digital methods initiative

Big Data

Digital Methods Summer School 2011

27 June - 8 July 2011
New Media & Digital Culture
University of Amsterdam
Turfdraagsterpad 9
1012 XT Amsterdam
Rooms 0.13 & 0.04

I Studying Big Data

Rogers, Richard (2012). *Digital Methods*. Cambridge, MA: MIT Press. Excerpt from forthcoming book.

Bollier, David (2010). "The Promise and Peril of Big Data." (excerpt) http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Latour, Bruno (2009). "Tarde's idea of quantification." *The Social After Gabriel Tarde: Debates and Assessments*. Ed. Mattei Candea. Routledge, London, pp. 145-162. http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf

Shirky, Clay (2005). "Ontology is Overrated." http://www.shirky.com/writings/ontology_overrated.html

II Digital Humanities

Anderson, Chris (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

Berry, David M. (2011). "The Computational Turn: Thinking About the Digital Humanities." Culture Machine. Vol 12. http://www.culturemachine.net/index.php/cm/article/view/440/470

Manovich, Lev. "Trending: The Promises and the Challenges of Big Social Data." Debates in the Digital Humanities, edited by Matthew K. Gold. The University of Minnesota Press, forthcoming 2012. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Tooling Up for Digital Humanities: Digitization. Stanford University, 2011. http://toolingup.stanford.edu/?page_id=123

III Big Data Spaces

Queries

Mohebbi et. al. (2011). "Google Correlate Whitepaper." http://correlate.googlelabs.com/whitepaper.pdf

Streams

Berry, David (2011). *Philosophy of Software*. London: Palgrave Macmillan. Excerpt.

Platforms

Caplan, Paul (2011). "Software Tunnels Through the Rags 'n Refuse: Object Oriented Software Studies and Platform Politics". Presented at Platform Politics conference in Cambridge, 13 May 2011. http://theinternationale.com/blog/2011/05/software-tunnels-through-the-rags-n-refuse/

Commentspaces

Shah and Yazdani Nia (2011). "Politics 2.0 with Facebook – Collecting and Analyzing Public Comments on Facebook for Studying Political Discourses." *The Journal of Information Technology and Politics Annual Conference*. http://scholarworks.umass.edu/jitpc2011/3/

Fora

Bernstein et al (2011). "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community." *Association for the Advancement of Artificial Intelligence*. http://projects.csail.mit.edu/chanthropology/4chan.pdf

Location-based services

Cheng, Zhiyuan et. al (2011). "Exploring Millions of Footprints in Location Sharing Services." *Fifth International AAAI Conference on Weblogs and Social Media*. 17-21 July 2011, Barcelona, Spain. http://students.cse.tamu.edu/kyumin/papers/cheng11icwsm.pdf

IV Big Data Spaces: Wikileaks

Lynch, Lisa (2010). "A Toxic Archive of Digital Sunshine: Wikileaks and the Archiving of Secrets." Paper presented at the MIT6 conference, Cambridge, MA. http://web.mit.edu/comm-forum/mit6/papers/Lynch.pdf

Lovink, Geert and Patrice Riemens (2010). "Twelve Theses on WikiLeaks," *Eurozine.* http://www.eurozine.com/articles/2010-12-07-lovinkriemens-en.html

Stalder, Felix (2010). "Contain this! Leaks, whistle–blowers and the networked news ecology." *Eurozine.* http://www.eurozine.com/articles/2010-11-29-stalder-en.html

Sterling, Bruce (2010), "The Blast Shack," Webstock. http://www.webstock.org.nz/blog/2010/the-blast-shack/

Žižek, Slavoj (2011). "Good Manners in the Age of WikiLeaks," The London Review of Books, 33:2, 9-10. http://www.lrb.co.uk/v33/n02/slavoj-zizek/good-manners-in-the-age-of-wikileaks

Video

Manuel Castells (2011), "From WikiLeaks to Wiki-revolutions," SONIC Media, Technology and Society Speaker Series, Lecture on 8 March 2011 at Northwestern University, Video Registration available at: http://lecture.soc.northwestern.edu/mediasite/Viewer/?peid=4e192796ace943fabf8172b463ce74381d

I

Studying Big Data

Rogers, Richard (2012). *Digital Methods*. Cambridge, MA: MIT Press. Excerpt from forthcoming book.

Bollier, David (2010). "The Promise and Peril of Big Data." (excerpt) http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Latour, Bruno (2009). "Tarde's idea of quantification." *The Social After Gabriel Tarde: Debates and Assessments*. Ed. Mattei Candea. Routledge, London, pp. 145-162. http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf

Shirky, Clay (2005). "Ontology is Overrated." http://www.shirky.com/writings/ontology_overrated.html

7

After Cyberspace: Big Data, Small Data

The aim of this chapter is to examine the question of the Web's status as source of data, big and small.[1] The argument made here is an overall case for taking the Internet far more seriously than we have in the past, specifically in terms of what it as to offer for social and cultural research. The first step is to dispense with the ideas of cyberspace and the virtual as primary points of departure for Internet-related research, or in fact reposition those terms to reflect the conceptual opportunities they currently offer. Cyberspace, with its origins in science fiction literature and its legacy in cybercultural studies, most recently has become a specific realm of inquiry in Internet security studies, with the U.S. military, for example, creating a 'Cyber Command' in 2009, and in the same year the U.S. Air Force phrasing its mission as "fly, fight and win in air, space and cyberspace."[2] Similarly, the virtual, a term with a rich theoretical history, refers less to the Internet generally than it does most poignantly to virtual worlds such as Second Life and game environments such as World of Warcraft.[3] Studies of the virtual, as in those of specific types of online worlds and environments, would thus become a subset of Internet-related research just as cyberspace studies also refers to niche areas, cyber-war together with cyber-espionage and cyber-crime.[4]

Second, we may wish to reconsider the primacy of treating the Web as a site for the study of amateur production practices and user-generated content, for we would be re-running the 'online quality' debates. Arguably the Internet has seen recurrences of such debates, the first in the 1990s on the value of information online, where the Web was treated as a rumor mill and as breeding ground for conspiracy theory.[5] In the mid 2000s, the second such debate referred to the quality of content, where the Web became a free amateur content space

[1] The title is a variation on Derek de Solla Price's classic monograph, *Little Science, Big Science* (1963), where at issue, among others, are the relationship between status and authority and the size of data as well as apparatus to handle it. An earlier version of this chapter was written with Erik Borra.
[2] U.S. Air Force, 2009.
[3] Shields, 2003; van Doorn, 2009.
[4] Information Warfare Monitor, 2009.
[5] Marres and Rogers, 2000.

threatening the paid professional.[6] I would like to argue that the Web continues to pose problems for the analysis of content. It disappoints those in search of traditional markers of quality, and an underlying interpretive apparatus more specifically.[7] Especially with the decline of surfing and with it hypertext as literary theory underpinning a surfer's space, the Web has lost some of its hermeneutic productivity.[8] I would like to put forward that the Web nowadays invites more of the stance popular culture and television researchers radically put forward decades ago regarding their relatively new object of study – that one can read and diagnose cultural concerns from the medium, beginning with the study of the TV Guide, and what is on (and what is not). The question, however, becomes the means and techniques by which to do so. As I have argued, 'digital methods' provides means distinctive from other contemporary approaches to the study of digital materials, such as cultural analytics and culturnomics, which both make use of the digitized over the natively digital, as I come to.[9] The approach put forward here also may be distinguished from dominant points of deparature to date in the computational social sciences and digital humanities, where there is a big data *drang*, or an urge to work with large data sets, and a desire to create accompanying infrastructures for them. Throughout the book I have sought to consider the productivity of modest tools and small web data, too. Here I partially embrace big data in the sense of a proposal to work with them, albeit with a digital methods outlook, as I explain.

Third, and most extensively, the argument recognizes that the Internet has reputational issues for researchers accustomed to thinking of it as cyberspace and virtual realm, as domain of rumors and self-publication, as well as, most recently, a site of messy data. The quality of information debate that was followed by the quality of (amateur) content debate has become the quality of data debate. Initial concerns had to do with incomplete Web archives, as discussed in chapter three. Additionally, multiple dates on Webpages and search engine indexing were unable to provide accurate results for date range queries; longitudinal analysis, a marker of quality research, was thought to be doomed.[10] Questions now arise

[6] Keen, 2007; Thelwall and Hasler, 2007.
[7] Hayles, 2004.
[8] Elmer, 2001
[9] Apart from content analysis, another candidate approach would be user studies.
[10] Hellsten et al., 2006.

about the robustness of so-called user-generated data such as social bookmarks, tags, comments, likes and shares.[11] How could it all possibly be cleaned?

The Web as data

New Web data sources are increasingly becoming available, yet they suffer from an overall reputational problem, in the long line of such problems online. The concerns about Web data still stem from their historical association with a free-for-all 'cyberspace,' and an epistemology based on a do-it-yourself medium of self-publication, with an absence of editors performing quality control. Indeed, traditionally, the Web has been thought of as a source of *doxa*, or opinion yet to be substantiated. The substantiation of opinion 'floating around on the Internet' would require leaving the medium, for instance, by making a phone call, obtaining an eyewitness account, etc. Thus Web accounts could not stand alone as sources; they also could not serve as the crucial second source, confirming a claim, in a journalistic sense. Web claims required grounding off-line.

Above I mentioned how the Web initially arrived on the scene as infrastructure awaiting content. In a sense, it was a space of data only – command, communication and traffic data – with the content (traditionally speaking) under preparation. 'Under construction' sites or pages may be regarded as sources of nostalgia these days, like other aesthetics of the 1990s, such as starry blue nights as Website backdrops, the animated gif, or 'random site' links which invite surfers to navigate to unknown territories, and to jumpcut to another hyperspace, themes I touched on earlier.[12] The "I'm feeling lucky" button on Google is such a hyperspace jump-cut, and the names of the browsers (Netscape Navigator, Internet Explorer and Safari) elicit being a helmsman online. This is our Web 0.5 cyberspace, the precursor to what is now becoming historicized in business circles as Web 1.0 (info-Web) and Web 2.0 (social Web), and which I strove to stretch out through the discussion of the history of the Web as hyperspace, cyberspace, space of shapes (sphere, network) and grounded or locative space. In any case, in the new history of Web 1.0 followed by Web 2.0, the Web is seen as a succession of two stable software versions. Each version has had particular quality debates associated with it. Whether it is associated with fandom, porn and

[11] Thelwall et al., 2005.
[12] Espenschied and Lialina, 2009.

aliens, with imposters, conspiracy and self-publishers, in the first version, or with amateur production practices, user generated content and lolcats, in the second – the Internet has not borne sources of great standing. Thus it is likely that any introduction of it as source of study worthy of attention would be met with similar skepticism. Nowadays to view the Web as data sets for social and cultural research is to be confronted with a variety of issues about messy data. The Webometrician Mike Thelwall summarized the challenges of employing the Web for research as follows:

"One [issue] is the messiness of Web data and the need for data cleansing heuristics. The uncontrolled Web creates numerous problems in the interpretation of results (…). Indeed, a skeptical researcher could claim the obstacles (…) are so great that all Web analyses lack value. [O]ne response to this (…) is to demonstrate that Web data correlate significantly with some non-Web data in order to prove that the Web data are not wholly random."[13]

Here the general reputation problem about quality online is transformed, initially, into the question of how to clean up the data, since there is a lack of uniformity in how users fill in forms, fields, boxes, bars and other text entry spaces. In a sense the (unedited) Web is viewed as one large 'free text' space. There are misspellings. There are too few conventions. For example, different tags are used for the same content, with no clever means of disambiguating contents of such mass and scale. To Thelwall, this state of affairs makes many researchers simply renounce the Web as source, unless data sets come whole (all transactions in Second Life) and one studies online culture only (amateur production practices and user-generated content). Finally, if Web data are to be used, he argues that one must introduce offline data for comparative purposes – Web data should be correlated with non-Web data.

To the issues above, I would like to add a concern in the remark made by David Lazer, in a key text on the computational turn in the social sciences: "Perhaps the thorniest challenges exist on the data side, with respect to access and privacy."[14] Here the Web data are tainted for their ultimate capacity to identify persons without consent or expectation to be

[13] Thelwall et al., 2005: 81.
[14] Lazer et al. 2009: 722.

identified. Indeed, anonymization of the identifiable people in the data may fail, as was the case in the well-known AOL data release when a list of an anonymized users' search queries became a puzzle for investigative journalists to piece together people's identity (in the first instance, an older lady's). The lessons learned from the data release have had further consequences for the tidiness of Web data. Certain Web data now come degraded by design.

