twitter channels. In contrast to these other sources of information, Facebook Likes are unusual in that they are currently publicly available by default. However, those other digital records are still available to numerous parties (e.g., governments, developers of Web browsers, search engines, or Facebook applications), and, hence, similar predictions are unlikely to be limited to the Facebook environment.

The design of the study is presented in Fig. 1. We selected traits and attributes that reveal how accurate and potentially intrusive such a predictive analysis can be, including "sexual orientation," "ethnic origin," "political views," "religion," "personality," "intelligence," "satisfaction with life" (SWL), substance use ("alcohol," "drugs," "cigarettes"), "whether an individual's parents stayed together until the individual was 21 y old," and basic demographic attributes such as "age," "gender," "relationship status," and "size and density of the friendship network." Five Factor Model (9) personality scores ($n = 54,373$) were established using the International Personality Item Pool (IPIP) questionnaire with 20 items (25). Intelligence ($n = 1,350$) was measured using Raven's Standard Progressive Matrices (SPM) (26), and SWL ($n = 2,340$) was measured using the SWL Scale (27). Age ($n = 52,700$; average, $\mu = 25.6$; SD = 10), gender ($n = 57,505$; 62% female), relationship status ("single"/"in relationship"; $n = 46,027$; 49% single), political views ("Liberal"/"Conservative"; $n = 9,752$;

Author contributions: M.K. and T.G. designed research; M.K. and D.S. performed research; M.K. and T.G. analyzed data; and M.K., D.S., and T.G. wrote the paper.

Conflict of interest statement: D.S. received revenue as owner of the myPersonality Facebook application.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the myPersonality Project database (www.mypersonality.org/wiki).

[1]To whom correspondence should be addressed. E-mail: mk583@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218772110/-/DCSupplemental.

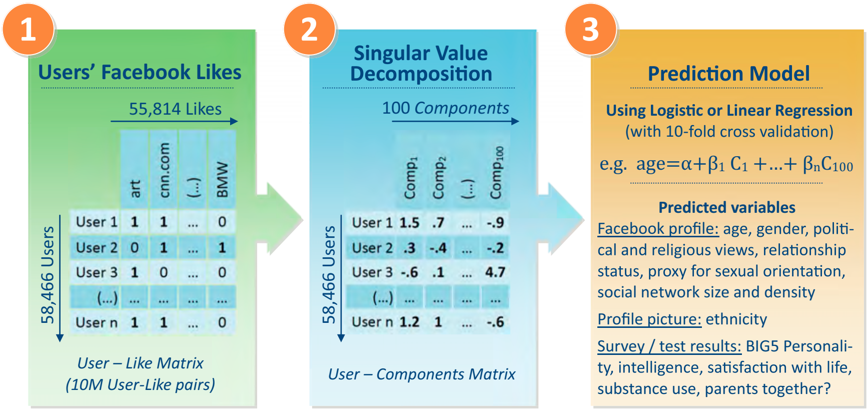

1
Users' Facebook Likes
55,814 Likes
art
cnn.com
(...)
BMW
58,466 Users
User 1 1 1 ... 0
User 2 0 1 ... 1
User 3 1 0 ... 0
(...) ... ...
User n 1 1 ... 0
User – Like Matrix
(10M User-Like pairs)
2
Singular Value Decomposition
100 Components
$Comp_1$
$Comp_2$
(...)
$Comp_{100}$
58,466 Users
User 1 1.5 .7 ... -.9
User 2 .3 -.4 ... -.2
User 3 -.6 .1 ... 4.7
(...) ... ...
User n 1.2 1 ... -.6
User – Components Matrix
3
Prediction Model
Using Logistic or Linear Regression
(with 10-fold cross validation)
e.g. $age=\alpha+\beta_1 C_1 +...+ \beta_n C_{100}$
Predicted variables
Facebook profile: age, gender, political and religious views, relationship status, proxy for sexual orientation, social network size and density
Profile picture: ethnicity
Survey / test results: BIG5 Personality, intelligence, satisfaction with life, substance use, parents together?

Fig. 1. The study is based on a sample of 58,466 volunteers from the United States, obtained through the myPersonality Facebook application (www.mypersonality.org/wiki), which included their Facebook profile information, a list of their Likes (n = 170 Likes per person on average), psychometric test scores, and survey information. Users and their Likes were represented as a sparse user–Like matrix, the entries of which were set to 1 if there existed an association between a user and a Like and 0 otherwise. The dimensionality of the user–Like matrix was reduced using singular-value decomposition (SVD) (24). Numeric variables such as age or intelligence were predicted using a linear regression model, whereas dichotomous variables such as gender or sexual orientation were predicted using logistic regression. In both cases, we applied 10-fold cross-validation and used the k = 100 top SVD components. For sexual orientation, parents' relationship status, and drug consumption only k = 30 top SVD components were used because of the smaller number of users for which this information was available.

65% Liberal), religion ("Muslim"/"Christian"; n = 18,833; 90% Christian), and the Facebook social network information [n = 17,601; median size, $\tilde{X}$ = 204; interquartile range (IQR), 206; median density, $\tilde{X}$ = 0.03; IQR, 0.03] were obtained from users' Facebook profiles. Users' consumption of alcohol (n = 1,196; 50% drink), drugs (n = 856; 21% take drugs), and cigarettes (n = 1211; 30% smoke) and whether a user's parents stayed together until the user was 21 y old (n = 766; 56% stayed together) were recorded using online surveys. Visual inspection of profile pictures was used to assign ethnic origin to a randomly selected subsample of users (n = 7,000; 73% Caucasian; 14% African American; 13% others). Sexual orientation was assigned using the Facebook profile "Interested in" field; users interested only in others of the same sex were labeled as homosexual (4.3% males; 2.4% females), whereas those interested in users of the opposite gender were labeled as heterosexual.

Results

Prediction of Dichotomous Variables. Fig. 2 shows the prediction accuracy of dichotomous variables expressed in terms of the area under the receiver-operating characteristic curve (AUC), which is equivalent to the probability of correctly classifying two randomly selected users one from each class (e.g., male and female). The highest accuracy was achieved for ethnic origin and gender. African Americans and Caucasian Americans were correctly classified in 95% of cases, and males and females were correctly classified in 93% of cases, suggesting that patterns of online behavior as expressed by Likes significantly differ between those groups allowing for nearly perfect classification.

Christians and Muslims were correctly classified in 82% of cases, and similar results were achieved for Democrats and Republicans (85%). Sexual orientation was easier to distinguish among males (88%) than females (75%), which may suggest a wider behavioral divide (as observed from online behavior) between hetero- and homosexual males.

Good prediction accuracy was achieved for relationship status and substance use (between 65% and 73%). The relatively lower accuracy for relationship status may be explained by its temporal variability compared with other dichotomous variables (e.g., gender or sexual orientation).

The model's accuracy was lowest (60%) when inferring whether users' parents stayed together or separated before users were 21 y old. Although it is known that parental divorce does have long-term effects on young adults' well-being (28), it is remarkable that this is detectable through their Facebook Likes. Individuals with parents who separated have a higher probability of liking statements preoccupied with relationships, such as "If I'm with you then I'm with you I don't want anybody else" (Table S1).

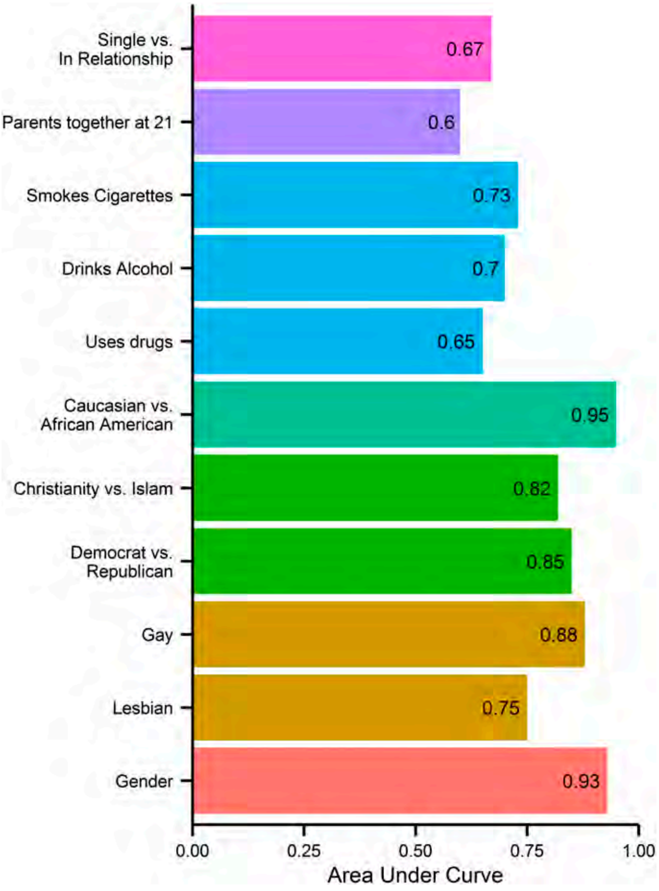


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

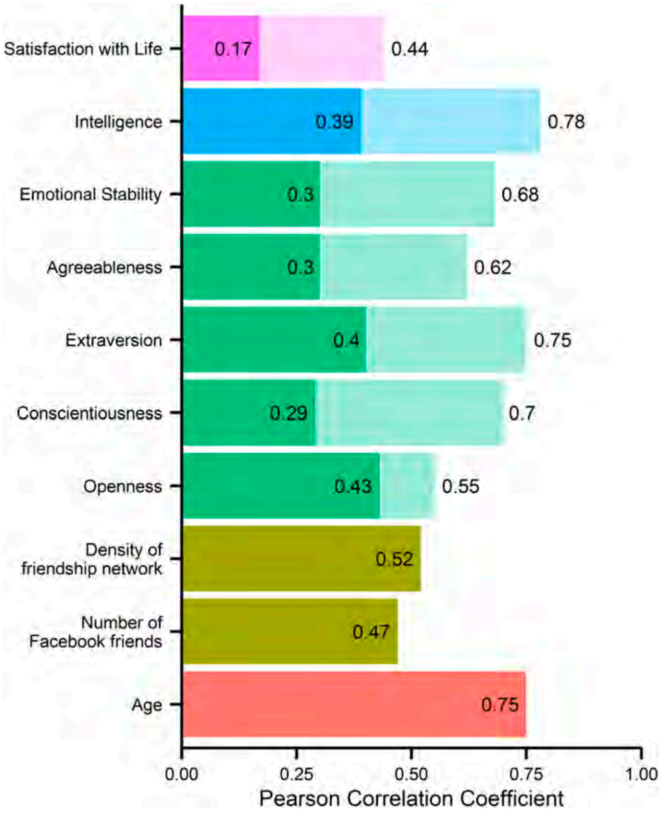


Fig. 3. Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values; all correlations are significant at the $P < 0.001$ level. The transparent bars indicate the questionnaire's baseline accuracy, expressed in terms of test–retest reliability.

Prediction of Numeric Variables. Fig. 3 presents the accuracy of predicting numeric variables as expressed by the Pearson product–moment correlation coefficient between the actual and predicted values. The highest correlation was obtained for age ($r = 0.75$), followed by density ($r = 0.52$) and size ($r = 0.47$) of the Facebook friendship network. Closely following were the personality traits of "Openness" ($r = 0.43$), "Extraversion" ($r = 0.40$), and "Intelligence" ($r = 0.39$). The remaining personality traits and SWL were predicted with somewhat lower accuracy ($r = 0.17$ to 0.30).

Psychological traits are examples of latent traits (i.e., traits that cannot be measured directly). As a consequence, their values can only be measured approximately, for example, by evaluating responses to questionnaires. The transparent bars presented in Fig. 3 indicate the accuracy of the questionnaires used as expressed by their test-retest reliabilities (Pearson product–moment correlation between the questionnaire scores obtained by the same respondent at two points in time). The correlation between the predicted and actual Openness score ($r = 0.43$) was very close to the test–retest reliability for Openness ($r = 0.50$). This indicates that for the Openness trait, observation of the user's Likes is roughly as informative as using their personality test score itself. For the remaining traits, prediction accuracies correspond to roughly half the questionnaire's test-retest reliabilities.

The relatively lower prediction accuracy for SWL ($r = 0.17$) may be attributable to the difficulty of separating long-term happiness (29) from mood swings, which vary over time. Thus, although the SWL score includes variability attributable to mood, users' Likes accrue over a longer period and, so, may be suitable only for predicting long-term happiness.

Amount of Data Available and Prediction Accuracy. The results presented so far rely on individuals for which between one and 700 Likes were available. The median number of Likes was 68 per individual (IQR, 152). Therefore, what is the expected accuracy given a random individual and how does prediction accuracy change with the number of observed Likes? Using a subsample ($n = 500$) of users for whom at least 300 Likes were available, we ran predictive models based on randomly selected subsets of $n = 1, 2, \ldots, 300$ Likes. The results presented in Fig. 4 show that even knowing a single random Like for a given user can result in nonnegligible prediction accuracy. Knowing further Likes increases the accuracy but with diminishing returns from each additional piece of information.

Predictive Power of Likes. Individual traits and attributes can be predicted to a high degree of accuracy based on records of users' Likes. Table S1 presents a sample of highly predictive Likes related to each of the attributes. For example, the best predictors of high intelligence include "Thunderstorms," "The Colbert Report," "Science," and "Curly Fries," whereas low intelligence was indicated by "Sephora," "I Love Being A Mom," "Harley Davidson," and "Lady Antebellum." Good predictors of male homosexuality included "No H8 Campaign," "Mac Cosmetics," and "Wicked The Musical," whereas strong predictors of male heterosexuality included "Wu-Tang Clan," "Shaq," and "Being Confused After Waking Up From Naps." Although some of the Likes clearly relate to their predicted attribute, as in the case of No H8 Campaign and homosexuality, other pairs are more elusive; there is no obvious connection between Curly Fries and high intelligence.

Moreover, note that few users were associated with Likes explicitly revealing their attributes. For example, less than 5% of users labeled as gay were connected with explicitly gay groups, such as No H8 Campaign, "Being Gay," "Gay Marriage," "I love Being

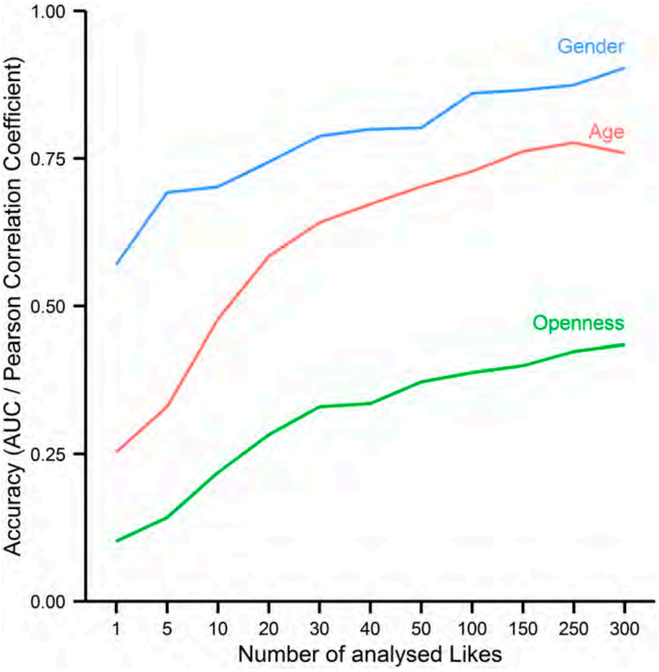


Fig. 4. Accuracy of selected predictions as a function of the number of available Likes. Accuracy is expressed as AUC (gender) and Pearson's correlation coefficient (age and Openness). About 50% of users in this sample had at least 100 Likes and about 20% had at least 250 Likes. Note, that for gender (dichotomous variable) the random guessing baseline corresponds to an AUC = 0.50.

Gay," "We Didn't Choose To Be Gay We Were Chosen." Consequently, predictions rely on less informative but more popular Likes, such as "Britney Spears" or "Desperate Housewives" (both moderately indicative of being gay).

This is further illustrated in Fig. S1, which shows the average levels of personality traits and age for several popular Likes. Each Like attracts users with a different average personality and demographic profile and, thus, can be used to predict those attributes. For example, users who liked the "Hello Kitty" brand tended to be high on Openness and low on "Conscientiousness," "Agreeableness," and "Emotional Stability." They were also more likely to have Democratic political views and to be of African-American origin, predominantly Christian, and slightly below average age. The same Likes were used to create Fig. S2, presenting their relative popularity in four groups: Democrats, Christians, Homosexuals, and African-American individuals. For example, although liking "Barack Obama" is clearly related to being a Democrat, it is also relatively popular among Christians, African Americans, and Homosexual individuals.

Conclusions

We show that a wide variety of people's personal attributes, ranging from sexual orientation to intelligence, can be automatically and accurately inferred using their Facebook Likes. Similarity between Facebook Likes and other widespread kinds of digital records, such as browsing histories, search queries, or purchase histories suggests that the potential to reveal users' attributes is unlikely to be limited to Likes. Moreover, the wide variety of attributes predicted in this study indicates that, given appropriate training data, it may be possible to reveal other attributes as well.

Predicting users' individual attributes and preferences can be used to improve numerous products and services. For instance, digital systems and devices (such as online stores or cars) could be designed to adjust their behavior to best fit each user's inferred profile (30). Also, the relevance of marketing and product recommendations could be improved by adding psychological dimensions to current user models. For example, online insurance advertisements might emphasize security when facing emotionally unstable (neurotic) users but stress potential threats when dealing with emotionally stable ones. Moreover, digital records of behavior may provide a convenient and reliable way to measure psychological traits. Automated assessment based on large samples of behavior may not only be more accurate and less prone to cheating and misrepresentation but may also permit assessment across time to detect trends. Moreover, inference based on observations of digitally recorded behavior may open new doors for research in human psychology.