Tidying the mess online

The sundry issues surrounding Web data, or at least the three Thelwall introduces (messiness, wholeness and off-line grounding of data), and the additional one from Lazer (anonymization), are being worked upon, though each 'solution' (to speak in enterprise software terms) reintroduces the complication that to date they have been addressed rather well by Google and other big data corporations. It is likely to exacerbate what one could call Google envy, that is, the capacity of search engine companies and social networking sites to collect data that approximates both the type and the scale social scientists would like to generate themselves, however much without the even finer grained texture that researchers may prefer (e.g., more demographic and ideological information). As if answering the calls issued by the authors cited above, Google Labs in 2011 made available server-side software that assists in cleaning data ("Google Refine") and software that correlates online data with offline data ("Google Correlate").[15] In the opening chapters I discussed the notion of scooping as it has been used in the sociology of science (as well as in journalism), whereby one's object of study comes to the conclusion the researcher had been working on, and publishes it first. The object of study thereby puts the researcher in the unexpected and sometimes unenviable position of having to confirm the object's prior art; Google has addressed the criticism that is being made of Web data, and has gone even further by inviting researchers to work with engine log data, in ways that differ both in form and in format from the AOL data release, as I come to. In doing so, the company follows the new media platform spirit I refered to earlier (make not the tool, but the tool-maker), and as such provides the underlying apparatus that enabled the making of the Google Flu Trends project, discussed in the opening chapters. Which queries and their places correlate with the findings made by the Centers for Disease Control and Prevention about the whereabouts

[15] Google Refine and Google Correlate were introduced in 2011 by Google Labs.

and incidence of flu? With Correlate, Google also may have initiated precisely what the compuational social scientists and Webometricians have called for: a large-scale data infrastructure of Web data (query logs), that one may use to compare with non-Web data (an offline baseline).

Digitized data and natively digital data

In the discussion of digital data, and their properties compared to others, information scientist Christine Borgman lists classic types of data as observational data, computational data, experimental data and records, and describes why they are considered of quality.[16] Good data are collected "as cleanly as possible and as early as possible in its life cycle;" they are captured regularly, and preferably over long periods of time.[17] Certain Web data, especially search engine logs, would fail miserably according to these criteria. Certain digitzed data sets would meet the criteria, however. As cases in point, I would like touch on two relatively novel undertakings in digital humanities, in order to make clear the current reliance on digitized data, which pass the above tests, and the challenges of natively digital data, which do not.

One program of cultural research relying on digitized data, cultural analytics, proposes to "start thinking of culture as data [...] that can be mined and visualized."[18] Indeed, new media theorist Lev Manovich and colleagues have performed longitudinal analyses of the changing properties of all of the front covers of *Time Magazine*, *Science* and *Popular Science Magazine*, as well as all Mark Rothko paintings. To the digitized artwork they apply computer vision techniques "to generate numerical descriptions of their structure and content" such as levels of grayscale, brightness, hue, saturation, and forms.[19] Another research undertaking along these lines, albeit larger, is called culturomics, a field of study of recent coinage that pursues a "quantitative analysis of culture" using as its initial corpora some 4% of all books published all time, now in digitized format from the Google scanning project.[20] In introducing the work, the founders of the initiative discuss the impossibility of actually

[16] National Science Board, 2005; Borgman, 2007; Borgman, 2010.
[17] Borgman, 2010: 13.
[18] Franklin, 2008: 1.
[19] Manovich, 2009:208; Hubner, 2010.
[20] Michel et al, 2010: 3.

reading the works they are now able to analyze through elaborate search. Generally speaking, the culturomics 'search as research' program examines the context and frequency of word usage over time and across world cultures, where there are intriguing lexicographical findings (American English is gradually taking over from British spelling of the same words) as well as broader cultural trends, such as an increasing proclivity to forget the past, or at least to refer to specific years in the past far less frequently. Celebrity is also becoming shorter-lived, in the sense that more and more celebrities are being referred to less and less as time goes by.[21]

Here I would like to introduce thought that considers the Web as more than an infrastructural platform for the storage of digitized data sets, yet also deploys insights from the study of digitized materials related above. That is, can the Web furnish its own data sets, and eventually become a privileged place from which to read and diagnose cultural and social change? In this regard, social theorist Noortje Marres has turned on its head the debate surrounding the reputation problem of the Web (and the messiness of its data) by arguing that "Web services incorporate social science methods like textual analysis, social network analysis, and geospatial analysis, arguably ordering data for (...) research."[22] In other words, it is the method incorporated into the Web services that is worthy of study for its ability to make sense of Web data. May the Web deliver structured data after all? In this way of thinking, Web services – search engines, collaboratively authored wikis and social networking platforms – become the data filterers, cleaning and ordering the data for end usage as well as perhaps for research.

I would like to concentrate on search engine log data, for it requires negotiated access, big infrastructure as well as skill in being able to handle big data. Often it is thought that the manner in which data are collected by devices such as search engines is unobtrusive. That is, for search engines and arguably also for collaboratively authored wikis and user-populated platforms such as social networking sites, the Web may be considered a site to make unobtrusive measurements, i.e., those less affected by the effects of other methodological

[21] Michel et al., 2010; Bohannon, 2010.
[22] Marres, 2009.

apparatuses.[23] Whilst the data may be collected without a clear and present methodological apparatus in front of the user, or the presence of an ethnographer casually listening in, any use of the data must be viewed against the backdrop of the scandal surrounding the data release in 2006 by the Internet service provider and search engine, AOL. AOL Research, the AOL scientific unit, made available some 650,000 users' engine queries over a three-month period for researchers, with lists of queries per numbered user and other data, such as URL clicked.[24] The *New York Times* was able to 'de-anonymize' one of the users by looking at the list of the queries, and performing relatively straightforward detective work, identifying user 441749 as Thelma Arnold, a 62-year-old woman residing in Lilburn, Georgia, USA.[25] She had searched for people with her last name, Arnold, and services in her home town, Lilburn (see Table one). All the search data that AOL made available to the scientific community during the SIGIR information retrieval conference in Seattle, was taken offline a few days after the release.[26] The data were described by computer scientist Jon Kleinberg as tainted, for "the number of things it reveals about individual people seems much too much."[27] Considered "naive" in the manner in which it was released, the data also were organized in a way that arguably invited detective work, and de-anonymization.[28] From a research point of view, the log data were formatted to facilitate a particular style of research, namely inquiry into search engine user behavior and ultimately the improvement of personalized search (see Table two).

Table Two. Fields and field descriptions of the AOL search engine user data set, 500k User Session Collection. Source: gregsadetsky.com/aol-data/US500k_README.text

AnonID - an anonymous user ID number.
Query - the query issued by the user, case shifted with most punctuation removed.
QueryTime - the time at which the query was submitted for search.
ItemRank - if the user clicked on a search result, the rank of the item on which they clicked

[23] Webb et al., 1966; Lewis et al., 2008; Spink et al., 2008.
[24] Pass et al., 2006.
[25] Barbaro and Zeller, 2006.
[26] Hafner, 2006.
[27] Anderson, 2006: 1.
[28] Poblete et al., 2008.

is listed.
ClickURL - if the user clicked on a search result, the domain portion of the URL in the clicked result is listed.

In particular, the AOL data fields that were provided suggest engine effectiveness research with relatively straightforward questions. Do engine users click the top results returned to them? This information is available in the ItemRank field. More nuanced is the question, do users find what they are looking for?[29] If users click on the top result for a query (ItemRank), and there is only one query in their session (QueryTime), the user will have found what he or she is looking for, and the engine presumably would be doing its job the best. Long sessions, repeatedly reformulated queries and URLs clicked that appear on the second or third results page (11-29 ItemRank) would prompt questions about the effectiveness of the search engine. A line of inquiry to follow would be an investigation into the characteristics of such queries.

In reaction to peculiar query-builders, researchers could seek to build elements into the algorithm that would help engines return results to such users (see Table Three). For example, users may pose engines actual questions. They may make remarks to engines. They may converse with them, or even confess to them. One such user is the subject of the "true and heartbreaking (search) history" of AOL search user 711391, entitled *I love Alaska*, a video art project that is named after a query.[30] Indeed 711391's query style is precisely the type that the AOL researchers envisaged when they wrote the read me text inviting the scientific community to work with "real query log data that is based on real users."[31] The data were to be employed for work on personalization as well as "query reformulation."[32] However, for our purposes here, it is of interest how the data were formatted: lists of queries per numbered individual. Indeed, the data set of the individual user 711391 and others may fill in the idea of engine logs' furnishing a "database of intentions."[33] According to the logs, AOL search engine users are victims of despicable acts. Or, they may be plotting revenge

[29] van Couvering, 2007.
[30] Engelberts and Plug, 2008.
[31] AOL, 2006; Pass et al., 2006.
[32] AOL, 2006.
[33] Battelle, 2005; Lernert and Sander, 2008.

and planning other acts, from the sensitive to the heinous.[34] In their complaint to the Federal Trade Commission, about the data release and what it revealed about users who did not expect to have their queries made public, and perhaps be personally identifiable, the Electronic Frontier Foundation writes:

The disclosure (...) made public extremely sensitive search queries such as "how to tell your family you're a victim of incest," "surgical help for depression," "how to kill your wife," "men that use emotional and physical abandonment to control their partner," "suicide by natural gas," "how to make someone hurt for the pain they caused someone else," "revenge for a cheating spouse," will I be extradited from ny to fl on a dui charge," and "my baby's father physically abuses me" (EFF, 2006: 4-5).

Table Three. AOL Search User 711391 Queries, 2006. Source: "User 711391," *Smith Magazine*, 10 August 2006, http://www.smithmag.net/2006/08/10/user-711391/.

cannot sleep with snoring husband … god will fulfill your hearts desires … online friendships can be very special … people are not always how they seem over the internet … gay churches in houston tx … who is crystal bernard romantically linked with … is crystal bernard bisexual … men need encouragement … how many online romances lead to sex … how many online romances lead to sex in person … the bible says be kind to one another … i cant stand dr. phil or his wife … is george clooney gay … how can i be a good example to an unsaved friend … farting preacher … who's the hottest porn star … devotions for women … hillary swanke nude … best nude scenes of 1999 … how to take your body measurements … jake gyllenhaal is hot … bleached pubes … oprah gained weight lately … star jones hubby is a flaming homosexual … how to make a good first impression … accepting your body … why do i weigh so much though i am in shape … the lord's table bible study … how can i tell if spouse is spying on me while i'm online … tempted to have an affair … extra maritial affairs are not the answer … staying calm while meeting an online friend … guilt cheating spouses feel … bryce howard nude … what the bible says about worry … female pirate costumes … symptoms of bladder infection … god will show you

[34] EFF, 2006.

future events … symptoms of herpes of the tongue … i don't want my ex back … why do christian men cheat … don't contact an ex if you want to get over them … christian men that feel guilty about cheating on their wives … if you are upset can it cause bad dreams … and after you have suffered a little while god will make you stronger than ever … kelly ripa is so annoying … how to forgive yourself … how to recover from internet affairs … denise richards is a bitch … reason for constant bad dreams … having an affair is a waste of time … how to make a man want you …

Many of the (subsequent) commentaries on the search log data focused on the privacy breaches as well as on the opportunities for law enforcement, and such attention led search engine companies only to grow more wary of the use of search logs for academic research purposes.[35] A main lesson drawn from the AOL search log data release, as pointed out elsewhere, is the improbability of anonymization of search engine users (NAIVE, xx). Providing each user with a number, and a list of queries, does not mask identities, but rather invites detective work.

There are two further consequences of the AOL search log release of relevance to social researchers employing Web data, which should be covered. The first is that search engine companies subsequently made pledges to protect user privacy through data anonymization and data destruction directives, downgrading the data it collected and also making it shorter lived – two of the benchmarks of 'good data' as discussed above, and one of the reasons why the Web data would fail the test of quality hands down.[36] Proper names may be replaced with random characters, so as to comply with data retention laws. In the event, Google decided to anonymize their engine users (in the logs) by removing the last few digits of the IP address of the searcher, whilst Microsoft related that it would scrub the IP addresses entirely.[37] The IP address, however, is also a geo-location marker of the user, which concerns the second consequence of the AOL search log release. Instead of focusing on the individual

[35] EFF, 2006.
[36] It goes without saying that search log data could pass 'good data' quality tests if it were not regularly destroyed, and if it were not regularly degraded through the insertion of gobbly-gook characters in the place of proper names and parts of speech thought to be personally identifiable.
[37] Sullivan, 2008.

user (anonymized by applying a number to each or by scrubbing all or part of the IP address), log research may direct its sights onto what we call the places of queries. Where are users searching for particular terms, such as a candidate in an upcoming election? Could the location of such collective query behavior be made of relevance to research? Thus a social research imagination and outlook, as I have argued in this book, potentially would transform not only how Web data are studied but how they are made available. It would be of less interest to have a list of queries by an individual user, than to have a list of queries from a place.

The major breakthrough in this respect has been the Google Flu Trends project, which, in the seminal piece, found that it can "track influenza-like illness in a population (…) accurately [estimating] the current level of weekly influenza activity in each region of the United States."[38] Google Flu Trends thus shifted the attention away from the individual user's privacy, and away from search effectiveness or personalized search research more generally as the main work to be undertaken with search log data. Instead, queries become a means for detecting trends, which has been a by-product of search engines for some time, with such services as Google Zeitgeist, listing the most popular searches at a given time (and place). However, with the addition of the data of place (via IP address or zip code for registered users of a toolbar or other service), the outlook changes as do the interests from zeitgeist and a marketing mentality to flu and social research.

IP address scrubbing, as mentioned above, is a means of anonymizing users; in the Google case only unauthenticated users (those not logged in) had their IP addresses rubbed out. Authenticated users, on the other hand, agree to allow search companies to retain their data for the purposes of improving search. A set of Yahoo! Research's search engine log data, from Yahoo's registered users, largely mirrors the demographics of the U.S. population.[39] Thus, as undertaken with the allrecipes.com data, discussed in the preface to the book, one could study the distribution of specific cultural and social preferences, employing three

[38] Ginsberg et al. 2009:1012.
[39] Weber and Castillo, 2010.

variables: query terms (including volume), query locations (zip codes), and the date stamps of the queries.[40]

Here the interest extends to social and political search. One could imagine an array of query log research projects, such as the time and place of the use of hate speech or extremist language, where one could strive to ground the findings made from archived websites about the hardening of Dutch culture (described in the first chapter) through additional Web data (query logs). Here one grounds (or at least triangulates) web findings with web findings. One also could query for candidates in the run up to an election, and candidates coupled with social issues. One of the crucial challenges is to inquire into the Web's ability (through analysis of search engine query data) to provide the place and time as well as intensity of cultural preference and political expression, compared to other means of finding the same. Research focusing on the places, times and intensities of queries has the social research methodological outlook I wish to describe (as opposed to the online detective's or to the personalized search engine builder's), as discussed above, for one is formulating questions concerning attitudes and preference as opposed to ones that may ultimately reveal personally identifiable data.[41] Such a research undertaking involves 'big' Web data.

[40] Spink et al., 2008; Baeza-Yates et al., 2009; Silvestri, 2010.
[41] Ess, 2002; Jones et al., 2007; boyd, 2010; Dutton and Piper, 2010.

The Promise and Peril of Big Data

David Bollier
Rapporteur

THE ASPEN INSTITUTE

Communications and Society Program
Charles M. Firestone
Executive Director
Washington, DC
2010

To purchase additional copies of this report, please contact:

The Aspen Institute
Publications Office
P.O. Box 222
109 Houghton Lab Lane
Queenstown, Maryland 21658
Phone: (410) 820-5326
Fax: (410) 827-9174
E-mail: publications@aspeninstitute.org

For all other inquiries, please contact:

The Aspen Institute
Communications and Society Program
One Dupont Circle, NW
Suite 700
Washington, DC 20036
Phone: (202) 736-5818
Fax: (202) 467-0790

Charles M. Firestone
Executive Director

Patricia K. Kelly
Assistant Director

The Aspen Institute
One Dupont Circle, NW
Suite 700
Washington, DC 20036

Published in the United States of America in 2010
by The Aspen Institute

Printed in the United States of America

ISBN: 0-89843-516-1

10-001

1762/CSP/10-BK

Contents

Foreword, *Charles M. Firestone* vii

The Promise and Peril of Big Data, *David Bollier*

How to Make Sense of Big Data? 3

Data Correlation or Scientific Models? 4
How Should Theories be Crafted in an Age of Big Data? 7
Visualization as a Sense-Making Tool 9
Bias-Free Interpretation of Big Data? 13
Is More Actually Less? 14
Correlations, Causality and Strategic Decision-making 16

Business and Social Implications of Big Data 20
Social Perils Posed by Big Data 23

Big Data and Health Care 25
Big Data as a Disruptive Force (Which is therefore Resisted) 28
Recent Attempts to Leverage Big Data 29
Protecting Medical Privacy 31

How Should Big Data Abuses be Addressed? 33
Regulation, Contracts or Other Approaches? 35
Open Source Analytics for Financial Markets? 37

Conclusion 40

Appendix

Roundtable Participants 45

About the Author 47

Previous Publications from the Aspen Institute Roundtable on Information Technology 49

About the Aspen Institute Communications and Society Program 55

The Promise and Peril of Big Data

David Bollier

It has been a quiet revolution, this steady growth of computing and databases. But a confluence of factors is now making Big Data a powerful force in its own right.

Computing has become ubiquitous, creating countless new digital puddles, lakes, tributaries and oceans of information. A menagerie of digital devices has proliferated and gone mobile—cell phones, smart phones, laptops, personal sensors—which in turn are generating a daily flood of new information. More business and government agencies are discovering the strategic uses of large databases. And as all these systems begin to interconnect with each other and as powerful new software tools and techniques are invented to analyze the data for valuable inferences, a radically new kind of "knowledge infrastructure" is materializing. A new era of Big Data is emerging, and the implications for business, government, democracy and culture are enormous.

…a radically new kind of "knowledge infrastructure" is materializing. A new era of Big Data is emerging….

Computer databases have been around for decades, of course. What is new are the growing scale, sophistication and ubiquity of data-crunching to identify novel patterns of information and inference. Data is not just a back-office, accounts-settling tool any more. It is increasingly used as a real-time decision-making tool. Researchers using advanced correlation techniques can now tease out potentially useful patterns of information that would otherwise remain hidden in petabytes of data (a petabyte is a number starting with 1 and having 15 zeros after it).

Google now studies the timing and location of search-engine queries to predict flu outbreaks and unemployment trends before official

government statistics come out. Credit card companies routinely pore over vast quantities of census, financial and personal information to try to detect fraud and identify consumer purchasing trends.

Medical researchers sift through the health records of thousands of people to try to identify useful correlations between medical treatments and health outcomes.

Companies running social-networking websites conduct "data mining" studies on huge stores of personal information in attempts to identify subtle consumer preferences and craft better marketing strategies.

A new class of "geo-location" data is emerging that lets companies analyze mobile device data to make intriguing inferences about people's lives and the economy. It turns out, for example, that the length of time that consumers are willing to travel to shopping malls—data gathered from tracking the location of people's cell phones—is an excellent proxy for measuring consumer demand in the economy.

The inferential techniques being used on Big Data can offer great insight into many complicated issues, in many instances with remarkable accuracy and timeliness. The quality of business decision-making, government administration, scientific research and much else can potentially be improved by analyzing data in better ways.

But critics worry that Big Data may be misused and abused, and that it may give certain players, especially large corporations, new abilities to manipulate consumers or compete unfairly in the marketplace. Data experts and critics alike worry that potential abuses of inferential data could imperil personal privacy, civil liberties and consumer freedoms.

Because the issues posed by Big Data are so novel and significant, the Aspen Institute Roundtable on Information Technology decided to explore them in great depth at its eighteenth annual conference. A distinguished group of 25 technologists, economists, computer scientists, entrepreneurs, statisticians, management consultants and others were invited to grapple with the issues in three days of meetings, from August 4 to 7, 2009, in Aspen, Colorado. The discussions were moderated by Charles M. Firestone, Executive Director of the Aspen Institute Communications and Society Program. This report is an interpretive synthesis of the highlights of those talks.

How to Make Sense of Big Data?

To understand implications of Big Data, it first helps to understand the more salient uses of Big Data and the forces that are expanding inferential data analysis. Historically, some of the most sophisticated users of deep analytics on large databases have been Internet-based companies such as search engines, social networking websites and online retailers. But as magnetic storage technologies have gotten cheaper and high-speed networking has made greater bandwidth more available, other industries, government agencies, universities and scientists have begun to adopt the new data-analysis techniques and machine-learning systems.

Certain technologies are fueling the use of inferential data techniques. New types of remote censors are generating new streams of digital data from telescopes, video cameras, traffic monitors, magnetic resonance imaging machines, and biological and chemical sensors monitoring the environment. Millions of individuals are generating roaring streams of personal data from their cell phones, laptops, websites and other digital devices.

The growth of cluster computing systems and cloud computing facilities are also providing a hospitable context for the growth of inferential data techniques, notes computer researcher Randal Bryant and his colleagues.[1] Cluster computing systems provide the storage capacity, computing power and high-speed local area networks to handle large data sets. In conjunction with "new forms of computation combining statistical analysis, optimization and artificial intelligence," writes Bryant, researchers "are able to construct statistical models from large collections of data to infer how the system should respond to new data." Thus companies like Netflix, the DVD-rental company, can use automated machine-learning to identify correlations in their customers' viewing habits and offer automated recommendations to customers.

Within the tech sector, which is arguably the most advanced user of Big Data, companies are inventing new services such that give driving directions (MapQuest), provide satellite images (Google Earth) and consumer recommendations (TripAdvisor). Retail giants like Wal-Mart assiduously study their massive sales databases—267 million transactions a day—to help them devise better pricing strategies, inventory control and advertising campaigns.

Intelligence agencies must now contend with a flood of data from its own satellites and telephone intercepts as well as from the Internet and publications. Many scientific disciplines are becoming more computer-based and data-driven, such as physics, astronomy, oceanography and biology.

Data Correlation or Scientific Models?

As the deluge of data grows, a key question is how to make sense of the raw information. How can researchers use statistical tools and computer technologies to identify meaningful patterns of information? How shall significant correlations of data be interpreted? What is the role of traditional forms of scientific theorizing and analytic models in assessing data?

Chris Anderson, the Editor-in-Chief of *Wired* magazine, ignited a small firestorm in 2008 when he proposed that "the data deluge makes the scientific method obsolete."[2] Anderson argued the provocative case that, in an age of cloud computing and massive datasets, the real challenge is not to come up with new taxonomies or models, but to sift through the data in new ways to find meaningful correlations.

> At the petabyte scale, information is not a matter of simple three and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising—it just assumed that better data, with better analytic tools, would win the day. And Google was right.

Physics and genetics have drifted into arid, speculative theorizing, Anderson argues, because of the inadequacy of testable models. The solution, he asserts, lies in finding meaningful correlations in massive piles of Big Data, "Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without

hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

J. Craig Venter used supercomputers and statistical methods to find meaningful patterns from shotgun gene sequencing, said Anderson. Why not apply that methodology more broadly? He asked, "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?"

Conference participants agreed that there is a lot of useful information to be gleaned from Big Data correlations. But there was a strong consensus that Anderson's polemic goes too far. "Unless you create a model of what you think is going to happen, you can't ask questions about the data," said William T. Coleman. "You have to have some basis for asking questions."

Researcher John Timmer put it succinctly in an article at the Ars Technica website, "Correlations are a way of catching a scientist's attention, but the models and mechanisms that explain them are how we make the predictions that not only advance science, but generate practical applications."[3]

Hal Varian, Chief Economist at Google, agreed with that argument, "Theory is what allows you to extrapolate outside the observed domain. When you have a theory, you don't want to test it by just looking at the data that went into it. You want to make some new prediction that's implied by the theory. If your prediction is validated, that gives you some confidence in the theory. There's this old line, 'Why does deduction work? Well, because you can prove it works. Why does induction work? Well, it's always worked in the past.'"

Extrapolating from correlations can yield specious results even if large data sets are used. The classic example may be "My TiVO Thinks I'm Gay." *The Wall Street Journal* once described a TiVO customer who gradually came to realize that his TiVO recommendation system thought he was gay because it kept recommending gay-themes films. When the customer began recording war movies and other "guy stuff" in an effort to change his "reputation," the system began recommending documentaries about the Third Reich.[4]

Another much-told story of misguided recommendations based on statistical correlations involved Jeff Bezos, the founder of Amazon. To demonstrate the Amazon recommendation engine in front of an audience, Bezos once called up his own set of recommendations. To his surprise, the system's first recommendation was *Slave Girls from Infinity*—a choice triggered by Bezos' purchase of a DVD of *Barbarella*, the Jane-Fonda-as-sex-kitten film, the week before.

Using correlations as the basis for forecasts can be slippery for other reasons. Once people know there is an automated system in place, they may deliberately try to game it. Or they may unwittingly alter their behavior.

It is the "classic Heisenberg principle problem," said Kim Taipale, the Founder and Executive Director of the Center for Advanced Studies in Science and Technology. "As soon as you put up a visualization of data, I'm like—whoa!—I'm going to 'Google bomb' those questions so that I can change the outcomes." ("Google bombing" describes concerted, often-mischievous attempts to game the search-algorithm of the Google search engine in order to raise the ranking of a given page in the search results.[5])

The sophistication of recommendation-engines is improving all the time, of course, so many silly correlations may be weeded out in the future. But no computer system is likely to simulate the level of subtlety and personalization that real human beings show in dynamic social contexts, at least in the near future. Running the numbers and finding the correlations will never be enough.

Theory is important, said Kim Taipale, because "you have to have something you can come back to in order to say that something is right or wrong." Michael Chui, Senior Expert at McKinsey & Company, agrees: "Theory is about predicting what you haven't observed yet. Google's headlights only go as far as the data it has seen. One way to think about theories is that they help you to describe ontologies that already exist." (Ontology is a branch of philosophy that explores the nature of being, the categories used to describe it, and their ordered relationships with each other. Such issues can matter profoundly when trying to collect, organize and interpret information.)

Jeff Jonas, Chief Scientist, Entity Analytic Solutions at the IBM Software Group, offered a more complicated view. While he agrees

that Big Data does not invalidate the need for theories and models, Jonas believes that huge datasets may help us "find and see dynamically changing ontologies without having to try to prescribe them in advance. Taxonomies and ontologies are things that you might discover by observation, and watch evolve over time."