On the other hand, the predictability of individual attributes from digital records of behavior may have considerable negative implications, because it can easily be applied to large numbers of people without obtaining their individual consent and without them noticing. Commercial companies, governmental institutions, or even one's Facebook friends could use software to infer attributes such as intelligence, sexual orientation, or political views that an individual may not have intended to share. One can imagine situations in which such predictions, even if incorrect, could pose a threat to an individual's well-being, freedom, or even life. Importantly, given the ever-increasing amount of digital traces people leave behind, it becomes difficult for individuals to control which of their attributes are being revealed. For example, merely avoiding explicitly homosexual content may be insufficient to prevent others from discovering one's sexual orientation.

There is a risk that the growing awareness of digital exposure may negatively affect people's experience of digital technologies, decrease their trust in online services, or even completely deter them from using digital technology. It is our hope, however, that the trust and goodwill among parties interacting in the digital environment can be maintained by providing users with transparency and control over their information, leading to an individually controlled balance between the promises and perils of the Digital Age.

ACKNOWLEDGMENTS. We thank Yoram Bachrach, Alan Blackwell, George Danezis, Stephen Emmott, David Good, Peter Key, Emre Kiciman, Pushmeet Kohli, Drew Purves, Jason Rentfrow, John Rust, and Duncan Watts for discussions about the topic of this study, as well as for comments on the manuscript. Demonstration of personality prediction based on individuals' Likes is available at http://www.youarewhatyoulike.com. M.K. received funding from Boeing Corporation and from Microsoft Research.

1. Lazer D, et al. (2009) Computational social science. *Science* 323(5915):721–723.
2. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.
3. Chen Y, Pavlov D, Canny JF (2009) Large-scale behavioral targeting. *International Conference on Knowledge Discovery and Data Mining*, pp 209–218.
4. Butler D (2007) Data sharing threatens privacy. *Nature* 449(7163):644–645.
5. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, pp 111–125.
6. Duhigg C (2012) *The Power of Habit: Why We Do What We Do in Life and Business* (Random House, New York).
7. İnce HO, Yarali A, Özsel D (2009) Customary killings in Turkey and Turkish modernization. *Middle East Stud* 45(4):537–551.
8. Fast LA, Funder DC (2008) Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *J Pers Soc Psychol* 94(2):334–346.
9. Costa PT, McCrae RR (1992) *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Manual* (Psychological Assessment Resources, Odessa, FL).
10. Gosling SD, Ko SJ, Mannarelli T, Morris ME (2002) A room with a cue: Personality judgments based on offices and bedrooms. *J Pers Soc Psychol* 82(3):379–398.
11. Hu J, Zeng H-J, Li H, Niu C, Chen Z (2007) Demographic prediction based on user's browsing behavior. *International World Wide Web Conference*, pp 151–160.
12. Murray D, Durrell K (1999) Inferring demographic attributes of anonymous Internet users. *Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, eds Masand BM, Spiliopoulou M (Springer, London), pp 7–20.
13. De Bock K, Van Den Poel D (2010) Predicting website audience demographics for Web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae* 98(1):49–70.
14. Goel S, Hofman JM, Sirer MI (2012) Who does what on the Web: Studying Web browsing behavior at scale. *International Conference on Weblogs and Social Media*, pp 130–137.
15. Kosinski M, Kohli P, Stillwell DJ, Bachrach Y, Graepel T (2012) Personality and website choice. *ACM Web Science Conference*, pp 251–254.
16. Marcus B, Machilek F, Schütz A (2006) Personality in cyberspace: Personal Web sites as media for personality expressions and impressions. *J Pers Soc Psychol* 90(6):1014–1031.
17. Rentfrow PJ, Gosling SD (2003) The do re mi's of everyday life: The structure and personality correlates of music preferences. *J Pers Soc Psychol* 84(6):1236–1256.
18. Quercia D, Lambiotte R, Kosinski M, Stillwell D, Crowcroft J (2012) The Personality of popular Facebook users. *ACM Conference on Computer Supported Cooperative Work*. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp 955–964.
19. Bachrach Y, Kohli P, Graepel T, Stillwell DJ, Kosinski M (2012) Personality and patterns of Facebook usage. *ACM Web Science Conference*. Proceedings of the ACM Web Science Conference, pp 36–44.
20. Quercia D, Kosinski M, Stillwell DJ, Crowcroft J (2011) Our Twitter profiles, our selves: Predicting personality with Twitter. *IEEE International Conference on Social Computing*. Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp 180–185.
21. Golbeck J, Robles C, Edmondson M, Turner K (2011) Predicting personality from Twitter. *IEEE International Conference on Social Computing*, pp 149–156.
22. Golbeck J, Robles C, Turner K (2011) Predicting personality with social media. *Conference on Human Factors in Computing Systems*, pp 253–262.
23. Jernigan C, Mistree BF (2009) Gaydar: Facebook friendships expose sexual orientation. *First Monday* 14(10).
24. Golub GH, Kahan W (1965) Calculating the singular values and pseudo-inverse of a matrix. *J Soc Ind Appl Math* 2(2):205–224.
25. Goldberg LR, et al. (2006) The international personality item pool and the future of public-domain personality measures. *J Res Pers* 40(1):84–96.
26. Raven JC (2000) The Raven's progressive matrices: Change and stability over culture and time. *Cognit Psychol* 41(1):1–48.
27. Diener E, Emmons RA, Larsen RJ, Griffin S (1985) The satisfaction with life scale. *J Pers Assess* 49(1):71–75.
28. Musick K, Meier A (2010) Are both parents always better than one? Parental conflict and young adult well-being. *Soc Sci Res* 39(5):814–830.
29. Schimmack U, Diener E, Oishi S (2002) Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. *J Pers* 70(3):345–384.
30. Nass C, Lee KM (2000) Does computer-generated speech manifest personality? An experimental test of similarity-attraction. *J Exp Psychol* 7(3):171–181.

Studying Facebook via Data Extraction: The Netvizz Application

Bernhard Rieder
University of Amsterdam
Turfdraagsterpad 9
1012TX Amsterdam
rieder@uva.nl

ABSTRACT

This paper describes Netvizz, a data collection and extraction application that allows researchers to export data in standard file formats from different sections of the Facebook social networking service. Friendship networks, groups, and pages can thus be analyzed quantitatively and qualitatively with regards to demographical, post-demographical, and relational characteristics. The paper provides an overview over analytical directions opened up by the data made available, discusses platform specific aspects of data extraction via the official Application Programming Interface, and briefly engages the difficult ethical considerations attached to this type of research.

Author Keywords

research tool, social networking services, Facebook, data extraction, social network analysis, media studies

ACM Classification Keywords

J.4 Social and Behavioral Sciences

INTRODUCTION

In October 2012, Facebook announced that it had reached the symbolic number of one billion monthly active users. [4] This arguably makes it one of the biggest media organizations in the history of humankind, contested only by Google's collection of services in terms of daily worldwide audience size and engagement. Traditional corporations dwarf these massive Internet companies when it comes to the size of their workforce – Facebook employed a mere 4500 people at the end of 2012 – but the sheer number of "[p]eople [who] use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them" [4] is simply gigantic. It is no wonder, then, that researchers from many areas of the human and social sciences have moved quickly to study the platform: a recent review article [19] identified 412 peer-reviewed research papers that follow empirical approaches, not counting the numerous publications employing conceptual and/or critical approaches. While traditional empirical methods such as interviews, experiments, and observations are widely used, a growing number of studies rely on what the authors call "data crawling", i.e. "gleaning information about users from their profiles without their active participation" [19]. This paper presents a software tool, Netvizz, designed to facilitate this latter approach.

Research methods using software to capture, produce, or repurpose digital data in order to investigate different aspects of the Internet have been used for well over a decade. Datasets can be exploited to analyze complex social and cultural phenomena and *digital methods* [12] have a number of advantages compared to traditional ones: advantages concerning cost, speed, exhaustiveness, detail, and so forth, but also related to the rich contextualization afforded by the close association between data and the properties of the *media* (technologies, platforms, tools, websites, etc.) they are connected with; data crawling necessarily engages these media through the specifics of their technical and functional structure and therefore produces data that can provide detailed views of the systems and the use practices they host. The study of social networking services (SNS) like Facebook, however, introduces a number of challenges and considerations that makes the scholarly investigation of these services, their users, and the various forms of content they hold significantly different from the study of the open Web. This paper discusses some of the possibilities and difficulties with the data crawling approach applied to Facebook and introduces a tool that allows researchers to generate data files in standard formats for different sections of the Facebook social networking service without having to resort to manual collecting or custom programming. I will first introduce some of the approaches to data extraction on SNS, in order to situate the proposed tool. I will then introduce the Netvizz application and provide a number of short examples for the type of analysis it makes possible. Before concluding, I will discuss two further aspects that are particularly relevant to the matter at hand: research via Application Programming Interfaces (API) and the question of privacy and research ethics. While this paper contains technical descriptions, it is written from a media studies perspective and therefore focuses on aspects most relevant to media scholars.

WebSci'13, May 2–4, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1889-1....$10.00.

STUDYING FACEBOOK THROUGH DATA EXTRACTION

The study of Internet platforms via data extraction has seen fast growth over the last two decades and the recent excitement around the concept of *big data* seems to have added additional momentum to efforts going into this direction. [9] For researchers from the humanities and social sciences, the possibility to analyze the expressions and behavioral traces from sometimes very large numbers of individuals or groups using these platforms can provide valuable insights into the arrays of meaning and practice that emerge and manifest themselves online. Besides merely shedding light on a "virtual" space, supposedly separate from "real life", the Internet can be considered as "a source of data about society and culture" [12] at large. The promise of producing *observational* data, i.e. data that documents what people do rather than what they say they do, without having to manually protocol behavior, expressions, and interactions is particularly enticing to researchers. SNS in general, and the gigantic Facebook platform in particular, can be likened, on a certain level, to observational devices or even to experimental designs: the "captured" data are closely related to meticulously constructed technical and visual forms – functionalities, interfaces, data structures, and so forth – that function as "grammars of action" [1], enabling and directing activities in distinct ways by providing and circumscribing possibilities for action and expression. Even if the design of this large-scale social experiment is specified neither by nor for social scientists and humanists, the delineated and parametered spaces provided by SNS confer a controlled frame of reference to gathered data. No wonder that Cameron Marlow, one of the research scientists working at Facebook considers the service to be "the world's most powerful instrument for studying human society" [16]. In order to better understand how such data can be gathered, a short overview of existing approaches is indispensable.

Existing Approaches

The already mentioned review paper [19] distinguishes five categories of empirical Facebook research: descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions, and privacy and information disclosure. It is not difficult to see how approaches gathering data from or through the platform can be useful for each of these areas of investigation. The question, then, is what data can actually be accessed and how this is to be done, considering that the particular technique chosen has important repercussions for the scope of what can be realistically acquired.

One can largely distinguish two general orientations when it comes to collecting digital data from SNS through software-based tools: first, researchers can recruit participants, through Facebook itself or from the outside, and gather data by asking them to fill out questionnaires, often via so called Facebook applications[1]. [11] While this method certainly differs from traditional ways of recruiting participants in terms of logistics and sampling procedures, it is not fundamentally different from online surveying in general.[2] Second, data can be retrieved in various ways from the pools of information that the Facebook platform *already* collects as part of its general operation. This latter approach, which is the focus of this paper, is fueled by data derived from both sides of the distinction Schäfer makes between "implicit and explicit participation" [14], referring to the difference between information and content deliberately provided by users, e.g. by filling out their profiles, and the data collected and produced by logging users' actions in sometimes minute detail. While Facebook members share content, write messages, and curate their profiles, they also click, watch, read, navigate, and so forth, thereby providing additional data points that are stored and analyzed. Because these activities revolve around elements that have cultural significance – liking a page of a political party is more than "clicking" – these data are not simply behavioral, but allow for deeper probing into *culture*. For research scholars, there are three ways by which to gain access to these data, with significant differences between approaches in terms of technical requirements and institutional positioning:

Direct database access to the company's servers is reserved to in-house researchers or cooperation between a SNS and a research institution. [17] Certain companies also make data "donations", for example Twitter deciding to transfer its complete archive to the Library of Congress, albeit with a significant delay. The data made accessible in these ways are generally very large and well structured, but often anonymized or aggregated. Partnering with a platform owner is certainly the only (legal) way to gain access to *all* collected data, at least in theory.

Access through sanctioned APIs makes use of the machine interfaces provided by many Web 2.0 services to third-party developers with the objective of stimulating application development and integration with other services in order to provide additional functionality and utility to users. These interfaces also provide well-structured data, but are generally limited in terms of which data, how much data, and how often data can be retrieved. Conditions can vary significantly between services: in contrast to Twitter, for example, Facebook is quite restrictive in terms of what data can be accessed, but imposes few limits on request frequency. Companies also retain the right to modify or close their data interfaces, which can lead to substantial problems for researchers.

[1] A Facebook application is a program that is provided by a third-party but integrates directly into the platform.

[2] One should note that studies using questionnaires on Facebook often access profile data as well.

User interface crawling can be done manually, but usually employs so-called *bots* or *spiders* that read the HTML documents used to provide graphical interfaces to users, either directly at the HTTP protocol level or via browser automation from the rendered DOM.[3] [8] These techniques can circumvent the limitations of APIs, but often at the price of technical and legal uncertainties if a platform provider's permission is not explicitly granted. In the case of Facebook, bot detection mechanisms are in place and suspicious activity can quickly lead to account suspension.

If performed on a large scale, all of these approaches require either custom programming or considerable amounts of manual work. The focus points and requirements for research and teaching do, however, bear marks of resemblance and Facebook itself is designed around a limited number of functionalities or "spaces". One can therefore argue that general-purpose tools may be envisioned that provide utility to a variety of research projects and interests. Several such *data extractors* targeting Facebook have been developed over the last years, invariably using sanctioned APIs for data gathering. These tools generally export data in common formats and they focus on specific sections of the platform – partly by choice, partly due to limitations imposed by the platform itself. Their goals are also similar: to lower the technical and logistical requirements for empirical research via data analysis in order to further the ability of researchers to study a medium that unites over a billion users in a system that is essentially conceived as a walled garden. In what follows, I describe the Netvizz application[4], a tool designed to help research scholars in extracting data from Facebook.

Similar Work

The enormous success of Facebook has prompted the emergence of a large number of analytics tools for marketing purposes, which often focus on *pages*, the section of Facebook that brand communication and consumer relations rely on, due to their public showcase character. Because these tools are generally built for monitoring marketing campaigns, they target page *owners* rather than researchers interested in studying a page. For this reasons – and the sheer number of tools available – I will leave these applications to the side.

There are, however, two tools that function as general-purpose data extractors for researchers studying Facebook. *NameGenWeb*[5] originated at the Oxford Internet Institute and provides the possibility of exporting a user's friendship network, i.e. all of the user's friends, the friendship connections between them, and a wide array of variables for each user account extracted. Another application, the *Social Network Importe*r[6], a plug-in for the *NodeXL* network analysis and visualization toolkit developed by an international group of scholars, provides similar functionality for downloading personal networks, but also a means to extract extensive data from Facebook pages, including monopartite[7] networks for users and posts, based on co-like or co-comment activities, and bipartite networks combining the two in a single graph. One should also mention Wolfram Alpha's "Facebook report"[8] in this context: while it does not make raw data available, and therefore limits in-depth analytics using statistical or graph theoretical approaches, the tool provides a large number of analytical views on personal networks.

The Netvizz application provides "raw" data for both personal networks and pages, but provides data *perspectives* not available in other tools, e.g. comment text extraction; it also provides data for groups, a third functional space on Facebook. Running as a Web application, Netvizz does not require the use of Microsoft Excel on Windows like *NodeXL* and thereby further lowers the threshold to engagement with Facebook's rich data pools. The next section will introduce the application and its different data outputs in more detail.

THE NETVIZZ APPLICATION

The Netvizz application was initially developed by the author in 2009 as a practical attempt to study Facebook's API as a new media object in its own right[9] and to gauge the potential of using natively digital methods [12] to study SNS. Because of the positive reactions and high uptake, the application was developed into a veritable data extractor that provides outputs for different sections of Facebook in standard formats.[10] Before introducing the different

[3] The latter approach has become more common due to the fact that sites are increasingly using programming languages (mostly JavaScript) to assemble pages client-side rather than sending finished documents described in a markup language (mostly HTML).

[4] https://apps.facebook.com/netvizz/

[5] https://apps.facebook.com/namegenweb/

[6] http://socialnetimporter.codeplex.com/

[7] Monopartite graphs contain nodes that are all of the same kind (e.g. users). Bipartite graphs include two types of nodes (e.g. users and posts), and so forth.

[8] http://www.wolframalpha.com/facebook/

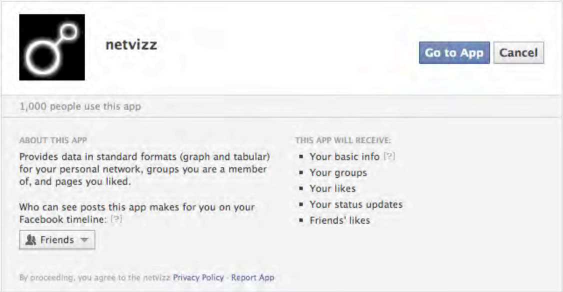
[9] APIs as *objects* of research for new media scholars are only slowly coming into view, despite their importance for the Web as data ecosystem. A separate publication will detail empirical approaches to studying APIs from a critical media studies perspective.