John Clippinger, Co-Director of the Law Lab at Harvard University, said: "Researchers have wrestled long and hard with language and semantics to try to develop some universal ontologies, but they have not really resolved that. But it's clear that you have to have some underlying notion of mechanism. That leads me to think that there may be some self-organizing grammars that have certain properties to them—certain mechanisms—that can yield certain kinds of predictions. The question is whether we can identify a mechanism that is rich enough to characterize a wide range of behaviors. That's something that you can explore with statistics."

How Should Theories be Crafted in an Age of Big Data?

If correlations drawn from Big Data are suspect, or not sturdy enough to build interpretations upon, how then shall society construct models and theories in the age of Big Data?

Patrick W. Gross, Chairman of the Lovell Group, challenged the either/or proposition that either scientific models or data correlations will drive future knowledge. "In practice, the theory and the data reinforce each other. It's not a question of data correlations versus theory. The use of data for correlations allows one to test theories and refine them."

That may be, but how should theory-formation proceed in light of the oceans of data that can now be explored? John Seely Brown, Independent Co-Chair of Deloitte Center for the Edge, believes that we may need to devise new methods of theory formation: "One of the big problems [with Big Data] is how to determine if something is an outlier or not," and therefore can be disregarded. "In some ways, the more data you have, the more basis you have for deciding that something is an outlier. You have more confidence in deciding what to knock out of the data set—at least, under the Bayesian and correlational-type theories of the moment."

But this sort of theory-formation is fairly crude in light of the keen and subtle insights that might be gleaned from Big Data, said Brown: "Big Data suddenly changes the whole game of how you look at the ethereal odd data sets." Instead of identifying outliers and "cleaning" datasets, theory formation using Big Data allows you to "craft an ontology and subject it to tests to see what its predictive value is."

He cited an attempt to see if a theory could be devised to compress the English language using computerized, inferential techniques. "It turns out that if you do it just right—if you keep words as words—you can compress the language by *x* amount. But if you actually build a theory-formation system that ends up discovering the morphology of English, you can radically compress English. The catch was, how do you build a machine that actually starts to invent the ontologies and look at what it can do with those ontologies?"

"…The more data there is, the better my chances of finding the 'generators' for a new theory."

John Seely Brown

Before huge datasets and computing power could be applied to this problem, researchers had rudimentary theories about the morphology of the English language. "But now that we have 'infinite' amounts of computing power, we can start saying, 'Well, maybe there are many different ways to develop a theory.'"

In other words, the data once perceived as "noise" can now be reconsidered with the rest of the data, leading to new ways to develop theories and ontologies. Or as Brown put it, "How can you invent the 'theory behind the noise' in order to de-convolve it in order to find the pattern that you weren't supposed to find? The more data there is, the better my chances of finding the 'generators' for a new theory."

Jordan Greenhall suggested that there may be two general ways to develop ontologies. One is basically a "top down" mode of inquiry that applies familiar philosophical approaches, using *a priori* categories. The other is a "bottom up" mode that uses dynamic, low-level data and builds ontologies based on the contingent information identified through automated processes.

For William T. Coleman, the real challenge is building new types of machine-learning tools to help explore and develop ontologies: "We

have to learn how to make data tagged and self-describing at some level. We have to be able to discover ontologies based on the questions and problems we are posing." This task will require the development of new tools so that the deep patterns of Big Data can be explored more flexibly yet systematically.

Bill Stensrud, Chairman and Chief Executive Officer of InstantEncore, a website that connects classical music fans with their favorite artists, said, "I believe in the future the big opportunity is going to be non-human-directed efforts to search Big Data, to find what questions can be asked of the data that we haven't even known to ask."

"The data *is* the question!" Jeff Jonas said. "I mean that seriously!"

Visualization as a Sense-Making Tool

Perhaps one of the best tools for identifying meaningful correlations and exploring them as a way to develop new models and theories, is computer-aided visualization of data. Fernanda B. Viégas, Research Scientist at the Visual Communications Lab at IBM, made a presentation that described some of the latest techniques for using visualization to uncover significant meanings that may be hidden in Big Data.

Google is an irresistible place to begin such an inquiry because it has access to such massive amounts of timely search-query data. "Is Google the ultimate oracle?" Viégas wondered. She was intrigued with "Google Suggest," the feature on the Google search engine that, as you type in your query, automatically lists the most-searched phrases that begin with the words entered. The feature serves as a kind of instant aggregator of what is on people's minds.

Viégas was fascinated with people using Google as a source of practical advice, and especially with the types of "why?" questions that they asked. For example, for people who enter the words "Why doesn't he…" will get Google suggestions that complete the phrase as "Why doesn't he *call*?", "Why doesn't he *like me*?" and "Why doesn't he *love me*?" Viégas wondered what the corresponding Google suggestions would be for men's queries, such as "Why doesn't *she*…?" Viégas found that men asked similar questions, but with revealing variations, such as "Why doesn't she *just leave*?"

Viégas and her IBM research colleague Martin Wattenberg developed a feature that visually displays the two genders' queries side by side, so that the differences can be readily seen. The program, now in beta form, is meant to show how Google data can be visually depicted to help yield interesting insights.

While much can be learned by automating the search process for the data or by "pouring" it into a useful visual format, sometimes it takes active human interpretation to spot the interesting patterns. For example, researchers using Google Earth maps made a striking discovery—that two out of three cows (based on a sample of 8,510 cattle in 308 herds from around the world) align their bodies with the magnetic north of the Earth's magnetic field.[6] No machine would have been capable of making this starting observation as something worth investigating.

Viégas offered other arresting examples of how the visualization of data can reveal interesting patterns, which in turn can help researchers develop new models and theories. Can the vast amount of data collected by remote sensors yield any useful patterns that might serve as building blocks for new types of knowledge? This is one hope for "smart dust," defined at Wikipedia as a "hypothetical wireless network of tiny microelectromechanical (MEMS) sensors, robots, or devices that can detect (for example) light, temperature, or vibration."

To test this idea with "dumb dust"—grains of salt and sand—scientists put the grains on the top of a plate to show how they respond when the frequency of audio signals directed at the bottom of the plate is manipulated. It turns out that the sand self-organizes itself into certain regular patterns, which have huge implications for the study of elasticity in building materials. So the study of remote sensor data can "help us understand how vibration works," said Viégas. It engendered new models of knowledge that "you could take from one domain (acoustics) and sort of apply to another domain (civil engineering)."

Visualization techniques for data are not confined to labs and tech companies; they are becoming a popular communications tool. Major newspapers such as *The New York Times* and *The Washington Post* are using innovative visualizations and graphics to show the significance of otherwise-dry numbers. Health websites like "Patients Like Me" invite people to create visualizations of their disease symptoms, which then become a powerful catalyst for group discussions and further scrutiny of the data.

Visualizations can help shine a light on some improbable sorts of social activity. Viégas describes a project of hers to map the "history flow" of edits made on Wikipedia articles. To learn how a given Wikipedia entry may have been altered over the course of months or years, Viégas developed a color-coded bar chart (resembling a "bar code" on products) that illustrates how many people added or changed the text of a given entry. By using this visualization for the "abortion" entry, Viégas found that certain periods were notable for intense participation by many people, followed by a blank "gash" of no color. The gash, she discovered, represented an "edit war"—a period of intense disagreement about what the text should say, followed by vandalism in which someone deleted the entire entry (after which Wikipedia editors reverted the entry to the preexisting text).

The visualizations are useful, said Viégas, because they help even the casual observer see what the "normal" participation dynamics are for a given Wikipedia entry. They also help researchers identify questions that might be explored statistically—for example, how often does vandalism occur and how quickly does the text get reverted? "This visualization tool gave us a way to do data exploration, and ask questions about things, and then do statistical analyses of them," said Viégas.

Stensrud agreed that visualization of Big Data gives you a way "to find things that you had no theory about and no statistical models to identify, but with visualization it jumps right out at you and says, 'This is bizarre.' "

Or as Lise Getoor, Associate Professor in the Department of Computer Science at the University of Maryland, articulated, visualizations allows researchers to "'explore the space of models' in more expansive ways. They can combine large data sets with statistical analysis and new types of computational resources to use various form functions in a systematic way and explore a wider space."

After exploring the broader modeling possibilities, said Getoor, "you still want to come back to do the standard hypothesis testing and analysis, to make sure that your data is well-curated and collected. One of the big changes is that you now have this observational data that helps you develop an initial model to explore."

Kim Taipale of the Center for Advanced Studies in Science and Technology warned that visualization design choices drive results every bit as much as traditional "data-cleaning" choices. Visualization tech-

niques contain embedded judgments. In Viégas' visualization models of Wikipedia editing histories, for example, she had to rely upon only a fraction of the available data—and the choices of which entries to study ("abortion" and "chocolate," among others) were idiosyncratic. Taipale believes disputes about the reliability of visualization designs resemble conversations about communications theory in the 1950s, which hosted similar arguments about how to interpret signal from noise.

Jesper Andersen, a statistician, computer scientist and Co-Founder of Freerisk, warned about the special risks of reaching conclusions from a single body of data. It is generally safer to use larger data sets from multiple sources. Visualization techniques do not solve this problem. "When you use visualization as an analytic tool, I think it can be very dangerous," he said. "Whenever you do statistics, one of the big things you find is spurious correlations"—apparent relationships or proximities that do not actually exist.

"You need to make sure the pattern that you *think* is there, is actually there," said Andersen. "Otherwise, the problem gets worse the bigger your data is—and we don't have any idea how to handle that in visualization because there is a very, very thin layer of truth on the data, because of tricks of the eye about whether what you see is actually there. The only way that we can solve this problem right now is to protect ourselves with a model."

So how can one determine what is accurate and objective? In a real-world business context, where the goal is to make money, the question may be moot, said Stephen Baker, *Business Week* journalist and author of *The Numerati.* "The companies featured in Amazon's recommendations don't have to be right. They just have to be better than the status quo and encourage more people to buy books—and in that way, make more money for the company," he said.

Baker noted that companies are often built "on revenue streams that come from imprecise data methods that are often wrong." The company may or may not need to decide whether to "move from what works to truth." It may not be worth trying to do so. This leads Baker to wonder if "truth could be just something that we deal with in our spare time because it's not really part of the business model."

Bias-Free Interpretation of Big Data?

Andersen's point is part of a larger challenge for those interpreting Big Data: How can the numbers be interpreted accurately without unwittingly introducing bias? 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?"

"Cleaning the data"—i.e., deciding which attributes and variables matter and which can be ignored—is a dicey proposition, said Jesper Andersen, because "it removes the objectivity from the data itself. It's a very opinionated process of deciding what variables matter. People have this notion that you can have an agnostic method of running over data, but the truth is that the moment you touch the data, you've spoiled it. For any operation, you have destroyed that objective basis for it."

"'Bad data' is good for you."

Jeff Jonas

The problems of having an objective interpretation of data are made worse when the information comes from disparate sources. "Every one of those sources is error-prone, and there are assumptions that you can safely match up two pieces together. So I think we are just magnifying that problem [when we combine multiple data sets]. There are a lot of things we can do to correct such problems, but all of them are hypothesis-driven."

Responding to Andersen, Jeff Jonas of the IBM Software Group believes that "'bad data' is good for you. You *want* to see that natural variability. You *want* to support dissent and disagreement in the numbers. There is no such thing as a single version of truth. And as you assemble and correlate data, you have to let new observations change your mind about earlier assertions."

Jonas warned that there is a "zone" of fallibility in data, a "fuzzy line" between actual errors and what people choose to hear. For example, he said, "My brother's name is 'Rody' and people often record this as 'Rudy' instead. In this little zone, you can't do peer review and you can't read everybody's mind. And so to protect yourself, you need to keep natural variability and know where every piece of data comes from—and then

allow yourself to have a complete change of mind about what you think is true, based on the presence of new observations."

Or as Bill Stensrud of InstantEncore put it, "One man's noise is another man's data."

Is More Actually Less?

One of the most persistent, unresolved questions is whether Big Data truly yields new insights—or whether it simply sows more confusion and false confidence. Is more actually less?

Perhaps Big Data is a tempting seduction best avoided, suggested Stefaan Verhulst, Chief of Research at the Markle Foundation. Perhaps "less is more" in many instances, he argued, because "more data collection doesn't mean more knowledge. It actually means much more confusion, false positives and so on. The challenge is for data holders to become more constrained in what they collect." Big Data is driven more by storage capabilities than by superior ways to ascertain useful knowledge, he noted.

"One man's noise is another man's data."

Bill Stensrud

"The real challenge is to understand what kind of data points you need in order to form a theory or make decisions," said Verhulst. He recommends an "information audit" as a way to make more intelligent choices. "People quite often fail to understand the data points that they actually need, and so they just collect everything or just embrace Big Data. In many cases, less is actually more—if data holders can find a way to know what they need to know or what data points they need to have."

Hal Varian, Chief Economist at Google, pointed out that small samples of large data sets can be entirely reliable proxies for the Big Data. "At Google, we have a system to look at all the data. You can run a day's worth of data in about half an hour. I said, no, that's not really necessary. And so the engineers take one-third of a percent of the daily data as a sample, and calculate all the aggregate statistics off my representative sample."

"I mean, the reason that you've got this Big Data is you want to be able to pick a random sample from it and be able to analyze it. Generally,

you'll get just as good a result from the random sample as from looking at everything—but the trick is making sure that it's *really* a random sample that is representative. If you're trying to predict the weather in New England from looking at the weather patterns in California, you'll have a problem. That's why you need the whole system. You're not going to need every molecule in that system; you might be able to deal with every weather station, or some degree of aggregation that's going to make the analysis a lot easier."

Bill Stensrud took issue with this approach as a general rule: "If you know what questions you're asking of the data, you may be able to work with a 2 percent sample of the whole data set. But if you don't know what questions you're asking, reducing it down to 2 percent means that you discard all the noise that could be important information. What you really want to be doing is looking at the whole data set in ways that tell you things and answers questions that you're not asking."

"The more people you have playing with the data, the more people are going to do useful things with it."

Kim Taipale

Abundance of data in a time of open networks does have one significant virtue—it enables more people to crunch the same numbers and come up with their own novel interpretations. "The more people you have playing with the data, the more people are going to do useful things with it," argued Kim Taipale.

The paradox of Big Data may be that it takes more data to discover a narrow sliver of information. "Sometimes you have to use *more* to find *less*," said Jeff Jonas of IBM Software Group. "I do work helping governments find criminals within. You really don't want to stare at less data. You want to use more data to find the needle in the haystack, which is really hard to find without a lot of triangulation. But at some point, less becomes more because all you are interested in doing is to prune the data, so that you can stare at the 'less.'"

Esther Dyson, Chairman of EDventure Holdings, believes that sifting through "more" to distill a more meaningful "less" represents a huge market opportunity in the future. "There is a huge business for third parties in providing information back to consumers in a form that is meaningful," she said. One example is a company called Skydeck,

which helps you identify your cell phone calling patterns, based on the data that your phone company provides on your behalf.

…sifting through "more" to distill a more meaningful "less" represents a huge market opportunity….

Esther Dyson

The lesson of Big Data may be "the more abundance, the more need for mediation," said Stefaan Verhulst. There is a need for a "new mediating ecosystem."

John Liechty, Associate Professor of Marketing and Statistics at Pennsylvania State University, agreed: "It really comes down to what tools we have to deal with Big Data. We're trying to get to a system where you can begin to extract meaning from automated systems…. Less is more only if we are able to reduce large sets of data down, and find ways to think about the data and make decisions with it. Ultimately, you have to have some extraction of the data in order to deal with it as a human being."

Correlations, Causality and Strategic Decision-making

The existence of Big Data intensifies the search for interesting correlations. But correlation, as any first-year statistics student learns, does not establish causality. Causality requires models and theories—and even they have distinct limits in predicting the future. So it is one thing to establish significant correlations, and still another to make the leap from correlations to causal attributes. As Bill Stensrud put it, "When you get these enormously complex problems, I'm not sure how effective classic causal science ends up being. That's because the data sets are so large and because it is difficult to establish causality because of the scale of the problem."

That said, there are many circumstances in which correlations by themselves are eminently useful. Professor Lise Getoor of the University of Maryland pointed out that for tasks like collaborative filtering, group recommendations and personalization, "correlations are actually enough to do interesting things."

For Sense Networks, Inc., which evaluates geo-location data for mobile phone providers, establishing correlations is the primary task. "We analyze really large data sets of location data from mobile

phones and carriers and handset manufacturers," said Greg Skibiski, Co-Founder and Chief Executive Officer of the company. "So we see tens of millions of people moving around, and really all we care about, in the end, is the correlations. The problem is, we have to make some really core theory decisions at the very, very beginning of [analyzing] these data sets."

For example, said Skibiski, how should analysts define "place?" Is place defined by the amount of time that people spend there, or by the number of visits that they make daily, weekly or monthly? Or does one try to establish a more subjective, idiosyncratic definition?

In the end, Skibiski said the size of a database tends to resolve such definitional quibbles: "If the 'lift curves' look good and the false-negatives and false-positives match up, that's the end of the story for us." However the techniques are refined, it is clear that they are enabling new sorts of inferences to emerge. A recent M.I.T. study found that geo-location data patterns can successfully predict people's future locations and social interactions.[7]

Correlations can be functional and useful in stimulating sales and making money, but they can also be highly imperfect. "If I order books for myself and my wife through Amazon," said John Seely Brown, "there are two different sets of data to be looked at, so sometimes Amazon will have to decide that, well, maybe there are two sets of buyers here." Bill Stensrud agreed, "Everything I buy from Amazon is a present for somebody else, and so their recommendation engine is meaningless to me. Some day they'll figure that out."

"Google doesn't know how many Jeff Jonas's there are," said Jeff Jonas of IBM Software Systems. "If you can't do correlations at atomic level construction [counting discrete identifiable units], you can't really do any form of meaningful predictions because you can't get trajectory or velocity [from the data]."

Even though correlations are inherently limited as predictors, they can be useful in many different ways. Some new businesses are based on sharing real-time correlations of data. City Sense tabulates the most active night life spots in a city, based on mobile phone and taxi traffic, giving subscribers a near real-time idea of "where the action is" at that very moment. Rapleaf, a San Francisco company, sifts through its massive pile of data from social networking websites to make suggestive

correlations, such as the association between one's friends and credit risk, reports Greg Skibiski. Even if your personal financial indicators give you a credit risk score of 550, but your friends have scores of 650, then your actual credit risk may well be closer to 650, according to some data analysts.

Data correlations are also useful in provoking interpretive stories, said John Seely Brown. "The quality of the story really matters in making sense of the data. You start to see stories clash against each other and get passed around, so that they may actually generate new insights." Data correlations can provoke people to develop a "new ecology of stories," which itself may shed light on the numbers.

For Joi Ito, the Chief Executive Officer of Creative Commons, the search for correlations is a trap to be avoided, at least in his capacity of a computer security expert and a venture capitalist. Ito says he is "always looking for unpredictable things that you can use opportunistically." As a venture capitalist, he is looking for the "subversive outlier" whose ideas could have a big upside. From a security perspective, Ito says he wants to be alert to the *unexpected* forms of intrusion and deceit, not to the ones whose correlations can be easily discovered using computers.

"I'm always very aggressively going outside of my comfort zone and looking for that tiny little data point that the statisticians won't see," said Ito. "Then I amplify it like crazy so that I can make as much money as quickly as possible. When you do that kind of analysis on, say, terrorist networks, you have to understand that Hezbollah is actively trying to continuously come up with patterns that they think you won't predict."

"Remember," said Ito, "the same technology that we're using to analyze Big Data enables these other actors to become more actively random. The people who are outliers, who used to sort of behave randomly, now have access to the same tools as the rest of us and are looking at the same data.

"People like me don't even look at the data. We go randomly to places that are completely absent of any data and we test and then we jump. That's why I'm in the Middle East—because it's completely random. [Ito recently moved to Abu Dhabi.] It's a big hole in which you can mess around and maybe find something. If you do find something,

then you start creating your own patterns and hook them back into the [mainstream]. *Then* you create this huge arbitrage. But the way that I do it is completely non-analytical. The more analytical you become, the more likely you're going to end up bumping into somebody who is already there. So it's much better to be completely random."

…linear extrapolations are "a little bit like saying, in 1950, 'What's the business model for big band music?'"

Jordan Greenhall

Ito's *modus operandi* may be especially well-suited to our times, in which data trends are not necessarily linear. As Jordan Greenhall explained, linear extrapolations are "a little bit like saying, in 1950, 'What's the business model for big band music?'" The human brain has a relatively high elasticity, he said, and the different experiences and technologies of today's generation mean that its brain neurology actually differs from that of previous generational cohorts. "As a result," said Greenhall, "we can't extrapolate from our own expectations of what makes sense to us, to what may make sense to a younger generation. So any decision made today has to plan forward ten years to make sure that it makes sense in the future reality."

To Greenhall, the pace of cultural (if not neurological) change is so great that we must recognize "the non-linearity of the space that we are dealing with." Simple correlations will be very crude tools—sensible by the terms of a more stable, classical worldview, but more problematic in their capacity to limn the future.

"Big Data is about exactly *right now*, with no historical context that is predictive," said Ito. "It's predictive of a linear thing—but you can use data collection to discover non-linearity as well. For example, I look on Twitter every day for any word that I don't know. Then I search Twitter for a whole feed and go to all the links. I always discover a new trend every morning that is completely non-linear from anything that I expected—because if I expect it, I just gloss over it…. It's important not to be obsessed with the old models that come from the old data. It's more important to be ignorant enough to come up with a new model of the future."

Business and Social Implications of Big Data

However one may quarrel about interpretive methodologies, there is little question that Big Data can help identify emerging trends, improve business decision making and develop new revenue-making strategies. Hal Varian, Chief Economist at Google, says that the growth of "computer-mediated transactions"—in which a computer sits between every buyer and seller—means that companies can "do many, many extra things."

"We have a lot of tools to look at data," said Varian. "Our basic operating procedure is to come up with an idea, build a simulation, and then go out and do the experimentation. At any given moment, Google is running hundreds and hundreds of experiments on both the search side [of data] and the ad side. We use a lot of different variables, and trade them off against others, sometimes explicitly and sometimes implicitly. If you're getting a 1 percent or 2 percent lift every two or three weeks, then after a few years you can build up an advantage."

Many innovative uses of Big Data could be called "now-casting," said Varian. This term refers to the use of real-time data to describe contemporaneous activities *before* official data sources are available. "We've got a real-time variable, Google search queries, which are pretty much continuous," said Varian. "Even if all you've got is a contemporaneous correlation, you've still got a six-week lead on the reported values" for certain types of data.

One of the most noted examples of now-casting is a service known as Google Flu Trends. By tracking the incidence of flu-related search terms, this Google spinoff service can identify possible flu outbreaks one to two weeks earlier than official health reports. When the Google data are correlated with actual flu cases compiled by the Centers for Disease Control, the Google estimates are 97 percent to 98 percent accurate.[8]

Varian noted that Google search queries for jobs and welfare can also indicate future economic trends. When first-time claims for unemployment benefits drop, for example, it has historically signaled the end of a recession. Google data can reveal such trends a week or two earlier than official government statistics. In fact, Varian has made the rounds in Washington "to make the case that government agencies should use Google tools to better draw current snapshots of consumer sentiment, corporate health and social interests," according to *The Washington Post*.[9]

As Varian told the Aspen conference participants, "In about nine months, Fed Chairman Ben Bernanke is going to have to decide whether to raise interest rates. He will look at a whole lot of variables—the last month's economic reports, retail sales data, initial claims from unemployment, and so on. To the extent that this data is more up-to-date, you could potentially make a better decision about whether to move on one issue or another."

"To make money, you've got to predict two things—what's going to happen and what people *think* is going to happen."

Hal Varian

American Express has used its sizeable database of consumer behavior to identify early trends and craft appropriate responses. For example, Amex has found that people who run up large bills on their American Express card and then register a new forwarding address in Florida have a greater likelihood to declare bankruptcy. That is because Florida has one of the most liberal bankruptcy laws, which makes it a favorite destination for debtors who are financially troubled. Identifying such correlations in the data—a soaring credit card balance and a re-location to Florida—can trigger an inquiry into the actual solvency of the cardholder.

There are many types of real-time data streams that can now be assembled and analyzed. Besides search engine queries, data for credit card purchases, the trucking and shipping of packages, and mobile telephone usage are all useful bodies of information. Much of this data is becoming available on a near real-time basis, which leads Varian to predict that credit card data will be compiled and sold on a daily basis at some point in the future. "Real-time economic indicators" will be possible, he said. "The hope is that as you take the economic pulse in real time, you'll be able to respond to anomalies more quickly."

By revealing a new genre of ultra-timely information, now-casting enables new types of arbitrage. If a company or investor can use real-time data to out-perform the market by just a few percentage points, it can reap that much more revenue in the marketplace. "The problem is not whether your predictions are more accurate, it's whether they beat the consensus," said Varian. "To make money, you've got to predict two things—what's going to happen and what people *think* is going to happen. You only make money by beating that spread."

"Playing the percentages" can be especially important in advertising and marketing, several participants noted. As more consumers migrate from traditional mass media to online media, questions arise. How should marketing budgets be allocated in this marketing landscape? What has greater influence on consumer purchases—public relations or advertising?

Aedhmar Hynes, Chief Executive Officer of Text 100 Public Relations, reported that "for high-impact brands" in the United States—that is, brands where the purchase is a large expenditure that consumers may research or ponder—"the impact of public relations on the brand was 27 percent, whereas the impact of advertising on such purchases was less than 1 percent. The reverse was true for low-impact buying decisions. If you're going to buy a piece of gum, the likelihood is that advertising will influence that decision far more than if you're going to buy a notebook computer."

Hynes speculated that data-analysis might be especially influential in making more effective media buys. But she also wondered if the interpretations would be useful in non-U.S. countries where the use of computers and search engines is less pervasive than in the U.S.