[10] Data formats were chosen for their generality and simplicity. Network outputs use the GDF format introduced with the GUESS graph analysis toolkit. Tabular outputs use a simple tab separated format that can be opened in virtually all spreadsheet applications and statistical packages.

features, it is necessary to briefly discuss the Facebook API and those characteristics that are relevant to research procedures and data quality.

Data Access via the Facebook API

As indicated above, Netvizz is a simple Facebook application written in PHP that runs on a server provided by the Digital Methods Initiative[11]. It is part of Facebook's app directory and can be found by typing the name into the platform's main search box. Like any other Facebook application, it requires users to log in with an existing Facebook account to be able to access any data at all.



1,000 people use this app

ABOUT THIS APP

Provides data in standard formats (graph and tabular) for your personal network, groups you are a member of, and pages you liked.

Who can see posts this app makes for you on your Facebook timeline: [?]

THIS APP WILL RECEIVE:

- Your basic info [?]
- Your groups
- Your likes
- Your status updates
- Friends' likes

By proceeding, you agree to the netvizz Privacy Policy · Report App

Figure 1. The Netvizz app permission request page.

A vast SNS that deals with intimate and potentially sensitive matters is likely to implement rather strict privacy policies and this is – to a certain extent – also the case with Facebook. The construction of the Facebook API reflects these concerns in at last four ways that are significant here:

First, every probe into the data pool is "signed" with the credentials of a Facebook user whose actual status on the platform defines the scope of which data can be accessed. For example, detailed user data can generally only be extracted from accounts a user is friends with and one has to be a member of a group to extract any data from it.

Second, users' privacy settings play a role in what data can be exported. If one user excludes another from seeing certain elements on his or her profile, an application operating with the latter's credentials will also be blocked from accessing those elements.

Third, every application is required to explicitly ask for permission to access different data elements.[12] These requests are displayed to the user when she first uses the application. Figure 1 shows the permission dialogue for the Netvizz application. While these permissions have to be given for the application to work, users can limit the data made available to applications used by their friends in their preferences.

Fourth, certain elements that are visible on the level of the user interface are not available through the API. The user view count displayed on each post in a group, for example, is (currently) not retrievable and certain data elements, such as friends' email addresses, are equally off limits by design.

While we can expect scholars using the Netvizz application to grant all the permissions[13] it asks for – it will simply not work otherwise – users' privacy settings are indeed relevant when it comes to interpreting the retrieved data: from a technical perspective, it is not possible to know whether an empty field is empty because the user has not filled in the specific data or because the privacy settings prohibit access. This must be taken into account when making assumptions on the basis of missing data. User profile data, in particular, should be handled with prudence. Other data, such as page engagement and friendship connections in personal networks and groups, can be considered robust, however.

Overview

The Netvizz application currently extracts data from three different sections of the Facebook platform:

Personal networks are considered in two different ways. First, the friendship network feature provides a simple undirected graph file where the friends of the logged user are nodes and friendship connections edges. Sex, interface language, and a ranking based on the account creation date[14] are provided for each user and counts for posts and likes can be requested as an option. Friendship networks often cluster around significant places in a user's life, e.g. geographies or institutions such as high school, university, workplaces, clubs, and so forth. Second, a bipartite "like network" can be generated that formalizes both users and liked entities (all elements already represented in Facebook's Open Graph[15] are extracted) as nodes, a user liking a page generating an edge. This network, examined via a graph analysis toolkit, will arrange both users and liked objects around cultural affinity patterns, foregrounding *post-demographic* [13] variables.

Groups can be explored in a similar fashion as friendship networks, although the API currently limits the number of users one can retrieve from a group to 5000. For larger groups, a random subset of users is provided. A second

[11] https://www.digitalmethods.net

[12] For details concerning the permission structure refer to: http://developers.facebook.com/docs/reference/login/

[13] The Netvizz application does not store or aggregate any of the extracted data in a database and the generated files are deleted in regular intervals.

[14] The unique identifiers for accounts on Facebook are numbered consecutively, which means that the lower the number, the older the account. Netvizz simply adds a ranking to the output that orders accounts by their age.

[15] For more information on how Facebook represents entities in the *Open Graph* concept modeling system, see: https://developers.facebook.com/docs/concepts/opengraph/.

feature also provides a *social* graph, but one that is based on interactions between group members through the posts sent to a group. If one user likes or comments on another user's post, a directed edge between the two users is created, each interaction adding weight to the edge.

Pages are represented as a bipartite network, with both posts (up to 999 latest posts) and users as nodes. If a user comments on or likes a post, a directed edge between user and post is created. This way, one can not only detect the most active users, but also identify the posts that produced the highest amount of engagement. The latter data are also provided in a tabular data file, ready for statistical analysis. To make content analysis easier, a third file containing user comments, grouped per post, is generated. The application allows selecting whether posts made by users should be included, in addition to posts made by the page owner.

ANALYTICAL DIRECTIONS

The two types of data files provided by Netvizz – network files and tabular files – already indicate basic directions for analytical approaches, the former allowing for the application of concepts and methods from Social Network Analysis [15] and Network Science [18], while the latter points towards more traditional statistical techniques. Before describing analytical approaches in more detail, a short comment on modes of analysis – and in particular visualization – is in order.

Analysis and Visualization

One of the reasons for choosing simple and common file formats for outputs in Netvizz was the need to compensate for the lack of an actual visual and analytical interface in the application itself. There are, indeed, a number of Facebook applications available that produce direct visual representations, generally of personal networks, which greatly facilitates the initial encounter with the data in question for researchers with little or no training in quantitative research. Because these tools are mostly visualization widgets that do not target researchers and offer little to no analytical methodology beyond the visual display itself, one of the initial intentions was to design Netvizz as a bridge between Facebook data and the various network analysis toolkits available today, such as GUESS[16], Pajek[17] or the very easy to use Gephi[18]. The last program, in particular, must be credited with significant lowering the threshold to working with network analysis and visualization. Netvizz voluntarily inscribes itself in a movement, epitomized by tools such as gephi and the work of the Amsterdam-based Digital Methods Initiative[19] and other groups, that aims at bringing data-driven analysis to a wider audience and, specifically, to an audience that includes those regions of the social sciences and humanities that have been shunning quantitative and computational methods because of the epistemological and methodological commitments often associated with quantification and formalization. Lowering the threshold to using computer-based analytical methods is therefore not simply a service to long-time practitioners, but an attempt to see in what way and how far these methods can be useful in contexts where the dominant "styles of reasoning" [7] are based on interpretation, argumentation, and speculation, and build on conceptualizations of human beings and their practices that simply cannot be formalized as easily as theoretical frameworks like behaviorism or social exchange theory.

In this context, visualization has been presented as a means to profit from the analytical capacities afforded by software without having to invest years into the acquisition of skills in statistics or graph theory. While the data provided by Netvizz can certainly be used to calculate correlation coefficients as well as network metrics, focus was put on facilitating analysis through visualization. There is, however, no need to juxtaposition mathematical and visual forms of analysis; as Figure 2 demonstrates, the latter can not only help in communicating the results provided by the former, but adds a way of relating to the data that can provide a significant epistemic surplus.

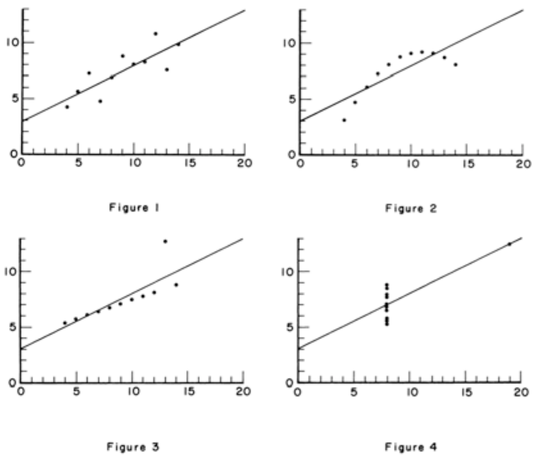


Figure 2. Four scatter-plots from [2]. They have identical values for number of observations, mean of the x's, mean of the y's, regression coefficient of y on x, equation of regression line, sum of squares of x, regression sum of squares, residual sum of squares of y, estimated standard error of bi, and multiple r2. Yet, the differences between the dataset are strikingly obvious to our eyes. Anscombe uses this example to make an argument for the usefulness of visualization in statistics beyond the communication to a larger audience.

Independently of its application to actual empirical analysis of Facebook data, Netvizz should thus be considered a pedagogical tool that can help in getting started with quantitative methodology, network analysis, and the

[16] http://graphexploration.cond.org

[17] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[18] https://gephi.org

[19] https://digitalmethods.net

required software. While one could argue that network visualizations are images and therefore intuitively accessible and "readable", there are also arguments that point into the opposite direction. It is easy to show how different graph layout algorithms highlight particular properties of a network and familiarity with a dataset can go far in helping novice users understand what is actually happening when they use software to work with graph data. Because many people are intimately familiar with their Facebook networks, they can more easily see what the software does, and what kind of epistemic surplus one can potentially derive from network analysis.

Analytical Perspectives

In actual research settings, Netvizz can provide data relevant to many different approaches and research questions. One can also consider different embeddings in the logistics of research projects: it is imaginable that a study recruits users to investigate patterns in social relations, but instead of asking them for access to their accounts, they encourage them to run the Netvizz application from their profile and share the data with the researchers. Descriptive approaches to user profiling could thus complement traditional socio-economic descriptors with *post-demographic* properties [13] in the form of like data and the *relational* data represented by friendship networks. It is worth mentioning that Netvizz uses the unique Facebook account identifiers as "keys" for nodes in the GDF format; this means that all network files can be combined to form larger networks because the same user appearing in two different files will be a single node if the networks are combined, e.g. in gephi.

The group and page features also enable or facilitate data-driven approaches to studying Facebook users and uses without requiring access to individual accounts. In the case of groups, one needs to be a member to access its data; in the case of pages, liking it is enough to make it show up in the Netvizz interface. The analytical possibilities afforded by the second perspective are explored in more detail via two short case studies in the following section, but one could classify analytical dimensions along a series of very basic questions:

Who? This concerns studies of users (profile data), their relations (friendship patterns and interactions), and the larger social spaces emerging through groups and pages.

What? For personal networks, this relates mainly to *likes*, while pages allow for an investigation into *posts*, in particular concerning media types and audience engagement.

Where? For all outputs containing information about users, interface language is provided in a comprehensive way, because users do not have the possibility to prevent applications from receiving this information. While interface language is certainly not a perfect stand-in for locality, it allows engaging the question of geography in interesting ways.

When? Temporal data is limited to pages, but here, a timestamp for each post and comment is provided, allowing for investigating page and user activity over time.

EXAMPLES

To make the provided directions for analysis more tangible, this section briefly outlines two case studies investigating the use of Facebook in political activism online, more precisely its use by the anti-Islam movements that have grown at a rapid pace, in particular since the 9/11 attacks. The first example focuses on a group and the second on a page. Both examples mobilize concepts and techniques from Social Network Analysis (SNA), which developed out of the work of social psychologists Jacob Moreno and Kurt Lewin in the 1930s and 1940s. Although its tight relationship with social exchange theory [3] has granted a certain amount of visibility to SNA, it is only the wide availability of relational data and the software tools to analyze these data that the approach has gained the popularity it enjoys today. The main tenant of SNA is to envision groups and other social units as *networks*, that is, as connected ensembles that emerge from tangible and direct connections (friendships, work relationships, joint leisure, direct interactions, etc.) rather than as social *categories* that are constructed on the bases of shared (socio-economic) properties instead of actual interactions. This approach is particularly promising when applied to Facebook groups.

The "Islam is Dangerous" Group

The "Islam is Dangerous" group is an "open" group on Facebook, which means that its shared posts and members are visible to every other Facebook user. At the time of writing, the group had 2339 members and was mainly dedicated to sharing information about atrocities, crimes, infractions or simply deviations from cultural standards by Muslims.

A first approach used Netvizz for extracting all friendship connections between all the members of the group. While it is difficult to imagine an "average" Facebook group, a first finding is constituted by what seems to be a relatively high network density of 0.019. An average degree of 39.7 is a second indicator that this is group hosts a tightly knit collective rather than a loosely associated group merely sharing information on a subject. Friendship patterns are, however, not evenly distributed. While 18.3% of the group members have no friendship connection with other members – a population attracted by the subject matter rather than through social contacts? – 37.2% have at least 20 connections and 14.8% 100 or more.

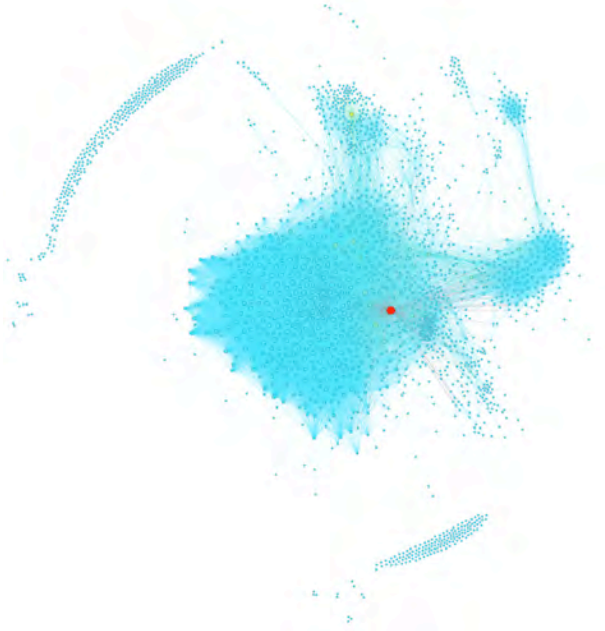


Figure 3. Friendship graph for the "Islam is Dangerous" group, colors represent betweenness centrality via a heat scale (blue => yellow => red).

While counting connections may be one way to identify leaders in a group, network analysis provides an extensive arsenal of techniques to analyze graphs in more specific ways. Figure 3 shows a spatialized visualization of the group (using gephi) and points to our ability to use advanced graph metrics to further analyzed the dataset by coloring nodes with a metric called *betweenness centrality*. This measure expresses a node's positioning in the larger topology of a graph and it can be very useful for detecting strategic positioning rather than popularity or social status. A person having high betweenness centrality is considered to be able to "influence the group by withholding or distorting information in transmission" [5] because he or she is located as a passage point between different sections of a network. While there are caveats to consider, betweenness centrality can be likened to Robert Putnam's concept of "bridging" social capital [10], which denotes the capacity to connect separate groups. In our case, this metric identifies the group administrator as the central *bridger*, which points to a group structure that, despite its high connectivity, is held together by a central figure.

The application of betweenness centrality can be seen as an example – a large number of techniques are now available to investigate structure, demarcate subgroups or qualify users in terms of their position in the network. Graph analysis software generally provides implementations of these metrics to researchers.

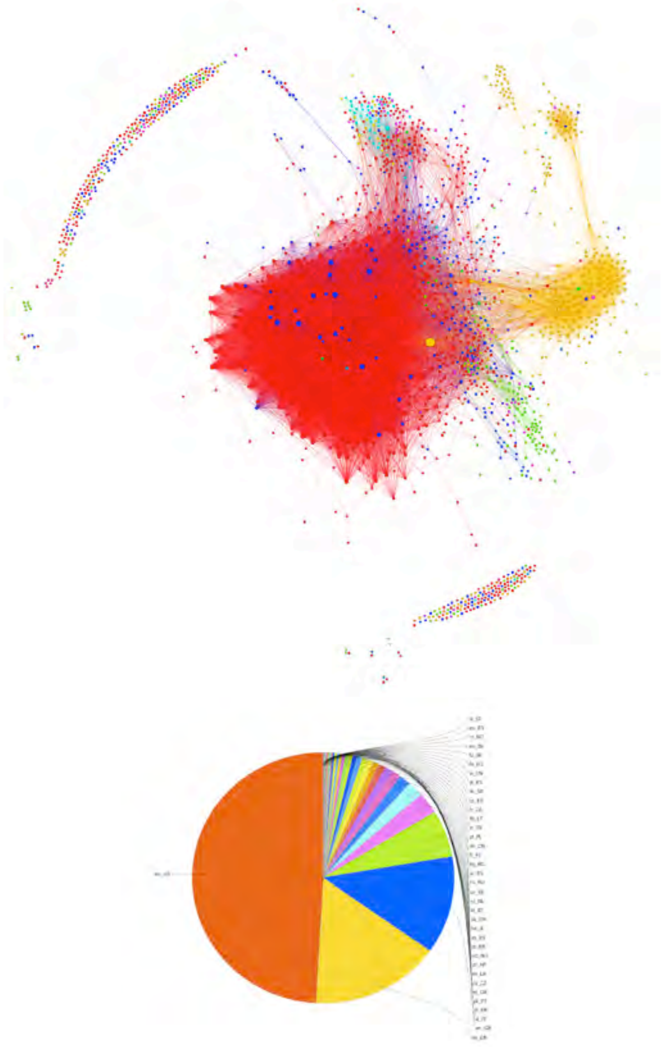


Figure 4. Friendship graph for the "Islam is Dangerous" group, colors represent "locale", i.e. the language of the Facebook interface for a given user.