Certainly one new marketing frontier that Big Data will enable is the use of real-time data correlations to drive business decisions. Jacques Bughin of McKinsey & Company reported that his firm had discovered "that the time frame between search and buy has been reduced in some market segments." This suggests a greater opportunity to influence potential buyers.

It is also possible to discover some non-intuitive correlations in consumer buying patterns, such as the fact that people who do search-engine queries for "weddings" also tend to do searches for "diets." (An apocryphal correlation asserts that people who search for "diapers" also search for "six packs of beer," a hypothetical correlation made in a speech that later ripened into an urban legend.)

Some correlations are entirely verified, yet extremely difficult to interpret. What are we to make of the fact that in the three days preceding the bankruptcy of Bear Stearns, the nightlife patterns in seven cities—people staying out late in restaurants and bars—was a "five Sigma event," according to Sense Networks, Inc., meaning a significant deviation from the statistical mean. "A lot of people were out extremely late on those evenings, like you've never seen in years of data," said Greg Skibiski.

According to Jacques Bughin of McKinsey and Company, the real insight is that Big Data has the potential to discover new laws of macrobehaviors totally overlooked with the paucity of data of the past. Although social influence is and remains large, a new type of market power behavior has emerged, one that is not necessarily firm driven, but consumer driven. Today, relatively few consumers are able to influence others via social media and other interactive platforms. Businesses that are able to target or link to those influencers will have an edge, says Bughin. For instance, more and more firms using this new power curve are opening their business systems to users and suppliers to co-create products and services, or they are leveraging their brand and platforms for third parties (think Apple with the iphone). The more companies that master the skills for open collaboration with users—and successfully deliver—the higher the probability to leverage the influencers to their benefit.

Social Perils Posed by Big Data

As Big Data becomes a more common tool in corporate decisions, a number of new social perils arise. The most obvious is the risk of privacy violations. "Is personalization something that is done *to* you or *for* you?" wondered Kim Taipale of the Center for Advanced Studies in Science and Technology. A business with economic motives is driving the process of data-driven personalization, but consumers have far less knowledge of what is going on, and have far less ability to respond. The benefits of personalization tend to accrue to businesses but the harms are inflicted on dispersed and unorganized individuals, Taipale noted.

Marc Rotenberg, Executive Director of the Electronic Privacy Information Center, admits that there are two sides to personalization. When Amazon and iTunes use their databases of consumer purchases to make recommendations to prospective customers, most people welcome the advice. It may help them identify just the book or music that they want. On the other hand, "people start getting very uneasy when buying suggestions are made based on how much we know about this particular person, a practice that takes us into the realm of behavioral targeting"—the "my TiVO thinks I'm gay" phenomenon.

One independent survey of adult Internet users—by two professors at the University of Pennsylvania and the University of California, Berkeley, in September 2009—found that two-thirds of users object

to online tracking by advertisers. Respondents particularly disliked behavioral advertising, in which commercial websites tailor ads based on an individual's Web behavior. "I do think we're at the cusp of a new era, and the kinds of information that companies share and have today is nothing like we'll see ten years from now," said Professor Joseph Turow, the lead author of the study. "The most important thing is to bring the public into the picture, which is not going on right now."[10]

Citizens are also legitimately worried about Internet service providers (ISPs) who may use "deep packet inspection" techniques to analyze the data flowing through their wires, to determine what websites you may be visiting and what purchases you may be making. "In recent years," said Rotenberg, "ISPs have recognized that there is commercial value in their networks, beyond any security issues. Some of the same tools that can be used to identify spam can be used to figure out who's interested in buying a new SUV or who's planning on traveling." One solution might be to allow ISPs to use deep packet inspection for assuring the security of their networks, but to prohibit use of the data for commercial purposes, he said.

"Vendors are using Big Data to try to acquire the consumer," said Bill Stensrud, Chairman and Chief Executive Officer of InstantEncore, "and they are doing that by using technologies that are beyond the reach of the consumer by orders of magnitude." He noted that three responses have been suggested—counter-responses by hackers to harass companies that violate their privacy; schemes to "monetize" people's private data so that they can control it and sell it; and government regulation to protect individual privacy.

"None of these answers leave me very satisfied," said Stensrud. "I guess one of the questions that I have is how does the consumer get armed to fight on an even ground with the big companies who can make nano-second stock market trades and personalize their marketing?"

Joi Ito of Creative Commons took issue with the whole framing of the discussion as one between vendors and consumers; there are non-market actors who are influential as well. "We don't use the word 'consumer' in our group [Creative Commons], which acts collectively in the way that people in the hacker, open source and Wikipedia worlds do. We believe that there are ways for us to take control, so that your business models don't matter. *We're* in charge."

Ito cited the hundreds of thousands of fake Twitter accounts that hackers created in the course of a few hours after Ashton Kutcher and CNN announced their hopes of being the first to amass one million Twitter followers. A website of disruptive hackers, 4chan, quickly gamed the system to amass a huge following for a notorious criminal. For Ito, attacks by "smart mobs" demonstrate the ability of non-commercial actors to influence trends—"a power that is getting stronger and stronger and stronger," he said.

Big Data and Health Care

As researchers apply extreme inference techniques to public health, disease research, drug research, genetic engineering, and much else, the implications are both heartening and frightening. Identifying new correlations in data can improve the ways to develop drugs, administer medical treatments and design government programs. But it can also introduce new frustrations and complexities because solutions must overcome existing incentive systems for physicians, insurers and patients.

Patients are beginning to take charge of their own health care by doing research about their injuries and illnesses and joining social networks….

Stefaan Verhulst, the Chief of Research at the Markle Foundation, gave an overview of the health care trends that involve Big Data. The first, most significant is the American Recovery and Reinvestment Act, the so-called ARRA, which is President Obama's stimulus package for dealing with the financial crisis. Approximately $19 billion of that stimulus legislation is earmarked to encourage physicians to adopt electronic medical recordkeeping systems. (Some people argue that, if additional funds managed by the Department of Health and Human Services and other sources are included, some $26 billion is being directed at this issue.)

This law is potentially significant because the American health care system is plagued by a fragmented, inefficient system of paper-based recordkeeping. Digitizing records could make health care recordkeeping vastly more efficient and versatile, especially in assembling large pools of data and evaluating them for new insights.

Conclusion

Big Data presents many exciting opportunities to improve modern society. There are incalculable opportunities to make scientific research more productive, and to accelerate discovery and innovation. People can use new tools to help improve their health and well-being, and medical care can be made more efficient and effective. Government, too, has a great stake in using large databases to improve the delivery of government services and to monitor for threats to national security.

Large databases also open up all sorts of new business opportunities. "Now-casting" is helping companies understand the real-time dynamics of certain areas of life—from the diffusion of diseases to consumer purchases to night-life activity—which will have many long-term reverberations on markets. New types of data-intermediaries are also likely to arise to help people make sense of an otherwise-bewildering flood of information. Indeed, data-intermediaries and interpreters could represent a burgeoning segment of the information technology sector in the years ahead.

But Big Data also presents many formidable challenges to government and citizens precisely because data technologies are becoming so pervasive, intrusive and difficult to understand. How shall society protect itself against those who would misuse or abuse large databases? What new regulatory systems, private-law innovations or social practices will be capable of controlling anti-social behaviors—and how should we even define what is socially and legally acceptable when the practices enabled by Big Data are so novel and often arcane?

These are some of the important open questions posed by the rise of Big Data. This report broaches some of the more salient issues that should be addressed. In the coming years, government, business, consumers and citizen groups will need to devote much greater attention to the economic, social and personal implications of large databases. One way or another, our society will need to take some innovative, imaginative leaps to ensure that database technologies and techniques are used effectively and responsibly.

Notes

1. Randal Bryant, Randy H. Katz and Edward D. Lazowska, "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society," December 2008, pp. 1-15, at http://www.cra.org/ccc/docs/init/Big_Data.pdf.

2. Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 23, 2008, at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

3. John Timmer, "Why the Cloud Cannot Obscure the Scientific Method," Ars Technica, June 25, 2008, at http://arstechnica.com/old/content/2008/06/why-the-cloud-cannot-obscure-the-scientific-method.ars.

4. Jeffrey Zaslow, "If TiVO Thinks You Are Gay, Here's How to Set It Straight," *Wall Street Journal*, November 26, 2002, p. 1, at http://online.wsj.com/article_email/SB1038261936872356908.html.

5. See http://en.wikipedia.org/wiki/Google_bomb.

6. Thomas H. Maugh II, "Cows Have Magnetic Sense, Google Earth Images Indicate," *Los Angeles Times*, August 26, 2008, at http://articles.latimes.com/2008/aug/26/science/sci-cows26.

7. N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255-268.

8. Alice Park, "Is Google Any Help in Tracking an Epidemic?" *Time* magazine, May 6, 2009, at http://www.time.com/time/health/article/0,8599,1895811,00.html.

9. Cecilia Kang, "Google Economist Sees Good Signs in Searchers," *The Washington Post*, September 12, 2009, at http://www.washingtonpost.com/wp-dyn/content/article/2009/09/11/AR2009091103771.html?hpid%3Dmoreheadlines&sub=AR.

10. Stephanie Clifford, "Tracked for Ads? Many Americans Say No Thanks," September 30, 2009, p. B3.

11. See, e.g., Richard Platt, M.D., et al., "The New Sentinel Network—Improving the Evidence of Medical-Product Safety," *New England Journal of Medicine*, August 13, 2009, p. 645-647.

12. Esther Dyson, "Health 2.0 could shock the system," *Financial Times*, August 12, 2009, p. 41.

13. Gina Kolata, "Forty Years War: Lack of Study Volunteers Hobbles Cancer Fight," *The New York Times*, August 3, 2009, p. A1, at http://www.nytimes.com/2009/08/03/health/research/03trials.html?scp=1&sq=lack+of+study+volunteers&st=nyt.

14. For more, see http://www.sensenetworks.com/press/wef_globalit.pdf.

15. Daniel Roth, "Road Map for Financial Recovery: Radical Transparency Now!" *Wired*, February 23, 2009, available at http://www.wired.com/techbiz/it/magazine/17-03/wp_reboot.

16. See http://www.ce-nif.org.

Tarde's idea of quantification*

A chapter for Mattei Candea (editor)
The Social After Gabriel Tarde: Debates and Assessments
Bruno Latour, Sciences Po

"[Thanks to statistics] public broadsheets will be to the social world what the sensory organs are to the organic world"
(**Lois de l'imitation**).

Numbers, numbers, numbers. Sociology has been obsessed by the goal of becoming a quantitative science. Yet it has never been able to reach this goal because of what it has defined as being quantifiable within the social domain. The work of Gabriel Tarde has been resurrected for many reasons. One of them, to be sure, is an acknowledgement of the diminishing returns of "social explanations". In my view, however, it would be wrong to limit Tarde's contribution to the theme of the "end of the social".[i] If he has become so interesting, if he is read with such great avidity today, it is also because he engaged sociology, and more generally the human sciences —history, geography, archaeology, social psychology and above all economics— with a different definition of what it is for a discipline to be *quantitative*. (He also had an alternative definition of what it is to be a *science*, but this is another subject).

In the last century, the schism between those who deal with numbers and those who deal with qualities has never been bridged. This is a fair statement given that so many scholars have resigned themselves to being partitioned into those who follow the model of the "natural" sciences, and those who prefer the model of the "interpretive" or "hermeneutic" disciplines. All too often, fields have been divided between number crunching, devoid (its enemies claim) of any subtlety; and rich, thick, local descriptions, devoid (its enemies say) of any way to generalize from these observations. Many domains have abandoned the hope of proving any point by transforming quantities into qualities, and qualities into quantities. Many in history or anthropology, as well as in sociology or psychology have tried, but at every occasion, the difficulties of reconciling the two types of proof have been so great that it is impossible to transition smoothly from one to the other. Many have despaired, as a consequence, of ever being able to develop a

scientific social science; while others have claimed that this goal is no longer desirable, that the best that can be hoped for is to obtain some political or literary effects on readers.

What is so refreshing in Tarde (more than a century later!) is that he never doubted for a minute that it was possible to have a scientific sociology —or rather, an "inter-psychology", to use his term. And he espoused this position without ever believing that this should be done through a superficial imitation of the natural sciences.

1) Social sciences are more quantitative than natural ones

Tarde's reasoning goes straight to the heart of the matter: the natural sciences grasp their object from far away, and, so to speak, in bulk. A physicist deals with trillions upon trillions of gas molecules, a biologist with billions of cells. It is therefore quite normal that they should rely on a rough outline of the "societies" of gas and cells to make their observations. (Remember that for Tarde "everything is a society"). Resemblance is what appeals to the natural scientist. Individual differences can be safely neglected. Although the very distinction between a law or structure and its individual components is acceptable in natural sciences, it cannot be used as a universal template to grasp all societies. The distinction is an artifact of distance, of where the observer is placed and of the number of entities they are considering at once. The gap between overall structure and underlying components is the symptom of a *lack of* information: the elements are too numerous, their exact whereabouts are unknown, there exist too many hiatus in their trajectories, and the ways in which they intermingle has not been grasped. It would therefore be very odd for what is originally a *deficit of information* to be turned into the universal *goal* of any scientific inquiry. In the face of such a striking gap, it would make much more sense to tackle this limitation and to try to get more detailed information, instead of glowing with the belief that one has reached the level of an exact science.

Physicists and biologists may be forgiven for having so little information since for the most part, they continue to access their objects of study from a great distance. But those who deal with type of societies composed of many fewer elements, societies that can be observed from the inside, do not have this excuse. Consider sociologists who study human societies. (After all, what are a handful of billions of fellow humans when compared to the number of animalcules teaming in a drop of water?) Given the immense privilege of having proximity to their objects of study, sociologists should not be (mis)led into imagining that there could be a strict distinction between structural features and individual or sub-individual components.[ii] If they are, they have been engaged in the rather silly task of becoming voluntarily estranged from the societies they are studying. It implies that they are attempting to grasp them in the same way that astronomers deal with stars or biologists with cells. And yet, if the latter must handle their subject matter from far away, it is not because it is especially "scientific" to do so. It is because they have no other way to reach their objects of investigation.

Paradoxically, those in sociology who try to ape the natural sciences have mistaken the latter's constitutive lack of information for their principal virtue! Yet what is really scientific is to have enough information so as to *not* have to fall back upon the makeshift approximation of a structural law, distinct from what its individual components do. What is perfectly acceptable for "sociologists" of stars, atoms, cells and organisms, is inacceptable for the sociologists of the few billions of humans, or for the economists of a few millions of transactions. For in the latter cases, we most certainly have, or we should at least strive to possess, the information needed to dissolve the illusion of the structure.

This first point about replacing the idea of what a science should be is crucial to grasp the deeper reasons for the opposition between Tarde and Durkheim. The tension is not simply due to a difference of attitude, as though one was more inclined to follow the individual agents while the other became obsessed by the relationship of the actor to the overall society. To be sure, this opposition is present, as the encounter between Tarde and Durkheim reproduced in this volume, has made quite clear.[iii] Beyond this, however, the tension is a consequence of a completely different way of calibrating what should be expected from any science of any society. Durkheim deals only with human societies and borrows his ideal of science from natural scientists with whom he has little occasion to collaborate since, for him, human societies should remain radically different from biological and physical ones. Tarde's position is the reverse; for him there exist only societies. Human societies are but a particular subset of these societies because they exist in so few copies. But since human societies are accessible through their most intimate features, social scientists have no need to let natural scientists dictate what their epistemology should be.

The paradox is that it is Durkheim who imitates the natural sciences while at the same time distancing his discipline most radically from theirs. Meanwhile, Tarde, because he does not distinguish the ideal of science by separate domains, takes the greatest liberty in moving away from the customary ways of the natural sciences for presenting their objects. The shibboleth that distinguishes their attitudes is not that one is "for society" while the other is "for the individual actor". (This is what the Durkheimians have quite successfully claimed so as to bury Tarde into the individual psychology he always rejected.) The distinction is drawn by whether one accepts or does not accept that a structure can be qualitatively distinct from its components. In response to this test question, Durkheim answers "yes" for both kinds of societies. Tarde says "yes", for natural societies (for there is no way to do otherwise), but "no" for human societies. For human societies, and for only human societies, we can do *so much more*.

2) Bypassing the notion of structure

In the tired old debate pitting a naturalistic versus an interpretative social science, a strange idea appears: that if we stick to the individual, the local, the situated, you will detect only qualities, while if we move towards the structural and towards the distant, we will begin to gather quantities. For Tarde the situation is almost exactly the opposite: the more we get into the intimacy of the individual, the more discrete quantities we'll find; and if we move away from the individual

towards the aggregate we might begin to *lose* quantities, more and more, along the way because we *lack the instruments* to collect enough of their quantitative evaluations. And this is the second reason why a science of society is possible for Tarde: the very heart of social phenomena is quantifiable because individual monads are constantly evaluating one another in simultaneous attempts to expand and to stabilize their worlds. The notion of expansion is coded for him in the word "desire", and stabilization in the word "belief" (more on this below).[iv] Each monad strives to *possess* one another.

Most social scientists remain limited to the study of qualities when they handle only one entity, and quantification *begins*, so to speak, once they have collected large numbers of those entities. To the contrary, for Tarde, quantification began with the individual and was very *difficult to maintain* when shifting to aggregates. Consider this passage:

> *"But before we speak, think, or act as "they" speak, think, or act in our world, we begin by speaking, thinking, and acting as "he" or "she" does. And this "he" or "she" is always one of our own near acquaintances. Beneath the indefinite they, however carefully we search, we never find anything but a certain number of he's and she's which, as they have increased in number, have become mingled together and confused" p.25.*

He then added:

> *"The impersonal, collective character is thus the* product *rather than the* producer *of the infinitely numerous individual characters; it is their composite photograph, and must not be taken for their mask." (27-28).*[v]

The relationship of the element to the aggregate is not the same as that of an ingredient to a structure. A "composite photograph"[vi] is not more than its individual components; it is not a law of behavior to which they should submit, *minus individual variations*. An "impersonal collective character" does not produce a behavior; it is itself produced by a multiplicity of individual innovations. There is nothing more in the accumulation of traits than there is in the multiplicity of individual components; but there definitely a lot less since elements become "mingled together and confused". Or rather, there is perhaps more in the "they" than in the "he" and "she", but this is because one monad has succeeded in expressing and possessing the whole. [vii] So, if we jump too quickly to the idea that an altogether different type of entity has taken over the action, just what that supplement is will becomes obscured. It is readily apparent that *confusion increases* when moving from the "he" to the "they", instead of decreasing as might be expected following an introductory class in the methodology of the social sciences: "Gather more examples; forget individual traits; see things from farther away; from above; in bulk not in detail; for goodness sake, put it into a frame". According to Tarde, from those well meaning pieces of advice, only disorientation can ensue!

Does this mean that we should always stick to the individual? No, but we should find ways to gather the individual "he" and "she" without *losing out on* the specific ways in which they are able to mingle, in a standard, in a code, in a bundle of customs, in a scientific discipline, in a technology —but never in some overarching society. The challenge is to try to obtain their aggregation without

either shifting our attention at any point to a whole, or changing modes of inquiry. Composite photography is a very crude and primitive way that confuses all the criminals into a single type. Let's try to find a better, more sensible, and above all, more *traceable* way of doing social science. And it does exist: those who commit crimes *imitate* one another. They have to learn from one another, *modus operandi* per *modus operandi*, crime per crime, trick by trick.[viii] And the same can be said of the Ministry of Justice or of the police. By assembling file after file, case after case, identification after identification, they end up producing "types of criminal" out of which the science of criminology will emerge.[ix] Following the "imitative rays" will render the social traceable from beginning to end without limiting us to the individual, or forcing a leap up to the level of a structure.[x]

Tarde is often presented as a man with one idea —imitation. It is true that he became famous following the publication of his book, *The Laws of Imitation*, in 1890.[xi] Nevertheless, it is important to understand that imitation is not an obsession of his. Nor is his point a psychological argument about how humans imitate one another, as if Tarde had generalized from some observations to the rest of his social psychology.[xii] The situation was rather the opposite. He was searching for a route by which to bypass the ill-conceived notion of structure when he stumbled upon a plausible vocabulary, borrowed in part from medicine, and later from psychology.[xiii] Imitation, that is, literally, the "epidemiology of ideas". With this notion, he could render the social sciences scientific enough by following individual traits, yet without them getting confused when they aggregated to form seemingly "impersonal" models and transcendent structures. The term "imitation" may be replaced by many others (for instance, monad, Actor-Network or entelechy), provided these have the equivalent role: of tracing the ways in which individual monads conspire with one another without ever producing a structure.[xiv]

In opposition to the entire century of social theory that followed it, this often quoted passage summarizes what is at stake for sociology to be scientific:

> *"But, no matter how intimate, how harmonious a social group is, never do we see emerging* ex abrupto*, in the midst of its astonished associates, a collective self, which would be real and not only metaphoric, a sort of marvelous result, of which the associates would be the mere conditions. To be sure, there is always an associate that represents and personifies the group in its entirety, or else a small number of associates (the ministers in a State) who, each under a particular aspect, individualize in themselves the group in its entirety. But this leader, or those leaders, are always also members of that group, born from their own fathers and mothers and not born collectively from their subjects or their constituency." p. 68.* Monadologie et sociologie.

For Tarde, if we were to believe that the first duty of social science is to "reconcile the actor and the system" or to "solve the quandary of the individual versus society", we would have to abandon all hope of ever being scientific. This is tantamount to aping the natural sciences which are perfectly alright in getting by with discovering a structure and neglecting minor individual variations because they are much too far to observe whether or not a "collective self" emerges *ex abrupto* from "its astonished associates". Fortunately, in the case of human sciences,

we know this emergence is different. We can verify every day, alas, that "leaders" are "born from fathers and mothers" and not "collectively". This forces us to discover the real conduits through which any group is able to emerge. For instance, we might search for how associates might "individualize in themselves the group in its entirety" through legal or political vehicles. Once we have ferreted out what makes this phase transition possible we will be able to see with clarity, the difference between "individualizing a group" and "being an individual in a collective structure".[xv] Each case requires a completely different feel for the complex ecology of the situation.

If this requirement strikes you as less demanding, less empirically exacting, less "scientific" than the search for a structure, then it means that you will have abandoned, in effect, the search for quantification, for the real *quanta* that lie at the heart of each monad.

3) Tracing the social world anew

There is a third reason why Tarde believed in the scientific program of the social sciences: he thought that we could invent the instrumentation for capturing the inner quantification of individual entities. This implies that the great quandary of "the actor and the system" is but a consequence of a very patchy statistical apparatus; or, to put it more bluntly, that you have the social theory of your statistics.

Tarde, who is often derided for having been "literary" instead of "scientific" knew very well what he was talking about. The misunderstanding is always the same. We confuse quantitative social sciences with a historical way of doing statistics.[xvi] But those techniques have changed immensely over the years. Rather than trying to eliminate individual variations so that they don't perturb the overall result, many other ways of handling them have been discovered. The situation of the natural sciences, where individual variations remain inaccessible to any direct inquiry, and are far too numerous to record, is in no way the same as for the social sciences. For human societies, there is no reason to limit quantification to only some of the ways of doing statistics.[xvii]

This assessment of statistics is so close to the heart of Tarde's work that he actually moved from his position as a judge in the provincial town of Sarlat (which he had occupied since 1875 before moving to Paris in 1894), after proposing alternative ways of assembling, interpreting and publishing, criminal, civil and commercial statistics to the Minister of Justice. (By then Tarde was already well known as a criminologist.)[xviii] As he argued there is no reason to consider individual variations as deviations from a more stable law that statistics was charge with educing out of the morass of chaotic data. Individual variations are the only phenomenon worth looking at in societies for which there are comparatively few elements. We have (or should have) full access to the aggregated dynamic. What is called a "structural law" by some sociologists is simply the phenomenon of aggregation: the formatting and standardization of a great number of copies, stabilized by imitation and made available in a new form, such as a code, a dictionary, an institution, or a custom. According to Tarde, if it is wrong to consider individual variations as though they were deviations from a law, it is

equally wrong to consider individual variations as the only rich phenomenon to be studied by opposition with (or distance from) statistical results. It is in the nature of the individual agent to imitate others. What we observe either in individual variations or in aggregates are just two detectable *moments* along a trajectory drawn by the observer who is following the fate of any given "imitative ray". To follow those rays (or "actor-networks" if you feel more comfortable with some updated vocabulary) is to encounter, depending on the moment, individual innovations and then aggregates, followed afterwards by more individual innovations. It is the trajectory of what circulates that counts, not any of its provisional steps.

The importance of trajectory is the most clear with intellectual arguments, a domain of great fascination to Tarde. It is in the study of scientific practice that one can see how useless it is to drown individual contributions into statistical means (scientists are so few and so far between that any "whole" is provisional). Nonetheless, it would be just as silly to deny that, from individually made arguments in specific journals and specific times, aggregates are not produced in the end, by consensus formation and paradigm entrenchments that deeply modify how an individuals finds their way in an argument. This result is in no way due to a structural law suddenly overwhelming the diversity of negligible individual positions (the *ex abrupto* we saw above). In each of the scientists' laboratories, for each of the issue at hand, each individual converts to the consensus each for his or her peculiar reason. Later, they may once again re-differentiate themselves from any established dogma.

Of course, the wonderful thing about science, contrary to criminology or fashion where the traces are much more elusive, is that there exists —thanks to footnotes, references and citations— an almost *uninterrupted set of traces*, that allows us to move from each individual innovation, up to the aggregate, and then back again to the individual resistance that can develop in response to a given paradigm.

> *"When, during some universal exhibition, we realize retrospectively how means of transportation have appeared in succession, since the time of the sedan-chair and the chariot until the time of the suspension carriage, the locomotive, the automobile and the bicycle, we behave much like the naturalist in a museum who compares the long series of vertebrates along the course of geological times from the lancelet to man. And yet, there is this difference that in the first case we are able to date exactly the appearance of most links in the chain and determine very precisely the invention and inventor from which each specimen comes from, while in the second case we are restricted to mere conjectures about the way a species transformed itself into another*" (Psychologie économique, volume 1, p. 12.)