Another example for types of analysis makes use of the users' interface language ("locale"), one of the few data points available for every Facebook member. Figure 4 shows the same network diagram as above, but uses locale to color nodes. We can see that there is a densely connected cluster of English speakers (both US and UK) that dominates the group, but smaller subcommunities, in particular a German one in yellow, can be identified as well. We can make the argument that this group, despite its high level of connectivity retains a degree of national coherence.

The "Educate children about the evils of islam" page

The second example quickly analyzes the Facebook page entitled "Educate children about the evils of Islam", which had been liked by 1586 users at the time of writing.

When extracting data from pages, Netvizz essentially operates by iterating over the last n (< 999) posts, collecting the posts themselves, as well as all of the users that like and comment on them. These data can be analyzed in various ways, either as bipartite network (Figure 5) or in more traditional form trough statistical analysis (Figures 6 and 7).

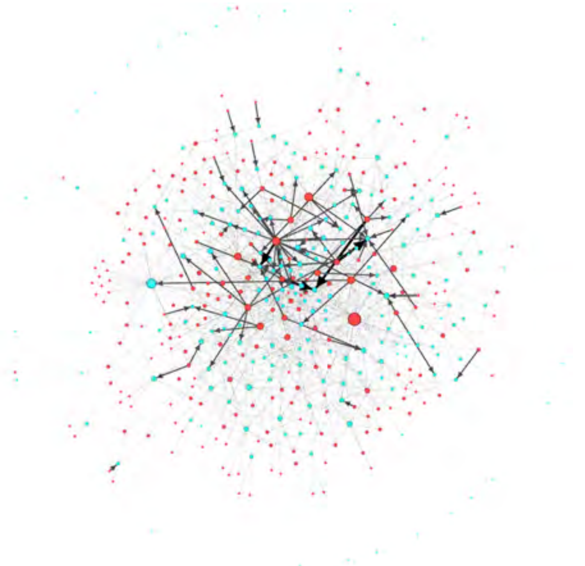


Figure 5. A network diagram showing the last 200 posts (turquoise), as well as the 253 users (red) liking and commenting them.

Network analysis maps interactions on a structural level and allows for the quick identification of particularly successful posts (in terms of engagement) and particularly active users. In this case, what emerges is a picture of a rather lively and intense conversational setting, with a core of loyal visitors that comment and react regularly.

Analyzing the posts over time (Figure 6), we can see that the 200 posts cover a period of less than four weeks, which indicates a high level of investment by the page owner, the only person allowed to post on the page.

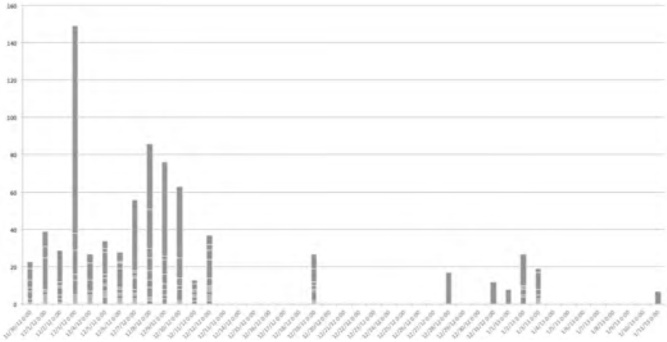


Figure 6. A stacked barchart showing the last 200 posts according to the days they were posted on; values indicate user engagement.

Because Facebook segments posts in content categories, we can also analyze content types, e.g. in relation to how particular types succeed in engaging users.

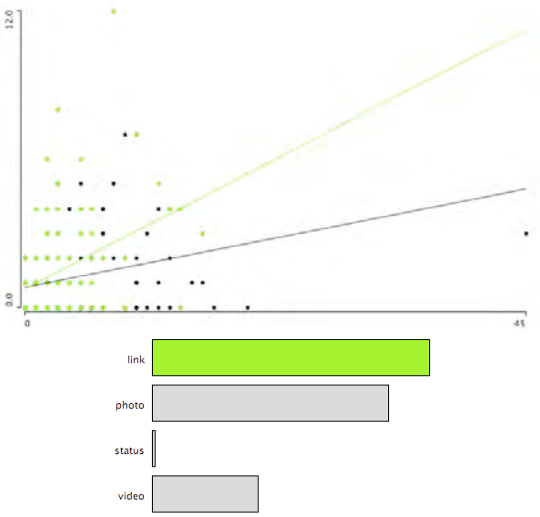


Figure 7. Visualization (using Mondrian) of the content types of the last 200 posts and how often they were liked (x-axis) and commented on (y-axis). Links are highlighted.

Figure 7 shows not only the distribution of content types over the last 200 posts (barchart), but also allows us to correlate these types to user activities. We can learn that links have a higher probability to receive comments, while photos are particularly likely to be liked.

These examples are mere illustrations of the analytical potential the in-depth data Facebook collects and Netvizz extracts. Many other types of analysis – from statistics to content analysis – are possible.

PRIVACY AND RESEARCH ETHICS CONSIDERATIONS

This final sections briefly sketches two aspects related to questions of privacy and research ethics, which would, however, merit a much more in-depth discussion that the space constraints allow.

The Facebook API as privacy challenge

Before discussing ethical considerations of data extraction on Facebook, it is useful to point out that part of the motivation for developing the Netvizz application was an exploration of the Facebook API itself, including the question how it governs access to data and what this means for users' capacity to limit or curate the way their data is accessible to others. This question is important because machine access needs to be treated differently than user interface access to data. While the latter is generally put to the front, the former allows for much more systematic forms of high speed and high volume data gleaning. Manual surveillance of activity is certainly possible, but I would argue that the largest part of user data collection by third parties on Facebook is performed via software that uses similar technological strategies as the Netvizz application. The application – and the knowledge gained by developing it – should therefore also be considered as an indicator of the types of information that other Facebook applications can get access to and certainly make extensive use of. While the fine-grained permission model holds the promise to limit third party access by asking users explicitly for permission, there is often no possibility for users to actually modulate which rights are granted: the application has to ask for detailed permissions for individual elements, but we can only acquiesce to all request or not use the platform. Access can be revoked *after* installation, but this means that applications can read that data at least once.

As Netvizz shows, a user granting rights to an application generally means that considerable access is given not only to her data, but also to *other* users' data. Application programming for research proposes is useful because of the analytical outcomes it produces or helps to produce, but it should also be considered as an investigation into the technological structures of platforms, which are as relevant to matters of privacy and beyond as they are understudied.

Research ethics

Social scientists have been confronted with the ethical dimension of empirical research well before the advent of the Internet. At no point have answers been easy or clear-cut. Recent debates amongst Internet researchers [20] have tended to put emphasis on the question of individual privacy. We should, however, note that there are significant cultural and political variations when it comes to arguing research ethics. Following Fuchs' critique [6] of the one-sided emphasis on a narrow definition of privacy, I would like to argue that research ethics navigate in a field defined by a number of tensions and competition between different ideals. Putting individuals' privacy on the top of the pyramid is a choice that can be traced to liberal sources of normative reasoning in particular, but we should not forget that these value sources are contingent and culturally colored. Competing ideals, such as the independence of research, larger social utility or the struggle against the encroaching of the private domain on publicness can equally be connected to established traditions in ethical reasoning.

It is clear that national traditions respond to these matters in different ways. While research ethics boards have become the norm in English-speaking countries, such an institutional governance of ethical decisions is hard to imagine in continental European countries such as France, where normative reasoning is concentrated both on the levels of the state and the individual, but only to a lesser degree on the layers in between. Similarly, the study of political extremism, and of the groups and individuals active in such movements, will not be framed in the same way in Germany and the United States, for obvious historical reasons.

What does that mean for Netvizz? Two decisions have been made: first, to anonymize all users for both groups and pages, simply because the number of accounts that can be collected this way is very large. For bigger pages, it is easy to quickly collect data for tens or even hundreds of thousands of user accounts. Second, Netvizz provides an option to anonymize accounts for personal networks. In this case, the complicated weighing of values and research ethics stays in the realm of the user/researcher and are only partially delegated to the programmer.

CONCLUSIONS

This paper has described the Netvizz application, a general-purpose data-extractor for different subsections of the Facebook platform. With a focus on questions relevant to media scholars, in particular, I have contextualized the application in a wider set of research concerns. With Facebook now counting over one billion active users, it is becoming urgent to develop and solidify research approaches to a service, largely constructed as a *walled garden*, that is part of an ongoing privatization of communication, both in terms of economics and accessibility. While there are important limits to what can be done without having to enter into a partnership with the company, the Netvizz application shows that certain parts of Facebook *are* amendable to empirical analysis, after all.

As Netvizz is continuously developed further, additional features will be added in the future. Providing more in-depth data on temporal aspects of user engagement with contents will certainly be one of the next steps.

ACKNOWLEDGMENTS

I would like to thank the attendants of the *Digital Methods Winter School 2013* for their useful comments, in particular Erik Borra and Jean-Christophe Plantin, as well as four anonymous reviewers.

REFERENCES

1. Agre, P.E. Surveillance and Capture: Two Models of Privacy. *The Information Society 10*, 2 (1994), 101-127.
2. Anscombe, F.J. Graphs in Statistical Analysis. *The American Statistician 27*, 1 (1973), 17-21.
3. Emerson, R.M. Social Exchange Theory. *Annual Review of Sociology 2*, (1976), 335-362.
4. Facebook Key Facts. http://newsroom.fb.com/Key-Facts.
5. Freeman, L.C. Centrality in Social Networks. Conceptual Clarification. *Social Networks 1*, 3 (1979), 215-239.
6. Fuchs, C. An Alternative View of Privacy on Facebook. *Information 2*, 4 (2011), 140-165.
7. Hacking, I. *Historical Ontology*. Harvard University Press, Cambridge, MA, USA, 2004.
8. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. and Christakis, N. Tastes, Ties, and Time: a New Social Network Dataset Using

Facebook.com. *Social Networks 30*, 4 (2008), 330-342.

9. Manovich, L. Trending: the Promises and the Challenges of Big Social Data. In Gold, M. *Debates in the Digital Humanities*. The University of Minnesota Press, Minneapolis, MN, USA, 2012, 460-475.

10. Putnam, R.D. *Bowling Alone: the Collapse and Revival of American Community*. Simon and Shuster, New York City, USA, 2000.

11. Quercia, D., Lambiotte, R., Kosinski, M., Stillwell, D., and Crowcroft, J. The Personality of Popular Facebook Users. In *Proc. CSCW 2012*, ACM Press (2012), 955-964.

12. Rogers, R. *The End of the Virtual*. Amsterdam University Press, Amsterdam, The Netherlands, 2009.

13. Rogers, R. Post-Democraphic Machines. In Dekker, A., Wolfsberger, A. *Walled Garden*. Virtual Platform, Amsterdam, The Netherlands, 2009, 29-39.

14. Schäfer, M.T. *Bastard Culture! How User Participation Transforms Cultural Production*. Amsterdam University Press, Amsterdam, The Netherlands, 2011.

15. Scott, J. Social Network Analysis. *Sociology 22*, 1 (1988), 109-127.

16. Simonite, T. What Facebook Knows. *MIT Technology Review*, June 13, 2012.

17. Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. The Anatomy of the Facebook Social Graph. *eprint arXiv:1111.4503*, 2011.

18. Watts, D.J. The 'New' Science of Networks. *Annual Review of Sociology 30*, 1 (2004), 243-270.

19. Wilson, R.E., Gosling, S.D. and Graham, L.T. A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science 7*, 3 (2012), 203-220.

20. Zimmer, M. 'But the Data Is Already Public': on the Ethics of Research in Facebook. *Ethics and Information Technology 12*, 4 (2010), 313-325.

Post-Demographic Machines

Richard Rogers
http://www.govcom.org

Richard Rogers holds the Chair in New Media & Digital Culture at the University of Amsterdam. He is also Director of the Govcom.org Foundation, Amsterdam, the group responsible for the Issue Crawler and other info-political tools, and Director of the Digital Methods Initiative, reworking method for Internet research. He is author of Information Politics on the Web, awarded the best 2005 book by the American Society for Information Science and Technology (ASIS&T). Current research interests include Internet censorship, googlization & Google art, the Palestinian-Israeli conflict on the Web as well as the technicity of content.

Post-demographics?

Leading research into social networking sites considers such issues as presenting oneself and managing one's status online, the different 'social classes' of users of MySpace and Facebook and the relationship between real-life friends and 'friended' friends (Boyd & Ellison, 2007). Another set of work, often from software-making arenas, concerns how to make use of the copious amounts of data contained in online profiles, especially interests and tastes. I would like to dub this latter work 'post-demographics'. Post-demographics could be thought of as the study of the data in social networking platforms, and, in particular, how profiling is, or may be, performed. Of particular interest here are the potential outcomes of building tools on top of profiling platforms, including two described below. What kinds of findings may be made from mashing up the data, or what may be termed meta-profiling? Elfriendo.com is an application that profiles a set of friends. It allows one to compare the tastes of a set of friends to those of another, using MySpace data. Which TV shows are most referenced by those who have friended Barack Obama? How do they differ from those shows as well as books, music and movies from John McCain's 'friends' online? (The small case study was performed prior to the U.S. presidential elections in November, 2008.) The second example of post-demographic work described here is the Leaky Garden Project (leakygarden.net), which furnishes a list of online services a particular user has subscribed to. One 'profiles' an individual (username) from the accounts taken out in Web 2.0 applications. Subsequently one sees the amount and also the details of the username's activity per platform, if, that is, the user's traces have been indexed by the major search engine, Google. These are 'leaks' in the so-called walled gardens, a term I return to.

Conceptually, with the 'post' prefixed to demographics, the idea is to stand in contrast to how the study of demographics organizes groups, markets and voters in a sociological sense. It also marks a theoretical shift from how demographics have been used 'bio-politically' (to govern bodies) to how post-demographics are employed 'info-politically,' to steer or recommend certain information to certain people (Foucault, 1998; Rogers, 2004). The term post-demographics also invites new methods for the study of social networks, where of interest are not the traditional demographics of race, ethnicity, age, income, and educational level – or derivations thereof such as class – but rather of tastes, interests, favorites, groups, accepted invitations, installed apps and other information that comprises an online profile and its accompanying baggage. As with Elfriendo and the Leaky Garden Project, the question concerns, which approaches and methods may be brought to bear in order to create new derivations from profile information, apart from niches and other, more specific products of behavioral marketing (Turow, 2006)?

Post-demographics is preferred over post-demography, as it recognizes popular usage of the notion of a 'demographic', referring to a segment or niche that may be targeted or polled. Crucially the notion attempts to capture the difference between how 'demographers' and, say, 'profilers' collect as well as use data. Demographers normally would analyze official records (births, deaths, marriages) and survey populations, with census taking being the most well known of those undertakings. Profilers, contrariwise, have users input data themselves in platforms that create and maintain social relations. They capture and make use of information from users of online platforms.

Perhaps another means of distinguishing between the two types of thought and practice is with reference to the idea of 'digital natives', those growing up with online environments, and unaware of life prior to the Internet, especially with the use of manual systems that came before it, like a library card catalogue (Prensky, 2001). The category of digital natives, however, takes a 'generational' view, and in that sense is a traditional demographic way of thinking. The post-demographic project would be less interested in new digital divides (digital natives versus non-natives) and the narratives that emerge around them (e.g., moral panics), but rather in how profilers recommend information, cultural products, events or other people ('friends') to users, owing to common tastes, locations, travel destinations and more. There is no end to what *could* be recommended, if the data are rich and stored.

Social networking sites as object of post-demographic study

'We define social networking websites here as sites where users can create a profile and connect that profile to other profiles for the purposes of making an explicit personal network' (Lenhart & Madden, 2007). Thus begins the study of American teenage use of such sites as MySpace and Facebook, conducted for the Pew Internet & American Life Project. 91% of the respondents use the sites to 'manage friendships'; less than a quarter use the sites to 'flirt'. Leaving behind surveys of user experiences for a moment, what is not as well known is what 'non-users' do with social network sites, with the occasional exception, such as the enquiry into how spammers leverage MySpace (Zinman & Donath, 2007). Non-users are those who do not manage friendships or flirt, but still visit the sites and read the profiles. They also may be interested in the data sets, and in automated means of capturing them, such as making use of the APIs (or application programming interface), or screen-scraping the pages. With 'post-demographics', the proposal is to make a contribution to the non-user studies – those profilers and researchers that both collect as well as harvest (or scrape) social networking sites' data for further analysis or software-making, such as mash-ups.[1]

How could one characterize the difference between the databases of online platforms and the databases of old (and new) that profile people to 'sort' them (Gandy, 1993)? Database philosophers were once deeply concerned about mandatory fields and field character limits – the number of letters and numbers that would fit on each line in the electronic or hard copy form. The paucity of fields and the limited space available for an entry would impoverish the self, similar to how bureaucracy transformed individuals into numbers (Poster, 1991). People could not describe themselves fittingly in a few fields and characters.

Other critiques of early database profiling practices pointed out that the 'anomaly' was the most significant output of analysis. Certain people (in the sense of data constructs) would stand out from the rest, owing to their lack of statistical normalcy. In a cultural theory sense, the database became the site to derive the other.

What may be derived from the new databases? More otherness? Now, with online platforms, there are longer character limits, more fields, and far greater agency to author oneself, or as one scholar aptly put it, 'to type oneself into being' (Sunden, 2003). 'Other', that last heading available on the form, standing for difference, or taxonomic indeterminacy, has been replaced, generally speaking, by 'more.' For example, the user is invited to 'write note', a freestyle field that provides opportunities for further self-definition and self-presentation. Now that the

[1] Non-users refer to profilers. Of course, profilers also may be users of the platforms, and most probably are, for one's sense of what may be mined, and how it may be analyzed or mashed up, would come from usage, with at least a minimal level of activity.

database is reaching out, providing you with more space to be yourself, questions may be posed. What does your form-filling say about you? Do you fill in the defaults only? Do you have many empty fields? What do your interests, and those of your friends, tell the profiler?

From a post-demographics perspective, the profile, together with the entities in orbit around it, lies at the core of research. Profilers are interested in what to do with all the 'interests' and 'favorites'.

You are media

What surrounds the profile? Generally, it has been observed that the Web, or at least a part of it, has new 'glue', or 'plasma' in the Latourian sense (Latour, 2005). Where once hyperlinks tied sites together, now the social networking sphere is viewed as less of a hypertext than a hyper-object space. From this perspective, the Web is more social than informational. The network has profiles as its nodes, with links between friends as well as social objects, not to mention 'social' third-party applications, socially derived recommendations as well as adverts (Knorr Cetina, 2001; Engeström, 2005). An initial question is how sociality is organized.

For one's profile, the user is invited to fill in certain personal information and list favorites. The fields for age, gender and location are still present; yet profiles invite the post-demographic, with requests for media listings, as favorite movies, music, TV shows, books, etc. It also asks for and stores media files, as pictures, clips and tunes. Once the profile has been completed (for the time being), the social linking begins. One 'friends' (the new verb), shares, joins groups and accepts invitations for events.

Sociality breeds more of it. The more social you are, the more prominent you become, in a presence sense. That is, your own activity boosts you on other (friends') pages, be it a tweet, wall writing, or comment, which may appear as running entries on other (friends') pages (Facebook). The platforms continually encourage more activity, inviting commentary on everything posted, and recommending to you more friends (who are friends of friends). With all the ties being made, and all the activity being logged, the opportunities for analysis, especially for social network researchers and profilers, appear to be boundless.

There are of course constraints. Certain of these concern the issues involved in harvesting the data, and making derivations. Which social networking sites are scrapable, and to which extent? When, and under which conditions, is it acceptable to harvest data? Apart from data collection, at issue is also data usage. The depersonalization of the data would be helpful in particular ethical discussions of social network site analysis, however much celebrated cases have shown 'why "anonymous" data sometimes isn't' (Schneier, 2007). There are norms for data usage, the most basic of which is user consent. When signing up, the user makes an agreement with the platform, and there are terms of use for both parties, as well as a service privacy policy. Of crucial importance however is the blurring of the line as to who is the primary agent of ensuring privacy. Arguably, on social networking sites, the user is assuming more and more responsibility for privacy, in the settings chosen. Whilst the services have thought through the default settings, the user is the one who lets his or her guard down, if you will, by changing the profile viewing setting from friends only, to friends of friends, which is the maximum exposure level inside Facebook.

How do social networking sites make available their data for profilers? Under the developers' menu item at Facebook, for example, one logs in and views the fields available in the API. Sample scripts are provided, as in 'get friends of user number x', where x is yourself. Thus the available scripts generally follow the privacy culture, in the sense that the user decides what the profiler can see. It becomes more interesting to the profiler when many users allow access, by clicking 'I agree' on a third-party application.

Another set of profiling practices are not interested in personal data per se, but rather in tastes and especially taste relationships. One may place many profiling activities in the category of depersonalized data analysis, including Amazon's seminal recommendation system, where it is not highly relevant which person also bought a particular book, but rather that people have done so. Supermarket loyalty cards and the databases storing purchase histories similarly employ depersonalized information analysis, where like Amazon, of interest is the quantity of particular items purchased as well as the purchasing relationships (which chips with which soft drink). Popular products are subsequently boosted. Certain combinations may be shelved together.

Post-demographic machines

Whilst they do not describe themselves as such, of course the most significant post-demographic machines are the social networking platforms themselves, collecting user tastes, and showing them to others, be they other friends, everyday 'people watchers' or profilers. Here however I would like to describe briefly two pieces of software built on top of machines, in the post-demographic analytical spirit, and the kinds of research practices that result.

Elfriendo.com is the outcome of thinking through how to make use of the profiles on the social networking platform, MySpace. At Elfriendo.com, enter a single interest, and the tool creates a new profile on the basis of the profiles of people expressing that single interest. One may also compare the compatibility of interests, i.e., whether one or more interests, tunes, movies, TV shows, books and heroes are compatible with other ones. Is Christianity compatible with Islam, in the sense that those people with one of the respective interests listen to the same music? Elfriendo answers those sorts of questions by analyzing sets of friends' profiles, and comparing interests across them. Thus a movie, TV show, etc. has an aggregate profile, made up of other interests. (To wit, Eminem, the rapper, appears in both the Christianity and Islam aggregate profiles, in early February 2009.)

One also may perform a semblance of post-demographic research with the tool, gaining an appreciation of relational taste analysis with a social networking site, more generally.[2]

It is instructive to state that MySpace is more permissive and less of a walled garden than Facebook, in that it allows the profiler to view a user's friends (and his/her friends' profiles), without you having friended anybody. Thus, one can view all of Barack Obama's friends, and their profiles. Here, in the example, one queries Elfriendo for Barack Obama as well as John McCain, and the profiles of their respective sets of friends are analyzed. The software counts the items listed by the friends under interests, music, movies, TV shows, books and heroes. What does this relational taste counting practice yield? The results provide distinctive pictures of the supporters of the two presidential candidates campaigning in 2008. The compatibility level between the interests of the friends of the two candidates is generally low. The two groups share few interests. (The tastes of the candidates' friends are not compatible for movies, music, books and heroes, though for TV shows the compatibility is 16%. See figure one.) There seem to be particular media profiles for each set of candidate's friends, where those of Obama for example watch the *Daily Show*, and those of McCain watch *Family Guy*, *Top Chef* and *America's Next Top Model*. Both sets of friends watch *Lost*.

The Leaky Garden Project

'Social networks require a degree of exclusion to work properly, (Shirky, 2003). Whilst commonly associated with certain social network sites, the term walled garden also refers to a business practice, notably in the software and hardware industries, where one firm's formats are incompatible with another's, thereby keeping the consumer 'locked in' (Arthur, 1989). Mobile phone rechargers come to mind, where Nokia's does not fit a Motorola phone, and vice versa. One of the arguments used in favor of lock-in is that dedicated hardware ensures the proper functioning of the technology. AT&T, with its historical slogan of 'one company, one system, universal service', made this argument repeatedly, in efforts to disallow 'foreign', or third party products and services, to run on the phone system, until the MCI lawsuit, and subsequent anti-trust work, finally unwound the Ma Bell monopoly in the 1970s and 1980s. With social networking sites, the notion of a walled garden cannot be applied as effortlessly. Social networking sites, especially Facebook, encourage third-party applications, in the new media style, with the realization that not only users' content, but also users' applications increase the value as well as levels of participation. This is the classic argument concerning the inversion of the 'value chain' in online games as well as in the entire Web 2.0 industry, summed up in the idea that the more who use it, and contribute to it, the better and more valuable it becomes (Shirky, 2008). (Like the now famous graphic by Bruce Clay that shows the dependencies between search engines, in a kind of data eco-system approach, see in figure two a rendition of the flows between leading 2.0 services, Facebook, Flickr and Twitter (Clay, n.d.).)

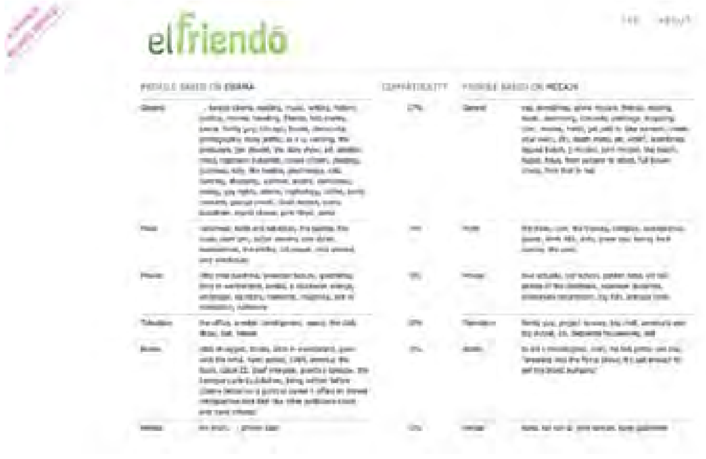
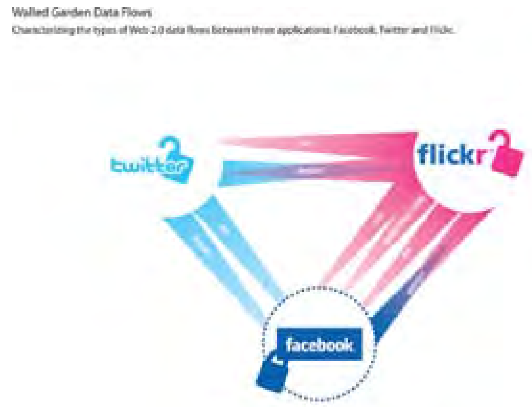


Figure one: The interests of Barack Obama's and John McCain's MySpace friends, 10 September 2008. Elfriendo.com, Govcom.org Foundation, Amsterdam, 2008.



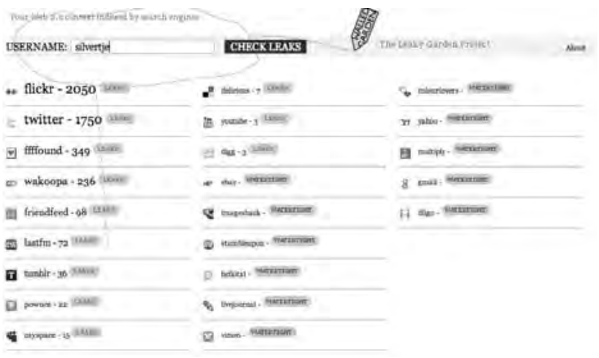
Walled Garden Data Flows

Characterizing the types of Web 2.0 data flows between three applications: Facebook, Twitter and Flickr.

Figure two: Walled Garden Data Flows. Digital Methods Initiative, Amsterdam, 2008.

[2] One gains only 'a sense' of how analysis may be performed, and the kinds of findings that may be made, because Elfriendo captures only the top 100 profiles, thus providing only an indication, as opposed to a grounded finding from a proper sampling procedure.

Here the question concerns, just how walled are these gardens? Apart from examining the data flows between applications, as above, the question of the permeability and penetrability of the platforms also may be approached by examining whether and to what extent each is indexed by search engines. In order to do so, leakygarden.net sits atop a machine that checks the availability of a particular username across a growing list of Web 2.0 applications. Usernamecheck.com is a useful service. When considering a new username, you may wish to know if and where it is taken, across the broader landscape of platforms. Here usernamecheck.com is repurposed, and in the first instance made into a profiling machine. Type in a username and check which services a person uses. Here the project researchers observed that generally speaking people seem to have two usernames, an alias as well as the real name (first and last name) as one word. Thus one may need to perform two queries for a fuller picture. Subsequently, leakygarden.net looks up references to the username. Does Google return pages from that username per platform? In all, the Leaky Garden Project shows which 'walled gardens' leak, and which are watertight (see Figure three).


USERNAME: silvertje

flickr - 2050
twitter - 1750
ffffound - 349
wakoopa - 236
friendfeed - 98
lastfm - 72
tumblr - 36
pownce - 22
myspace - 15
delicious - 7
youtube - 1
digg - 2
ebay -
imageshack -
stumbleupon -
livejournal -
vimeo -
colourlovers -
yahoo -
multiply -
gmail -
diigo -

Figure three: Username service subscription profile of 'silvertje' (Anne Helmond), including the 'leaks', or the amount of silvertje references per service, indexed by Google. Leakygarden.net, Govcom.org Foundation and the Digital Methods Initiative, Amsterdam, 2008.

Conclusion: What would Nielsen do?

Two methods dominate old media-style 'audience' research, the hand-written diary of a TV viewer or radio listener and the automated meter, registering how long a TV or radio channel is on, per household or household member. The diary technique is still in use, with the Nielsen company sending out a survey pack to its randomly selected families four times per year to record viewing habits during the so-called 'sweeps weeks'. Each person surveyed provides demographics, and a list of the shows they watch. Advertising is subsequently targeted to a TV show's demographic, with soap operas being the classic case of ads tied to a type of show. Because of survey effects, i.e., people changing their viewing habits owing to their need to keep a diary and fit a profile, an automated technique may be preferred (Stabile, 1995). In the United States, such recording devices were first employed for radio listeners, with the introduction in the 1940s of the Nielsen audimeter, which registered which frequency a radio was tuned to, and for how long (McLuhan, 1951). The results were useful for advertisers, and remain so. Of the initial study performed with the audimeter in 1942, *Time Magazine* wrote: 'When the star of one of radio's most popular nighttime shows said "Good night", listening dropped sharply. The sponsor's closing commercial was heard by only a fraction of the program's audience' (*Time Magazine*, 1943). Nielsen's automated television ratings began in the 1950s, and were taken to the next level with the black box known as the Storage Instantaneous Audimeter, which captured TV viewing of each set in the household, sending data back to headquarters daily through a phone line. 'People meters' have been employed since the 1980s, where each member of the household has his/her own button on the remote control. Behind the button, in the database, are the user's age and gender, and the meter on top of the television is tagged with a location.

TV shows are rated through a point system, with one point given per percentage of all households watching. Advertising rates are subsequently expressed in cost per point. A show has an expected rating (based on history) as well as an actual rating. Of interest to the advertisers is the 'post-buy' calculation of actual audience reach, that is, whether their advert actually had the expected audience types and numbers. Was the advert a good buy?

Should post-demographics emulate the Nielsen machines and metrics? Are there post-demographic equivalents to the machines and their metrics? Indeed, one may transfer the counting method from TV audience research to social net-working sites, using the available interest fields as well as basic demographic data (gender, age and location). Thus one may tally references to a particular interest across an entire social networking platform, as colleagues and I did for Hyves in the Netherlands in 2007 (see figure four). (No demographic data were used in the example.) Among the types of favorites at Hyves are brands, and Hyvers, as the users are called, fill in that field, albeit often without the care and diligence that would be demanded of a Nielsen family member.

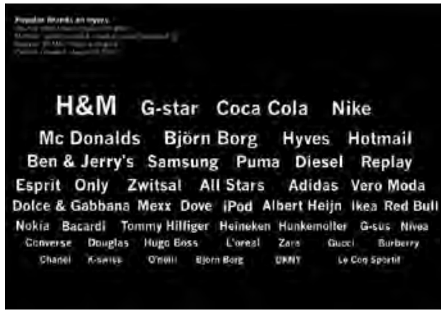


Figure four: Word cloud of the most referenced interests across the entire social networking platform Hyves, Govcom.org Foundation, Amsterdam, 2007.

Examples of 'non-cooperative' Hyvers' brands field (to 6 August 2007):

My Style is My Brand
ben geen merkentype
Houd er niet van ge(brand)merkt te worden
ik ben niet zo van de merken
I don't spend much time thinking about brands
Daar doe ik dus ff lekker niet aan mee he
Ik merk het
geen zin in aanvinken

How to tidy the data and make ratings? What would Nielsen do? One could strive to transfer the audience research technique to the new medium. Perhaps particular Hyvers would agree to become Nielsen social networkers, and provide meticulous up-to-date profiles. The fields would be monitored by Nielsen for changes in interests and tastes, and ratings could be provided with a point system, where fans are the equivalents of viewers.

As unlikely as the proposal may sound, it points up the larger question of whether and when to import standards methods of study onto the new medium. It also raises the question of the uses to be put to post-demographics.

Acknowledgments:

Richard Rogers would like to acknowledge the contributions made to earlier versions of the text by the Digital Methods Initiative, including Erik Borra, Sabine Niederer, Michael Stevenson, Marijn de Vries Hoogerwerff and Esther Weltevrede. The Leaky Garden Project software and graphics are the outcome of the Walled Garden event, organized by the Virtueel Platform, Amsterdam, November, 2008, with additional contributions by Marieke van Dijk, Andrei Fiore, Anne Helmond, Koen Martens, Auke Touwslager and Laura van der Vlies. The Elfriendo software is the product the Suggested Fields workshop by the Govcom.org Foundation at the Netherlands Media Art Institute, Amsterdam, January, 2008, for which the term post-demographics was coined.

References:

Arthur, W. Brian, 'Competing Technologies, Increasing Returns and Lock-in by Historical Events', *Economic Journal.* 99 (1989): 106-131.

Boyd, Danah and Ellison, Nicole, 'Social network sites: Definition, history, and scholarship'. *Journal of Computer-Mediated Communication.* 13(1) (2007), http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.

Clay, Bruce (n.d.). 'Search Engine Relationships Chart'. Bruceclay.com, http://www.bruceclay.com/searchenginerelationshipchart.htm.

Engeström, Jyri (2005). 'Why some social network services work and others don't. Or: the case for object-centered sociality'. Blog posting. Zengestrom.com, http://www.zengestrom.com/blog/2005/04/why_some_social.html.

Foucault, Michel, *The History of Sexuality Vol.1: The Will to Knowledge.* London: Penguin, 1998.

Gandy, Oscar, *The Panoptic Sort: A Political Economy of Personal Information.* Boulder, CO: Westview Press, 1999.

Knorr Cetina, Karin. 'Objectual Practice'. In: Schatzki, Theodor R., Knorr Cetina, Karin and von Savigny, Eike (eds.), *The Practice Turn in Contemporary Theory.* London: Routledge, 175-188, 2001.

Latour, Bruno, *Reassembling the Social: An Introduction to Actor-Network Theory.* Oxford: OUP, 2005.

Lenhart, Amanda and Madden, Mary, 'Social Networking Websites and Teens'. Pew Internet Project Data Memo (2007). Washington, DC: Pew Internet & American Life Project, http://www.pewinternet.org/pdfs/PIP_SNS_Data_Memo_Jan_2007.pdf.

McLuhan, Marshall, *The Mechanical Bride.* New York: Vanguard, 1951.

Poster, Mark, *The Mode of Information.* Chicago: University of Chicago Press, 1991.

Prensky, Marc, 'Digital Natives, Digital Immigrants'. *On the Horizon.* 9(5)(2001), http://pre2005.flexiblelearning.net.au/projects/resources/Digital_Natives_Digital_Immigrants.pdf.

Rogers, Richard, *Information Politics on the Web.* Cambridge, MA: MIT Press, 2004.

Schneier, Bruce, 'Why "Anonymous" Data Sometimes Isn't'. *Wired.* December, http://www.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1213, 2007.

Shirky, Clay (2003). People on page: YASNS… Blog posting. *Corante's Many-to-Many.* http://many.corante.com/archives/2003/05/12/people_on_page_yasns.php.

Shirky, Clay, *Here Comes Everybody.* New York: Penguin, 2008.

Stabile, Carol, 'Resistance, Recuperation and Reflexivity: The Limits of a Paradigm'. *Critical Studies in Mass Communication.* 12 (1995): 403-422.

Sunden, Jenny, *Material Virtualities: Approaching Online Textual Embodiment.* New York: Peter Lang, 2003.

Time Magazine (1943). 'Who Listens to What?' *XLI(1)*, 4 January.

Turow, Joseph, *Niche Envy.* Cambridge, MA: MIT Press, 2006.

Credits

ILLUSTRATION CREDITS

Page 9, 27, 28, 61, 71, 72, 73, 74, 84, 94, 95
Photos by Ward ten Voorde

Page 10
Photos by Anne Helmond, 2008, © All rights reserved

Page 11, 20, 21
Logo Walled Garden and badges Working Groups, designed by Studio Léon&Loes, Léon Kranenburg, 2008, © All rights reserved

Page 35, 36, 37
Images Elfriendo.com, Walled Garden data flows, Leakygarden.net and Word cloud Hyves, Govcom.org Foundation and the Digital Methods Initiative, 2007/2008

Page 46, 47
Walled Garden, Social and Semantic Serendipity working group, 2008 (made with http://prezicom)

Page 55, 62, 83, 96
Photos by Annette Wolfsberger

Page 68
Photo (top) by ib_odjov, http://www.flickr.com/photos/38392483@N00/385912858/sizes/o/in/photostream/
Photo (middle) by Gruntzooki, CC http://www.flickr.com/photos/doctorow/2732334638/sizes/l/
Photo (bottom) by Steinar Johnsen, CC http://www.flickr.com/photos/ess-jay/2530884062/sizes/l/

Page 93
Photo by Marijn de Vries Hoogerwerff, 2008, © All rights reserved

Page 110, 111
Image by Tom Klinkowstein and Irene Pereya, 2008, © All rights reserved

Page 112, 113
Photos by ANIMAE, 2008, © All rights reserved

Page 116, 117
Images by Artemesia/Celia Pearce, © All rights reserved

ACKNOWLEDGEMENTS

We would like to thank all the authors for their contribution to this book and the Walled Garden reporters and participants for their input:

PARTICIPANTS:
Emmy Alim
Kristina Andersen
Željko Blaće
Erik Borra
Maarten Brinkerink
Denis Burke
Marieke van Dijk
Janine Dijkmeijer
Sher Doruff
Maximillian von Duerckheim
Lucas Evers
Sofia Felix
Andrea Fiore
Lorna Goulden
Thomas Gray
Timothée Guicherd
Yolande Harris
Erik Hekman
Menno Heling
Anne Helmond
Liesbeth Huybrechts
Kirsten Krans
Frank Kresin
Corinne Kruger
Walter Langelaar
Maaike Lauwaert
Rogério Lira
Theo Mereboer
Bouke Mekel
Nicola Mullenger
Tom Oosterhuis
René Paré
Nuska Peszke Dako
Angela Plohman
Andrea Pozzi
Renée Ridgeway
Patrice Riemens
Tom Schofield
Joseph Scully
Karina Smrkovsky
Iskander Smit
Floor van Spaendonck
Auke Touwslager
Boudijn H. Uythof
Danja Vasiliev
Laura van der Vlies
Marijn de Vries Hoogerwerff
Esther Weltevrede
Dirk de Wit
Niels Wolf
Lotte Zwijnenburg

REPORTERS
Lisette van Blokland
Cathy Brickwood
Twan Eikelenboom
Maria Karagianni
Niels Kerssens
Annewil Neervens
Roman Tol

MODERATORS
Bronac Ferran
Tom Klinkowstein
Sabine Niederer and
Richard Rogers
Erin Manning
Aymeric Mansoux
Matt Ratto
Edward Shanken
Adam Somlai-Fischer and
Tapio Mäkelä

This book was made possible with the support of the Netherlands Ministry of Education, Culture and Science.

EDITED BY
Annet Dekker
Annette Wolfsberger

COPY EDITING
Cathy Brickwood

IMAGE EDITING
Niels Kerssens

DESIGN
Novak, Amsterdam

PRINTING
Lecturis, Eindhoven

VIRTUEEL PLATFORM
2009

“But the data is already public”: on the ethics of research in Facebook

Michael Zimmer

Published online: 4 June 2010

Abstract In 2008, a group of researchers publicly released profile data collected from the Facebook accounts of an entire cohort of college students from a US university. While good-faith attempts were made to hide the identity of the institution and protect the privacy of the data subjects, the source of the data was quickly identified, placing the privacy of the students at risk. Using this incident as a case study, this paper articulates a set of ethical concerns that must be addressed before embarking on future research in social networking sites, including the nature of consent, properly identifying and respecting expectations of privacy on social network sites, strategies for data anonymization prior to public release, and the relative expertise of institutional review boards when confronted with research projects based on data gleaned from social media.

Keywords Research ethics · Social networks · Facebook · Privacy · Anonymity

Introduction

In September 2008, a group of researchers publicly released data collected from the Facebook accounts of an entire cohort of college students. Titled “Tastes, Ties, and Time” (T3), the announcement accompanying the release noted the uniqueness of the data:

> The dataset comprises machine-readable files of virtually all the information posted on approximately 1,700 [Facebook] profiles by an entire cohort of students at an anonymous, northeastern American university. Profiles were sampled at 1-year intervals, beginning in 2006. This first wave covers first-year profiles, and three additional waves of data will be added over time, one for each year of the cohort’s college career.
>
> Though friendships outside the cohort are not part of the data, this snapshot of an entire class over its 4 years in college, including supplementary information about where students lived on campus, makes it possible to pose diverse questions about the relationships between social networks, online and offline. (N.A. 2008)

Recognizing the privacy concerns inherent with the collection and release of social networking data, the T3 research team took various steps in an attempt to protect the identity of the subjects, including the removal of student names and identification numbers from the dataset, a delay in the release of the cultural interests of the subjects, and requiring other researchers to agree to a “terms and conditions for use,” prohibiting various uses of the data that might compromise student privacy, and undergoing review by their institutional review board (Lewis 2008, pp. 28–29).

Despite these steps, and claims by the T3 researchers that “all identifying information was deleted or encoded” (Lewis 2008, p. 30), the identity of the source of the dataset was quickly discovered. Using only the publicly available codebook for the dataset and other public comments made about the research project, the identity of the “anonymous, northeastern American university” from which the data

M. Zimmer (✉)
School of Information Studies, University of Wisconsin-Milwaukee, 656 Bolton Hall, 3210 N. Maryland Ave, Milwaukee, WI 53211, USA
e-mail: zimmerm@uwm.edu

was drawn was quickly narrowed down to 13 possible universities (Zimmer 2008b), and then surmised to be Harvard College (Zimmer 2008a). Reminiscent of the ease at which AOL users were re-identified when the search engine thought the release of individuals' search history data was sufficiently anonymized (see Barbaro and Zeller Jr 2006), this re-identification of the source institution of the T3 dataset reveals the fragility of the presumed privacy of the subjects under study.[1]

Using the T3 data release and its aftermath as a case study, this paper will reveal numerous conceptual gaps in the researchers' understanding of the privacy risks related to their project, and will articulate a set of ethical concerns that must be addressed before embarking on future research similarly utilizing social network data. These include challenges to the traditional nature of consent, properly identifying and respecting expectations of privacy on social network sites, developing sufficient strategies for data anonymization prior to the public release of personal data, and the relative expertise of institutional review boards when confronted with research projects based on data gleaned from social media.

The "Tastes, Ties, and Time" project

Research in social networks has spanned decades, from Georg Simmel's foundational work in sociology (Simmel and Wolff 1964), to Barry Wellman's analyses of social networks in the emerging networked society of the late twentieth century (Wellman and Berkowitz 1988), to the deep ethnographies of contemporary online social networks by boyd (2008b). Indeed, the explosive popularity of online social networking sites such as MySpace, Twitter, and Facebook has attracted attention from a variety of researchers and disciplines (see boyd and Ellison 2008).[2] A primary challenge to fully understanding the nature and dynamic of social networks is obtaining sufficient data. Most existing studies rely on external surveys of social networking participants, ethnographies of smaller subsets of subjects, or the analysis of limited profile information extracted from what subjects chose to make visible. As a result, the available data can often be tainted due to self-reporting biases and errors, have minimal representativeness of the entire population, or fail to reflect the true depth and complexity of the information users submit (and create) on social networking sites.

Recognizing the data limitations faced by typical sociological studies of online social network dynamics, a group of researchers from Harvard University and the University of California—Los Angeles set out to construct a more robust dataset that would fully leverage the rich data available on social networking websites.[3] Given its popularity, the researchers chose the social network site Facebook as their data source, and located a university that allowed them to download the Facebook profiles of every member of the freshman class:

> With permission from Facebook and the university in question, we first accessed Facebook on March 10 and 11, 2006 and downloaded the profile and network data provided by one cohort of college students. This population, the freshman class of 2009 at a diverse private college in the Northeast U.S., has an exceptionally high participation rate on Facebook: of the 1640 freshmen students enrolled at the college, 97.4% maintained Facebook profiles at the time of download and 59.2% of these students had last updated their profile within 5 days. (Lewis et al. 2008, p. 331)

This first wave of data collection took place in 2006, during the spring of the cohort's freshman year, and data collection was repeated annually until 2009, when the vast majority of the study population will have graduated, providing 4 years of data about this collegiate social network. Each student's official housing records were also obtained from the university, allowing the researchers to "connect Internet space to real space" (Kaufman 2008a).

The uniqueness of this dataset is of obvious value for sociologists and Internet researchers. The data was extracted directly from Facebook without direct interaction with the subjects or reliance on self-reporting instruments, either of which could taint the data collected. The dataset includes demographic, relational, and cultural information on each subject, allowing broad analyses beyond more simple profile scraping methods. The inclusion of housing data for each of the 4 years of the study for analysis of any connection between "physical proximity, emerging roommate and friendship groups in the real world and the presence of these two types of relationships in their Facebook space" (Kaufman 2008a). Most importantly, the dataset represents nearly a complete cohort of college students, allowing the unique analysis of "complete social universe" (Kaufman 2008a), and it is longitudinal,

[1] While no individuals within the T3 dataset were positively identified (indeed, the author did not attempt to re-identify individuals), discovering the source institution makes individual re-identification much easier, perhaps even trivial, as discussed below.

[2] See also bibliography maintained by danah boyd at http://www.danah.org/SNSResearch.html.

[3] The research team includes Harvard University professors Jason Kaufman and Nicholas Christakis, UCLA professor Andreas Wimmer, and Harvard sociology graduate students Kevin Lewis and Marco Gonzalez.

providing the ability to study how the social network changes over time.

As a result of its uniqueness, the dataset can be employed for a number of research projects that have heretofore been difficult or impossible to pursue. As one of the "Tastes, Ties, and Time" researchers noted, "We're on the cusp of a new way of doing social science… Our predecessors could only dream of the kind of data we now have" (Nicholas Christakis, qtd in Rosenbloom 2007).

The dataset release

The "Tastes, Ties, and Time" project has been funded, in part, by a grant from the National Science Foundation,[4] who mandates certain levels of data sharing as a condition of its grants.[5] As a result, the Facebook dataset is being made available for public use in phases, roughly matching the annual frequency of data collection: wave 1 in September 2008, wave 2 in the fall of 2009, wave 3 in the fall of 2010, and wave 4 in the fall of 2011 (Lewis 2008, p. 3).

The first wave of data, comprising of "machine-readable files of virtually all the information posted on approximately 1700 FB profiles by an entire cohort of students at an anonymous, northeastern American university," was publicly released on September 25, 2008 (N.A. 2008).[6] Prospective users of the dataset are required to submit a brief statement detailing how the data will be used, and access is granted at the discretion of the T3 research team. Researchers are also required to agree to a "Terms and Conditions of Use" statement in order to gain access to the dataset, consenting to various licensing, use, and attribution provisions.

A comprehensive codebook was downloadable without the need to submit an application, which included detailed descriptions and frequencies of the various data elements (see Lewis 2008), including gender, race, ethnicity, home state, political views, and college major. For example, the codebook revealed that the dataset included 819 male and 821 female subjects, and that there were 1 self-identified Albanian, 2 Armenians, 3 Bulgarians, 9 Canadians, and so on.

The codebook also included an account of the steps taken by the T3 researchers in an attempt to protect subject privacy:

> All data were collected with the permission of the college being studied, the college's Committee on the Use of Human Subjects, as well as Facebook.com. Pursuant to the authors' agreement with the Committee on the Use of Human Subjects, a number of precautionary steps were taken to ensure that the identity and privacy of students in this study remain protected. Only those data that were accessible by default by each RA were collected, and no students were contacted for additional information. All identifying information was deleted or encoded immediately after the data were downloaded. The roster of student names and identification numbers is maintained on a secure local server accessible only by the authors of this study. This roster will be destroyed immediately after the last wave of data is processed. The complete set of cultural taste labels provides a kind of "cultural fingerprint" for many students, and so these labels will be released only after a substantial delay in order to ensure that students' identities remain anonymous. Finally, in order to access any part of the dataset, prospective users must read and electronically sign [a] user agreement… (Lewis 2008, p. 29)

These steps taken by the T3 researchers to remove identifying information reveal an acknowledgment of—and sensitivity to—the privacy concerns that will necessarily arise given the public release of such a rich and complete set of Facebook data. Their intent, as expressed by the project's principle investigator, Jason Kaufman, was to ensure that "all the data is cleaned so you can not connect anyone to an identity" (Kaufman 2008a). Unfortunately, the T3 researchers were overly optimistic.

Partial re-identification and withdrawal of dataset

Cognizant of the privacy concerns related to collecting and releasing detailed Facebook profile data from a cohort of college students, the T3 research team—in good faith—took a number of steps in an attempt to protect subject privacy, including review by their institutional review board, the removal of student names and identification numbers from the dataset, a delay in the release of the cultural interests of the subjects, and requiring other researchers to agree to a "terms and conditions for use" that prohibited any attempts to re-identify subjects, to disclose any identities that might be inadvertently re-identified, or otherwise to compromise the privacy of the subjects.

[4] See "Social Networks and Online Spaces: A Cohort Study of American College Students", Award #0819400, http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0819400.

[5] See relevant National Science Foundation Grant General Conditions (GC-1), section 38. Sharing of Findings, Data, and Other Research Products (http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gc109).

[6] The dataset is archived at the IQSS Dataverse Network at Harvard University (http://dvn.iq.harvard.edu/dvn/).

However, despite these efforts, the team's desire to ensure "all the data is cleaned so you can not connect anyone to an identity" fell short. On September 29, 2008, only 4 days after the initial data release, Fred Stutzman, a Ph.D. student at the University of North Carolina at Chapel Hill's School of Information and Library Science, questioned the T3 researchers' faith in the non-identifiability of the dataset:

> The "non-identifiability" of such a dataset is up for debate. A friend network can be thought of as a fingerprint; it is likely that no two networks will be exactly similar, meaning individuals may be able to be identified in the dataset post-hoc… Further, the authors of the dataset plan to release student "Favorite" data in 2011, which will provide further information that may lead to identification. (Stutzman 2008)

Commenting on Stutzman's blog post on the subject, Eszter Hargittai, an Associate Professor of Communication Studies at Northwestern University, sounded similar concerns:

> I think it's hard to imagine that some of this anonymity wouldn't be breached with some of the participants in the sample. For one thing, some nationalities are only represented by one person. Another issue is that the particular list of majors makes it quite easy to guess which specific school was used to draw the sample. Put those two pieces of information together and I can imagine all sorts of identities becoming rather obvious to at least some people. (Hargittai 2008)

Stutzman and Hargittai share a fear of the possible re-identification of the presumed anonymous Facebook dataset that has been made available to the public. Stutzman's concern over the ability to exploit the uniqueness of one's social graph to identify an individual within a large dataset has proven true in numerous cases (see, for example, Narayanan and Shmatikov 2008, 2009). Hargittai suggests that the uniqueness of the some of the data elements makes identifying the source of the data—and therefore some of the individual subjects—quite trivial. Hargittai's fears were correct.

Partial re-identification

Within days of its public release, the source of the T3 dataset was identified as Harvard College (see Zimmer 2008a, b). Most striking about this revelation was that the identification of the source of the Facebook data did not require access to the full dataset itself.

Using only the freely available codebook and referencing various public comments about the research, the source of the data was quickly narrowed down from over 2000 possible colleges and universities to a list of only seven (Zimmer 2008b). An examination of the codebook revealed the source was a private, co-educational institution, whose class of 2009 initially had 1640 students in it. Elsewhere, the source was identified as a "New England" school. A search through an online college database[7] revealed only seven private, co-ed colleges in New England states (CT, ME, MA, NH, RI, VT) with total undergraduate populations between 5000 and 7500 students (a likely range if there were 1640 in the 2006 freshman class): Tufts University, Suffolk University, Yale University, University of Hartford, Quinnipiac University, Brown University, and Harvard College.

Upon the public announcement of this initial discovery, and general criticism of the research team's attempts to protect the privacy of the subjects, Jason Kaufman, the principle investigator of the T3 research project, was quick to react, noting that, perhaps in justification for the amount of details released in the dataset, "We're sociologists, not technologists, so a lot of this is new to us" and "Sociologists generally want to know as much as possible about research subjects" (Kaufman 2008b). He then attempts to diffuse some of the implicit privacy concerns with the following comment:

> What might hackers want to do with this information, assuming they could crack the data and 'see' these people's Facebook info? Couldn't they do this just as easily via Facebook itself?
> Our dataset contains almost no information that isn't on Facebook. (Privacy filters obviously aren't much of an obstacle to those who want to get around them.) (Kaufman 2008b)

And then:

> We have not accessed any information not otherwise available on Facebook. We have not interviewed anyone, nor asked them for any information, nor made information about them public (unless, as you all point out, someone goes to the extreme effort of cracking our dataset, which we hope it will be hard to do). (Kaufman 2008c)

However, little "extreme effort" was needed to further "crack" the dataset; it was accomplished a day later, again without ever looking at the data itself (Zimmer 2008a). As Hargittai recognized, the unique majors listed in the codebook allowed for the ultimate identification of the source university. Only Harvard College offers the specific variety of the subjects' majors that are listed in the codebook, such as Near Eastern Languages and Civilizations,

[7] College Board, http://www.collegeboard.com.

Studies of Women, Gender and Sexuality, and Organismic and Evolutionary Biology. The identification of Harvard College was further confirmed after analysis of a June 2008 video presentation by Kaufman, where he noted that "midway through the freshman year, students have to pick between 1 and 7 best friends" that they will essentially live with for the rest of their undergraduate career (Kaufman 2008a). This describes the unique method for determining undergraduate housing at Harvard: all freshman who complete the fall term enter into a lottery, where they can designate a "blocking group" of between 2 and 8 students with whom they would like be housed in close proximity.[8]

In summary, the source of the T3 dataset was established with reasonable certainly in a relatively short period of time, without needing to download or access the dataset itself. While individual subjects were not identified in this process, the ease of identification of the source places their privacy in jeopardy given that the dataset contains a relatively small population with many unique individuals. The hopes by the T3 research team that "extreme effort" would be necessary to "crack" the dataset were, unfortunately, overly optimistic.

Withdrawal of the dataset

The announcement of this likely identification of the source of the Facebook dataset did not prompt a public reply by the T3 research team, but within 1 week of the discovery, the access page for the "Tastes, Ties, and Time" dataset displayed the following message, indicating that the dataset was, at least for the moment, no longer publicly available:

> Note: As of 10/8/08, prospective users may still submit requests and research statements, but the approval process will be delayed until further notice. We apologize for the inconvenience, and thank you for your patience.[9]

Then, in March 2009, the page was updated with a new message acknowledging the removal was in response to concerns over student privacy:

> UPDATE (3/19/09): Internal revisions are almost complete, and we expect to begin distributing again in the next 2–3 weeks. In the meantime, please DO NOT submit new dataset requests; but please check back frequently at this website for a final release notice. We again apologize for any inconvenience, and thank you for your patience and understanding as we work to ensure that our dataset maintains the highest standards for protecting student privacy.[10]

A full year after the initial release, the dataset remains unavailable, with the following message greeting interested researchers:

> UPDATE (10/2/09): The T3 dataset is still offline as we take further steps to ensure the privacy of students in the dataset. Please check back later at this site for additional updates- a notice will be posted when the distribution process has resumed.[11]

These messages noting the restricted access to the Facebook dataset to "ensure that our dataset maintains the highest standards for protecting student privacy" suggest that the re-identification of the source as Harvard College was correct, and that the T3 research team is re-evaluating their processes and procedures in reaction.

The insufficiency of privacy protections in the T3 project

The changing nature—and expectations—of privacy in online social networks are being increasingly debated and explored (see, for example, Gross and Acquisti 2005; Barnes 2006; Lenhart and Madden 2007; Nussbaum 2007; Solove 2007; Albrechtslund 2008; Grimmelmann 2009). The events surrounding the release of the Facebook data in the "Tastes, Ties, and Time" reveals many of the fault lines within these debates. Critically examining the methods of the T3 research project, and the public release of the dataset, reveals numerous conceptual gaps in the understanding the nature of privacy and anonymity in the context of social networking sites.

The primary steps taken by the T3 research team to protect subject privacy (quoted above), can be summarized as follows:

1. Only those data that were accessible by default by each RA were collected, and no students were contacted for additional information.
2. All identifying information was deleted or encoded immediately after the data were downloaded.
3. The complete set of cultural taste labels provides a kind of "cultural fingerprint" for many students, and so these labels will be released only after a substantial

[8] This process is described at the Harvard College Office of Residential Life website: http://www.orl.fas.harvard.edu/icb/icb.do?keyword=k11447&tabgroupid=icb.tabgroup17715.

[9] Screenshot of http://dvn.iq.harvard.edu/dvn/dv/t3 taken on October 22, 2008, on file with author.

[10] Screenshot of http://dvn.iq.harvard.edu/dvn/dv/t3 taken on March 27, 2009, on file with author. Webpage remains unchanged as of April 29, 2009.

[11] Screenshot of http://dvn.iq.harvard.edu/dvn/dv/t3 taken on November 1, 2009, on file with author. As of May 29, 2010, this message remains in place.

delay in order to ensure that students' identities remain anonymous.

4. In order to access any part of the dataset, prospective researchers must agree to a "terms and conditions for use" that prohibits any attempts to re-identify subjects, to disclose any identities that might be inadvertently re-identified, or otherwise to compromise the privacy of the subjects.
5. The entire research project, including the above steps, were reviewed and approved by Harvard's Committee on the Use of Human Subjects.

While each of these steps reveal good-faith efforts to protect the privacy of the subjects, each has serious limitations that expose a failures by the researchers to fully understand the nature of privacy in online social network spaces, and to design their research methodology accordingly. Each will be considered below, followed by a brief discussion of some of the public comments made by the T3 research team in defense of their methods and the public release of the dataset.

Use of in-network RAs to access subject data

In his defense of releasing subjects' Facebook profile data, Jason Kaufmann, the principle investigator of the T3 project, has stated that "our dataset contains almost no information that isn't on Facebook" and that "We have not accessed any information not otherwise available on Facebook" (Kaufman 2008c). Access to this information was granted by Facebook, but only through a manual process. Thus, research assistants (RA) from the source institution (presumably Harvard) were employed to perform the labor-intensive task of search for each first year student's Facebook page and saving the profile information. The dataset's codebook confirms that "Only those data that were accessible by default by each RA were collected, and no students were contacted for additional information" (Lewis 2008, p. 29).

The T3 codebook notes that of the 1,640 students in the cohort, 1,446 were found on Facebook with viewable profiles, 152 had a Facebook profile that was discoverable but not viewable by the RA, and 42 were undiscoverable (either not on Facebook or invisible to those not within their "friend" network) (Lewis 2008, p. 6).[12] Importantly, the codebook notes a peculiarity inherent with using in-network RAs to access the Facebook profile data:

> It is important to note that both undergraduate and graduate student RAs were employed for downloading data, and that each type of RA may have had a different level of default access based on individual students' privacy settings. In other words, a given student's information should not be considered objectively "public" or "private" (or even "not on Facebook")—it should be considered "public" or "private" (or "not on Facebook") from the perspective of the particular RA that downloaded the given student's data. (Lewis 2008, p. 6)

The T3 researchers concede that one RA might have different access to a student's profile than a different RA, and being "public" or "private" on Facebook is merely relative to that particular RAs level of access.

What appears to be lost on the researchers is that a subject might have set her privacy settings to be viewable to only to other users within her network, but to be inaccessible to those outside that sphere. For example, a Facebook user might decide to share her profile information only with other Harvard students, but wants to remain private to the rest of the world. The RAs employed for the project, being from the same network as the subject, would be able to view and download a subject's profile data that was otherwise restricted from outside view. Thus, her profile data—originally meant for only those within the Harvard network—is now included in a dataset released to the public. As a result, it is likely that profile information that a subject explicitly restricted to only "in network" participants in Facebook has been accessed from within that network, but then extracted and shared outside those explicit boundaries.

Given this likelihood, the justification that "we have not accessed any information not otherwise available on Facebook" is true only to a point. While the information was indeed available to the RA, it might have been accessible only due to the fact that the RA was within the same "network" as the subject, and that a privacy setting was explicitly set with the intent to keep that data within the boundaries of that network. Instead, it was included in a dataset released to the general public. This gap in the project's fundamental methodology reveals a troublesome lack of understanding of how users might be using the privacy settings within Facebook to control the flow of their personal information across different spheres, and puts the privacy of those subjects at risk.

Removal or encoding of "identifying" information

In an effort to protect the identity of the subjects, researchers note that "All identifying information was deleted or encoded immediately after the data were downloaded"

[12] Facebook allows users to control access to their profiles based on variables such as "Friends only", or those in their "Network" (such as the Harvard network), or to "Everyone". Thus, a profile might not be discoverable or viewable to someone outside the boundaries of the access setting.

(Lewis 2008, p. 29), and that "all the data is cleaned so you can not connect anyone to an identity" (Kaufman 2008a). Student names were replaced by "unique identification numbers" and any e-mail addresses or phone numbers that appeared in the Facebook profile data were excluded from the published dataset.

Yet, as the AOL search data release revealed, even if one feels that "all identifying information" has been removed from a dataset, it is often trivial to piece together random bits of information to deduce one's identity (Barbaro and Zeller Jr 2006). The fact that the dataset includes each subjects' gender, race, ethnicity, hometown state, and major makes it increasingly possible that individuals could be identified, especially those with a unique set of characteristics. Repeating Hargittai's concern: "I think it's hard to imagine that some of this anonymity would not be breached with some of the participants in the sample" (Hargittai 2008).

For example, the codebook reveals that each of these states has only a single student represented in the dataset: Delaware, Louisiana, Mississippi, Montana, and Wyoming. Similarly, there are only single instances of students identified as Albanian, Hungarian, Iranian, Malaysian, Nepali, Philippino, and Romanian. Their uniqueness very well might have resulted in publicity: it is possible that local media featured their enrollment at Harvard, or that the local alumni organization listed their name in a publicly-accessible newsletter, and so on. If such unique individuals can be personally identified using external sources, and then located within the dataset, one might also learn his/her stated political views or sexual preference, resulting in a significant privacy breach.

This reveals that even when researchers believe they have removed or encoded "all identifying information," there often remains information that could just as easily be used to re-identify individuals.[13] The T3 researchers' belief that stripping names alone is sufficient resembles the typical definition of "personally identifiable information" (PII) within the United States legal framework. As defined in California law, for example, PII is typically limited to an individual's name or other personally identifiable elements such as a social security number, a driver's license number, or a credit card number.[14] So long as these identifiers are removed from a dataset, it is presumed to be sufficiently anonymous.

However, others take a much broader stance in what constitutes personally identifiable information. The European Union, for example, defines PII much more broadly to include:

> [A]ny information relating to an identified or identifiable natural person…; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.[15]

Thus, while the T3 researchers might have felt simply removing or coding the subjects' names or other specific identifiers from the dataset was sufficient, had they followed the European Union's guidance, they would have recognized that many of the subjects' "physical, physiological, mental, economic, cultural or social identity" could also be used for re-identification. Even after removing the names of the subjects, since the dataset still includes race, ethnicity, and geographic data, re-identification of individual subjects remains a real possibility.

Delay in release of cultural taste data

Despite the apparent lack of use of the EU's more stringent definition of "personally identifiable information," the T3 researchers do recognize the unique nature of the cultural taste labels they have collected, referring to them as a kind of "cultural fingerprint". To protect subject privacy, the cultural tastes identified by the researchers have been assigned a unique number, and only the numbers will be associated with students for the initial data releases. The entire set of the actual taste labels will only be released in the fall of 2011, corresponding with the release of the wave 4 data.

The T3 researchers are right to recognize how a person's unique set of cultural tastes could easily identify her. Yet, merely instituting a "substantial delay" before releasing this personal data does little to mitigate the privacy fears. Rather, it only delays them, and only by 3 years. Researchers routinely rely on datasets for years after their initial collection: some influential studies of search engine behavior rely on nearly 10-year-old data (see, for example, Jansen and Resnick 2005; Jansen and Spink 2005), and these subjects' privacy needs do not suddenly disappear when they graduate from college in 2011.

Most surprisingly, despite the T3 researchers' recognition of the sensitive nature of the cultural data, they will

[13] Simply stripping names from records is rarely a sufficient means to keep a dataset anonymous. For example, Latanya Sweeny has shown that 87 percent of Americans could be identified by records listing solely their birth date, gender and ZIP code (Sweeney 2002).

[14] See, for example, the California Senate Bill 1386, http://info.sen.ca.gov/pub/01-02/bill/sen/sb_1351-1400/sb_1386_bill_20020926_chaptered.html.

[15] European Union Data Protection Directive 95/46/EC, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML.

provide immediate access to it on a case-by case basis. As the codebook reveals:

> In the meantime, if prospective users wish to access some subset of the taste labels, special arrangements may be made on a case-by-case basis at the discretion of the authors (send request and detailed justification to t3dataset@gmail.com). (Lewis 2008, p. 20)

No further guidance is provided as to what kinds of arrangements are made and what justifications are needed to make such an exception. If the T3 research team felt strongly enough that it was necessary to encode and delay the release of the subjects' "cultural fingerprints", it does not seem appropriate to announce that exceptions can be made for its release to selected researchers prior to the 3-year delay. If it is potentially privacy invading content, it simply should not be released.

Terms of use statement

Researchers wanting access to the T3 dataset must (electronically) sign a Terms and Conditions of Use statement. The statement includes various covenants related to protecting the privacy of the subjects in the dataset, including (as numbered in the original):

3. I will use the dataset solely for statistical analysis and reporting of aggregated information, and not for investigation of specific individuals or organizations, except when identification is authorized in writing by the Authors.
4. I will produce no links among the Authors datasets or among the Authors data and other datasets that could identify individuals or organizations.
5. I represent that neither I, nor anyone I know, has any prior knowledge of the possible identities of any study participants in any dataset that I am being licensed to use.
6. I will not knowingly divulge any information that could be used to identify individual participants in the study, nor will I attempt to identify or contact any study participant, and I agree to use any precautions necessary to prevent such identification.
7. I will make no use of the identity of any person or establishment discovered inadvertently. If I suspect that I might recognize or know a study participant, I will immediately inform the Authors, and I will not use or retain a copy of data regarding that study participant. If these measures to resolve an identity disclosure are not sufficient, the Authors may terminate my use of the dataset. (reproduced at Lewis 2008, p. 30)

The language within this statement clearly acknowledges the privacy implications of the T3 dataset, and might prove effective in raising awareness among potential researchers. However, studies have shown that users frequently simply "click through" such agreements without fully reading them or recognizing they are entering into a legally binding contract (Gatt 2002), and it is unclear how the T3 researchers specifically intend to monitor or enforce compliance with these terms. While requiring a terms of use is certainly a positive step, without enforcement it might have limited success in deterring any potential privacy-invasive use of the data.

IRB approval

As required of any research project involving human interaction, clearance for the research project and data release was provided by Harvard's intuitional review board (IRB), known as the Committee on the Use of Human Subjects in Research.[16] As Kaufman commented: "Our IRB helped quite a bit as well. It is their job to insure that subjects' rights are respected, and we think we have accomplished this" (Kaufman 2008c). Elsewhere he has noted that "The university in question allowed us to do this and Harvard was on board because we don't actually talk to students, we just accessed their Facebook information" (Kaufman 2008a).

Just as we can question whether the T3 researchers full understood the privacy implications of the research, we must critically examine whether Harvard's IRB—a panel of experts in research ethics—also sufficiently understood how the privacy of the subjects in the dataset could be compromised. For example, did the IRB recognize, as noted above, that using an in-network research assistant to pull data could circumvent privacy settings intended to keep that data visible to only other people at Harvard? Or did the IRB understand that individuals with unique characteristics could easily be extracted from the dataset, and perhaps identified? It is unclear whether these concerns were considered and discarded, or whether the IRB did not fully comprehend the complex privacy implications of this particular research project.[17] In either case, the potential privacy-invading consequences of the T3 data release suggest a possible lapse of oversight at some point of the IRB review process.

Other public comments

Beyond the shortcomings of the documented efforts to protect the privacy of the T3 dataset subjects, the researchers have made various public comments that reveal

[16] http://www.fas.harvard.edu/~research/hum_sub/.

[17] Attempts to obtain information about the IRB deliberations with regard to the T3 project have been unsuccessful.

additional conceptual gaps in their understanding of the privacy implications of the T3 research project.[18]

For example, when confronted with the potential re-identifiability of the dataset, Kaufman responded by pondering "What might hackers want to do with this information, assuming they could crack the data and 'see' these people's Facebook info?" and later acknowledging "Nonetheless, seeing your thought process—how you would attack this dataset—is extremely useful to us" (Kaufman 2008b). Kaufman's mention of "hackers", "attacking" the dataset, and focusing on what someone might "do" with this information exposes a *harm*-based theory of privacy protection. Such a position supposes that so long as the data can be protected from attack by hackers or others wishing to "do" something harmful once gaining access, the privacy of the subjects can be maintained. Such a position ignores the broader *dignity*-based theory of privacy (Bloustein 1964). Such a stance recognizes that one does not need to be a victim of hacking, or have a tangible harm take place, in order for there to be concerns over the privacy of one's personal information. Rather, merely having one's personal information stripped from the intended sphere of the social networking profile, and amassed into a database for external review becomes an affront to the subjects' human dignity and their ability to control the flow of their personal information.

The distinction between harm- and dignity-based theories of privacy are understood—and often debated—among privacy scholars, but when asked if they conferred with privacy experts over the course of the research and data release, Kaufman admits that "we did not consult [with] privacy experts on how to do this, but we did think long and hard about what and how this should be done" (Kaufman 2008c). Given the apparent focus on data security as a solution to privacy, it appears the T3 research team would have benefited from broader discussions on the nature of privacy in these environments.[19]

The T3 researchers also claim that there should be little concern over the ethics of this research since the Facebook data gathered was already publicly available. As Kaufman argues:

> On the issue of the ethics of this kind of research—Would you require that someone sitting in a public square, observing individuals and taking notes on their behavior, would have to ask those individuals' consent in advance? We have not accessed any information not otherwise available on Facebook. We have not interviewed anyone, nor asked them for any information, nor made information about them public… (Kaufman 2008c)

This justification presents a false comparison. The "public square" example depends on random encounters of people who happen to be in the square at the precise time as the researcher. Further, the researchers cannot observe everyone simultaneously, and instead must select which individuals to focus their attention, leaving some subjects out of the dataset. Finally, the data gathered is imprecise, and limited to the researchers ability to discern gender, age, ethnicity, and other physically-observable characteristics.

By contrast, the T3 researchers utilized an in-network research assistant to systematically access and download an entire cohort of college students' Facebook profile pages, each year for 4 years. They successfully targeted a specific and known group of students, obtaining a list of names and e-mail addresses of the students from the source university to improve their ability to gather data on the entire population. The data acquired included not only the subjects' self-reported gender and ethnicity, but also their home state, nation of origin, political views, sexual interests, college major, relational data, and cultural interests—data which would be considerably more difficult to obtain through observations in a public square. Suggesting that the two projects are similar and carry similar (and minimal) ethical dilemmas reveals a worrisome gap in the T3 research team's understanding of the privacy and ethical implications of their project.

The ethics of the "Tastes, Ties, and Time" project

The above discussion of the unsatisfactory attempts by the T3 researchers to protect subject privacy illuminates two central ethical concerns with the "Tastes, Ties, and Time" project: the failure to properly mitigate what amounts to violations of the subjects' privacy, and, thus, the failure to adhere to ethical research standards.

Privacy violations

The proceeding discussion notes numerous failures of the T3 researchers to properly understand the privacy implications of the research study. To help concretize these concerns, we can gather them into the following four

[18] This section is intended as an informal analysis of the discourse used when talking about the T3 project. It is meant to reveal gaps in broader understanding of the issues at hand, and not necessarily directed against a particular speaker.

[19] After the T3 research project was funded and well underway, Kaufman became a fellow at the Berkman Center for Internet & Society at Harvard University, an organization dedicated to studying a number of Internet-related issues, including privacy. While Kaufman presented preliminary results of his research to the Berkman community prior to joining the center (Kaufman 2008a), there is no evidence that others at Berkman were consulted prior to the release of the T3 dataset.

salient dimensions of privacy violations, as organized by Smith et al. (1996) and based on thorough review of privacy literature: the amount of personal information collected, improper access to personal information, unauthorized secondary use of personal information, and errors in personal information.[20] Viewing the circumstances of the T3 data release through the lens of this privacy violation framework helps to focus the ethical deficiencies of the overall project.

Amount of personal information collected

Privacy violations can occur when "extensive amounts of personally identifiable data are being collected and stored in databases" Smith et al. (1996, p. 172). Notably, the "Tastes, Ties, and Time" project's very existence is dependent on the extensive collection of personal data. The T3 project systematically, and regularly over a 4-year period, collected a vast amount of personal information on over 1,500 college students. Individual bits of data that might have been added and modified on a subject's Facebook profile page over time were harvested and aggregated into a single database, co-mingled with housing data from an outside source, and then compared across datafiles.

Improper access to personal information

Privacy violations might occur when information about individuals might be readily available to persons not properly or specifically authorized to have access the data. As described above, subjects within the T3 dataset might have used technological means to restrict access to their profile information to only members of the Harvard community, thus making their data inaccessible to the rest of the world. By using research assistants from within the Harvard community, the T3 researchers—whether intentional or not—would be able to circumvent those access controls, thereby including these subjects' information among those with more liberal restrictions.

Further, no specific consent was sought or received from the subjects in the study; their profile information was simply considered freely accessible for collection and research, regardless of what the subject might have intended or desired regarding its accessibility to be harvested for research purposes. Combined, these two factors reveal how a privacy violation based on improper access has occurred due to the T3 project.

Unauthorized secondary use

Unauthorized secondary use of personal information is the concern that information collected from individuals for one purpose might be used for another secondary purpose without authorization form the individual, thus the subject loses control over their information. Within Smith et al.'s. (1996) framework, this loss of control over one's personal information is considered a privacy violation. At least two incidences of unauthorized secondary use of personal information can be identified in the T3 project. First, the students' housing information and personal email addresses were provided to the T3 researchers to aid in their data collection and processing. These pieces of information were initially collected by the university to facilitate various administrative functions, and not for secondary use to assist researchers looking for students' profiles on Facebook. Second, the very nature of collecting Facebook profile information, aggregating it, and releasing it for others to download invites a multitude of secondary uses of the data not authorized by the students. The data was made available on Facebook for the purpose of social networking among friends and colleagues, not to be used as fodder for academic research. Without specific consent, the collection and release of Facebook data invariably brings about unauthorized secondary uses.

Errors in personal information

Finally, privacy concerns arise due to the impact of possible errors within datasets, which has lead to various policies ensuring individuals are granted the ability to view and edit data collected about them to minimize any potential privacy violations.[21] In the T3 project, subjects were not aware of the data collection nor provided any access to view the data to correct for errors or unwanted information.

Ethical research standards

Viewing the privacy concerns of the T3 data release through the lens of Smith et al.'s (1996) privacy violation framework helps to focus the ethical deficiencies of the overall project. In turn, our critique of the T3 project exposes various breeches in ethical research standards that, if followed, might have mitigated many of the privacy threats.

[20] I thank an anonymous reviewer for suggesting this organizing framework.

[21] See, for example, the United States Federal Trade Commission's Fair Information Practice Principles (http://www.ftc.gov/reports/privacy3/fairinfo.shtm), which include "Access" as a key provision, providing data subjects the ability to view and contesting inaccurate or incomplete data.

Ethical issues in human subjects research receive considerable attention, culminating in the scrutiny of research projects by Institutional Review Boards for the Protection of Human Subjects (IRB's) to review research according to federal regulations.[22] These regulations focus on research ethics issues such as subject safety, informed consent, and privacy and confidentiality. Others have then these broad standards and applied them specifically to Internet-based research and data collection. For example, the Association of Internet Researchers have issued a set of recommendations for engaging in ethical research online (see Ess and AoIR ethics working committee 2002), which places considerable focus on informed consent and respecting the ethical expectations within the venue under study.

As noted above, the T3 researchers did not obtain any informed consent by the subjects within the dataset (nor were they asked to do so by their Institutional Review Board). Further, as described in detail, the researchers failed to respect the expectations likely held by the subjects regarding the relative accessibility and purpose of their Facebook profile information. By failing to recognize that users might maintain strong expectations that information shared on Facebook is meant to stay on Facebook, or that only members of the Harvard network would ever have access to the data, the T3 researchers have failed in their duty to engage in ethically-based research.

Conclusion

The events surrounding the release of the Facebook data in the "Tastes, Ties, and Time" project –including its methodology, its IRB approval, the way in which the data was released, and the viewpoints publicly expressed by the researchers—reveals considerable conceptual gaps in the understanding of the privacy implications of research in social networking spaces. As a result, threats to the privacy of the subjects under study persist, despite the good faith efforts of the T3 research team.

The purpose of this critical analysis of the T3 project is not to place blame or single out these researchers for condemnation, but to use it as a case study to help expose the emerging challenges of engaging in research within online social network settings. These include challenges to the traditional nature of consent, properly identifying and respecting expectations of privacy on social network sites, developing sufficient strategies for data anonymization prior to the public release of personal data, and the relative expertise of institutional review boards when confronted with research projects based on data gleaned from social media.

As made apparent to the position of some of the T3 research team that their data collection methods were unproblematic since the "information was already on Facebook", future researchers must gain a better understanding of the contextual nature of privacy in these spheres (Nissenbaum 1998, 2004, 2009), recognizing that just because personal information is made available in some fashion on a social network, does not mean it is fair game for capture and release to all (see, generally, Stutzman 2006; Zimmer 2006; McGeveran 2007; boyd 2008a). Similarly, the notion of what constitutes "consent" within the context of divulging personal information in social networking spaces must be further explored, especially in light of this contextual understanding of norms of information flow within specific spheres. The case of the T3 data release also reveals that we still have not learned the lessons of the AOL data release and similar instances where presumed anonymous datasets have been re-identified. Perhaps most significantly, this case study has uncovered possible shortcomings in the oversight functions of institutional review boards, the very bodies bestowed with the responsibility of protecting the rights of data subjects.

Overcoming these challenges and conceptual muddles is no easy task, but three steps can be taken immediately to guide future research in social media spaces. One, scholars engaging in research similar to the T3 project must recognize their own gaps in understanding the changing nature of privacy and the challenges of anonymizing datasets, and should strive to bring together an interdisciplinary team of collaborators to help ensure the shortcomings of the T3 data release are not repeated. Two, we must evaluate and educate IRBs and related policy makers as to the complexities of engaging in research on social networks.[23] And three, we must ensure that our research methods courses, codes of best practices, and research protocols recognize the unique challenges of engaging in research on Internet and social media spaces.[24]

The "Tastes, Ties, and Time" research project might very well be ushering in "a new way of doing social

[22] See Part 46 Protection of Human Subjects of Title 45 Public Welfare of the Code of Federal Regulations at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

[23] See, for example, the "Internet Research Ethics: Discourse, Inquiry, and Policy" research project directed by Elizabeth Buchanan and Charles Ess (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0646591).

[24] An important movement in this direction is the recently funded "Internet Research and Ethics 2.0: The Internet Research Ethics Digital Library, Interactive Resource Center, and Online Ethics Advisory Board" project, also directly by Elizabeth Buchanan and Charles Ess (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0924604 and http://www.internetresearchethics.org/).

science", but it is our responsibility scholars to ensure our research methods and processes remain rooted in long-standing ethical practices. Concerns over consent, privacy and anonymity do not disappear simply because subjects participate in online social networks; rather, they become even more important.

Acknowledgments The author thanks the participants at the International Conference of Computer Ethics: Philosophical Enquiry in Corfu, Greece, as well as the Internet Research 10: Internet Critical conference in Milwaukee, Wisconsin, for their helpful comments and feedback. Additional thanks to Elizabeth Buchanan, Charles Ess, Alex Halavais, Anthony Hoffmann, Jon Pincus, Adam Shostack, and Fred Stutzman for their valuable insights and conversations, both online and off. The author also thanks the anonymous reviewers for their helpful suggestions and criticisms. This article would not have been possible without the research assistance of Wyatt Ditzler and Renea Drews. Finally, I would like to thank Jason Kaufman and Colin McKay at the Berkman Center for Internet & Society, for their valued and continued feedback regarding this work.

References

Albrechtslund, A. (2008). Online social networking as participatory surveillance. *First Monday* Retrieved 2008, March 3, from http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2142/1949.

Barbaro, M., & Zeller Jr, T. (2006). A face is exposed for AOL searcher no. 4417749. *The New York Times*, p. A1.

Barnes, S. (2006). A privacy paradox: Social networking in the United States. *First Monday* Retrieved October 12, 2007, from http://www.firstmonday.org/ISSUES/issue11_9/barnes/.

Bloustein, E. (1964). Privacy as an aspect of human dignity: An answer to Dean Prosser. *New York University Law Review, 39*, 962–1007.

boyd, D. (2008a). Putting privacy settings in the context of use (in Facebook and elsewhere). *Apophenia* Retrieved October 22, 2008, from http://www.zephoria.org/thoughts/archives/2008/10/22/putting_privacy.html.

boyd, D. (2008b). Taken out of context: American teen sociality in networked publics. Unpublished Dissertation, University of California-Berkeley.

boyd, D., & Ellison, N. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210–230.

Ess, C., & AoIR ethics working committee. (2002). Ethical decision-making and Internet research. Retrieved March 12, 2010, from http://www.aoir.org/reports/ethics.pdf.

Gatt, A. (2002). Click-wrap agreements the enforceability of click-wrap agreements. *Computer Law & Security Report, 18*(6), 404–410.

Grimmelmann, J. (2009). Facebook and the social dynamics of privacy. *Iowa Law Review, 95*, 4.

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. Paper presented at the 2005 ACM workshop on Privacy in the electronic society, Alexandria, VA.

Jansen, B. J., & Resnick, M. (2005). Examining searcher perceptions of and interactions with sponsored results. Paper presented at the Workshop on Sponsored Search Auctions at ACM Conference on Electronic Commerce, Vancouver, BC.

Jansen, B. J., & Spink, A. (2005). How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248–263.

Kaufman, J. (2008a). Considering the sociology of Facebook: Harvard Research on Collegiate Social Networking [Video].: Berkman Center for Internet & Society.

Kaufman, J. (2008b). I am the Principal Investigator… [Blog comment]. *On the "Anonymity" of the Facebook dataset* Retrieved September 30, 2008, from http://michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/.

Kaufman, J. (2008c). Michael—We did not consult… [Blog comment]. *michaelzimmer.org* Retrieved September 30, 2008, from http://michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/.

Lenhart, A., & Madden, M. (2007). Teens, privacy & online social networks. *Pew internet & American life project* Retrieved April 20, 2007, from http://www.pewinternet.org/pdfs/PIP_Teens_Privacy_SNS_Report_Final.pdf.

Lewis, K. (2008). Tastes, Ties, and Time: Cumulative codebook. Retrieved September 30, 2008, from http://dvn.iq.harvard.edu/dvn/dv/t3.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, Ties, and time: A new social network dataset using Facebook. com. *Social Networks, 30*(4), 330–342.

McGeveran, W. (2007). Facebook, context, and privacy. *Info/Law* Retrieved October 3, 2008, from http://blogs.law.harvard.edu/infolaw/2007/09/17/facebook-context/.

N.A. (2008). Tastes, Ties, and Time: Facebook data release. *Berkman Center for Internet & Society* Retrieved September 30, 2008, from http://cyber.law.harvard.edu/node/4682.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. Paper presented at the IEEE Symposium on Security and Privacy, 2008.

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing social networks. Paper presented at the 30th IEEE Symposium on Security and Privacy.

Nissenbaum, H. (1998). Protecting privacy in an information age: The problem of privacy in public. *Law and Philosophy, 17*(5), 559–596.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1), 119–157.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford, CA: Stanford University Press.

Nussbaum, E. (2007). Kids, the Internet, and the end of privacy. *New York Magazine* Retrieved February 13, 2007, from http://nymag.com/news/features/27341/.

Rosenbloom, S. (2007). On Facebook, scholars link up with data. *New York Times* Retrieved September 30, 2008, from http://www.nytimes.com/2007/12/17/style/17facebook.html?ref=us.

Simmel, G., & Wolff, K. H. (1964). *The sociology of Georg Simmel*. Glencoe, Ill: Free Press.

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly, 20*(2), 167–196.

Solove, D. (2007). *The future of reputation: Gossip, rumor, and privacy on the internet*. New Haven, CT: Yale University Press.

Stutzman, F. (2006). How Facebook broke its culture. *Unit Structures* Retrieved 2008, October 3, from http://chimprawk.blogspot.com/2006/09/how-facebook-broke-its-culture.html.

Stutzman, F. (2008). Facebook datasets and private chrome. *Unit Structures* Retrieved 2008, September 30, from http://fstutzman.com/2008/09/29/facebook-datasets-and-private-chrome/.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 10*(5), 557–570.

Wellman, B., & Berkowitz, S. D. (1988). *Social structures: A network approach*. Cambridge: University Press Cambridge.

Zimmer, M. (2006). More on Facebook and the contextual integrity of personal information flows. *michaelzimmer.org* Retrieved 2008, October 3, from http://michaelzimmer.org/2006/09/08/more-on-facebook-and-the-contextual-integrity-of-personal-information-flows/.

Zimmer, M. (2008a). More on the "Anonymity" of the Facebook dataset—It's Harvard College. *michaelzimmer.org* Retrieved October 3, 2008, from http://michaelzimmer.org/2008/10/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/.

Zimmer, M. (2008b). On the "Anonymity" of the Facebook dataset. *michaelzimmer.org* Retrieved September 30, 2008, from http://michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/.