We can understand from this passage what was meant earlier in pointing to the distinction between structure and ingredient as being due to a deficiency of information. If the researcher is in possession of this information, this chain of invention, this "imitative ray", then there is *no reason why* they can not follow the individual innovation *as well* as the aggregates, smoothly. If there is a map of a river catchment, there is no need to leap from the individual rivulets to the River, with a capital R. We will follow, one by one, each individual rivulet until they become a river —with a small r.

What is so striking in the sociology of science is even more evident with regard to the law. This might explain in part why such an original social theory finds its origins in the writings of a man who was a judge. For a practicing judge the difference between the slow process of Common Law is not that different from Code based law. In both case, and this is a peculiarity of legal reasoning, the rule does not give you an easy access to the individual case.[xix] A "juge d'instruction" (a strange mixture between a prosecutor, a judge and a lawyer, typical to the French "inquisitorial" tradition) is well placed to see that any "general opinion" grows case by case to form a "whole" that is nevertheless never superior to the case-law and that a reversal of precedent can easily reverse (well not easily, that's the whole point). For a judge, the Code (or the case law) is never seen as more than a reference, a summary, a memory, a "composite photograph", a guide; it is not a structure from which one could deduce any individual motif or to which individual behavior should obey. The law sits side by side with a multiplicity of cases and precedents.

Son of a judge and a judge himself for most of his active life, Tarde could feel the gap between rules and individual behavior every day. It is tempting to find within that longstanding judiciary practice the root of his deep seated diffidence to any structural account.[xx] When Tarde heard the words "laws of society" in Spencer or even Durkheim, or "laws of nature" when reading natural scientists, he knew, first hand, that this was, at best a loose legal metaphor, and that it could never truly be the way that elements and aggregates would conspire together.[xxi]

Although deeply fascinated by Darwin, Tarde avoided the temptation of social Darwinism (quite a feat at the end of the 19th century) and for the same reason. Just as there is no "collective self" in human society, it can not be expected to appear in any in animal or plant society. He could not believe for one minute that sociology could be "reduced" to biology since in both cases societies are made of the same stuff. Hence Tarde's powerful appropriation of Darwin's discovery that no clarification on the genealogy of, for instance, individual horses, could ever come from an appeal to any Idea of a Horse. Among "astonished associates", evolutionary biologist will never see the emergence *ex abrupto* of this "marvelous result": a "collective Horse" born "collectively" from no mare and no stallion! Tarde might be considered the only French Darwinian, the only one who saw that the problem of composing organisms was the same in human and biological assemblages. No overall scheme in one, no overall scheme in the other. And especially, no "law of the jungle".

A judge, an avid reader of Leibniz (witness his most daring article *Monadologie et sociologie*) and of Darwin, could not but be struck by the case by case, organism by organism nature of any genealogy. For him, in whichever domain —science, law, biology— any belief in a structure is nothing but the pre-scientific, pre-Darwinian infancy of the social sciences. Structure is what is imagined to fill the gaps when there is a deficit of information as to the ways any entity inherits from its predecessors and successors.

Tarde would not have been greatly surprised to learn that when we apply the same ideal of science to societies of apes, ants or cells, here too, we begin to shift from a gross, statistically produced structure, to a trajectory of individual

innovations. When primatologists learned how to recognize individual baboons, vervets or chimpanzees, they too had to abandon rough and ready notions of a "collective self". They began to follow how each organism managed to engender a highly unstable aggregate that had to be constantly surveyed and reassembled through interactions (grooming, following, fighting, copulating, etc).[xxii] Tarde would have been even more thrilled when the discovery was made that the study of bacteria, marked so as to individualize them, produces different results from those obtained by studying them in bulk. What was lost in the idea of a law plus minor individual variation was the rather amazing differentiation between individual bacterial contributions to reproductive success.[xxiii] The scientist who was clever enough to succeed in inventing an instrument able to capture the contributions of each bacteria (the same has been done with ants), has produced a much more accurate picture of their aggregates.

Here again the opposition is not between a holistic view of the societies (bacteria, ants, monkeys or humans) and an individualist ones. It is between a first approximation through crude statistical records that loses most of the inner quantification of the organism, and a more refined one that has learned how to follow how *each* of those organisms inherits and transmits its own individual innovations. Change the instruments, and you will change the entire social theory that goes with them. The only thing to lose is the notion of a structure, distinct from its incarnations, this artifact that compensates for a deficit of information.

4) A monad, not an atom

The more we focus on the individual monad the more quantitative evaluation we will get. As long as we have not grasped this point, which seems at first so counterintuitive, the main difficulty of Tarde's idea of quantification will remain, despite radically improved instruments. This is especially true in economics, a science to which Tarde dedicated his last years[xxiv] in an attempt to render it more quantitative *and* more psychological: "The tendency to mathematize economic science and the tendency to psychologize it, far from being irreconcilable, should rather, in our view, lend each other mutual support."[xxv] He would add:

> *"No man, no people has ever failed to seek, as a prize for relentless efforts, a certain growth either of wealth, or glory, or truth, or power, or artistic perfection; nor has he failed to fight against the danger of a* decrease *of all of these assets. We all speak and write as though there existed a scale of these different orders of magnitude, on which we can place different peoples and different individuals higher or lower and make them rise or fall continuously. Everyone is thus implicitly and intimately convinced that all these things, and not only the first, are, in fact, real quantities. Not to recognize this truly quantitative – if not measurable de jure and de facto — aspect of power, of glory, of truth, of beauty, is thus to go against the constant of mankind and to set as the goal of universal effort a chimera."* (Psychologie économique Tome 1. p. 67*).*

Here resides the fourth and final reason why Tarde's sociology seems so original and so fresh for us today. A judgment of taste, an inflexion in the way we speak, a slight mutation in our habits, a preference between two goods, a decision

taken on the spur of the moment, an idea flashing in the brain, the conclusion of a long series of inconclusive syllogisms, etc —what appears most qualitative is actually where the greatest numbers of calculations are being made among "desires" and "beliefs". So, in principle, for Tarde, this is also the locus where we should be best able to quantify. Providing, that is, that we have the instruments to capture what he calls "logical duels".[xxvi]

The quantitative nature of all associations will seem bizarre if we mistakenly impute an idea of the individual element seen as an *atom* to Tarde. But the very idea of an individual as an atom is a consequence of the social theory he is fighting against. It is an outcome, as we just saw, of the statistical instruments that were available back then. In this traditional view, quantification starts when we have assembled enough individual atoms so that the outline of a structure begins to appear, first as a shadowy aggregate, then as a whole, and finally as a law dictating how to behave to the elements. The division between a qualitative and a quantitative social science is in essence *the same* as the division between individuals and society, tokens and type, actors and system. This is why no one has ever succeeded in "overcoming" the dichotomy between holistic and individualistic social theories.

But for Tarde, the whole scene is entirely different. The reason why there is no need for an overarching society is because there is no individual to begin with, or at least no individual atoms. [xxvii] The individual element is a monad, that is, a representation, a reflection, or an interiorisation of a whole set of other elements borrowed from the world around it. If there is nothing especially structural in the "whole", it is because of a vast crowd of elements *already present* in every single entity. This is where the word "network" —and even actor-network— captures what Tarde had to say much better than the word "individual". Contrary to what is often said, there is not even a hint of "methodological individualism" in this argument. There is no psychologism, nor of course any temptation toward "rational choice".

Hesitation is the great focus of Tarde's work. When any actor is found to be hesitating it is not because they are an atom taken in different fields of forces pressing on them from the outside. An actor hesitates as a monad which has already gathered within itself vast numbers of other elements to which it offers the stage for an indefinite number of logical duels to take place. In other words, if we are able to quantify an individual "one", it is because this instance is already "many". Behind every "he" and "she", one could say, there are a vast numbers of other "hes" and "shes" to which they have been interrelated.[xxviii] When Tarde insists that we detect specific embranchments and bifurcations behind every innovation, he is not saying that we should celebrate individual genius. It is rather that geniuses are made of a vast crowd of neurons!

> *"In a society no individual may act socially without the collaboration of a vast number of other individuals, most often ignored. The obscure workmen who, through the accumulation of small facts, have prepared the apparition of a grand scientific theory formulated by a Newton, a Cuvier, a Darwin, compose, if one may say so, the organism of which this genius is the soul; their obscure works are the cerebral vibrations of which this theory is the conscience. Conscience* means cerebral glory,

> *so to speak, of the most influential and most powerful element of the brain. Left to itself, a monad is powerless. This is the most important fact, and it leads immediately to explain another one:* the tendency of monads to aggregate. *(...) If ego is nothing but a directing monad among myriads of monads commensally aggregated under the same skull, what reason do we have to think that they are inferior? Is a monarch necessarily more intelligent that his ministers and subjects?"* (Monadologie et sociologie, p. 28)

A monarch is to his people, what conscience is to the brain, what ego is to the neurons, what Darwin is to the thousands of naturalists through the obscure work on which he depends for his "glory"! Once again, the "one" piggy backs on top of the "many" but without composing a "they". This is where Tarde's originality resides: everything is individual and yet there is no individual in the etymological sense of that which can not be further divided. This loss is a paradox, but only for those who would begin by opposing the structure and the elements.

Tarde derives his position from Leibniz' solution: there are monads all the way down, and God is in charge of regulating the connections between all of them without any of them acting directly on any other. For Tarde, of course, there is no God; therefore no pre-established harmony, no transcendence of any sort. (Tarde is probably the most systematic atheist there has ever been since he rejects even the transcendence of a "collective self" emerging *ex abrupto* from its associates.)[xxix] If there are monads but no God, the only solution is to let monads *penetrate* one another freely. Tarde's monads are a cross between Leibniz and Darwin: each monad has to get by in order to interpret or "reflect" (Leibniz's term) all of the others, to spread as far and as quickly as possible.

Tarde devises his notions of "desire", "belief" and "possession" very early on to code those relationships of interpenetration and competition from which all quantification resides in the end. The question "how many" is as essential to a monarch representing his people without any already existing political structure to hold them, as it is to Darwin's theory of evolution emerging out of the myriads of factoids assembled by his numerous collaborators toiling to collect samples in obscurity. How many entities can one entelechy reach? —That is *desire*. Mow many can they stabilize, order, fix or keep in place? —That is *belief*. No providence whatsoever can produce any harmony over and above the interplay of desire and belief in each monad, let loose on the world.[xxx]

This is precisely the reason why quantification is so important: not only does it capture internal logical duels, but it is the only way for monads to *coordinate* their actions externally with others in the absence of any providence. In a very strict sense *in Tarde's atheist monadology the practice of quantification plays the role of Leibniz' God*. With extreme avidity (a term Tarde prefers to that of 'identity'), all monads will seize every possible occasion to grasp one another in a quantitative manner. This accelerates and also simplifies their aggregation and cohesion; it modifies them and gives them another turn and another handle. It is in this sense that Tarde can be considered as the inventor of the notion that producing instruments and formalisms plays an active role in making the social visible to itself; and that such production offers many new handles so that the social can be performed anew.[xxxi] Examine what he says about how the advent of the press facilitates all judgments:

> *"[…] The development of the press had the effect of giving moral values a quantitative character that was more and more marked and better and better suited to justify their comparison with the exchange value. The latter, which must also have been quite confused in the centuries before the common use of currency, became better defined as currency spread and became more unified. It was then able to give rise, for the first time, to political economy. Similarly, before the advent of the daily press, the notions of the scientific or literary value of writing, of people's fame and reputation, were still vague, as the awareness of their gradual waxings and wanings could barely be felt; but with the development of the press, these ideas became clearer, were accentuated, became worthy of being the objects of philosophical speculations of a new sort." (***Psychologie économique***-*1, p. 76*).*

When Tarde says there is no "whole" transcendent to its instantiations, and when he says that any quantification deployed by various statistical or metrological instruments will have huge influence on the way all monads cohere and conspire, he is repeating the same argument twice. This is why his theory of science is so original: science is *in* and *of* the world it studies. It does not hang over the world from the outside. It has no privilege. This is precisely what makes science so immensely important: it performs the social together with all of the other actors, all of whom try to turn new instruments to their own benefits.

The continuity between the inner and the outer quantification is so complete that Tarde goes even further. He assimilates the quantitative apparatus of so many social sciences to the biological senses. He imagines a progressive fusion between the technologies of statistical instruments and the very physiology of perception. A day will come, he argues, when the standardization and development of statistics will be so complete that we will begin to follow the trajectory of some data about the social world in the same way as we follow the flight of a swallow with out eyes.[xxxii] Does this strike you as poetry? History is not yet finished, so we must wait and see. A century from now we may well read those predictions in a very different light: data gathering instrumentations will have changed again, and so will the social theories associated with them.

5) Digital traceability… Tarde's vindication?

The amazing chapter devoted to statistics in *The Laws of Imitation* is inescapably connected to the digital world to which we now have access.

> *"If Statistics continues to progress as it has done for several years, if the information which it gives us continues to gain in accuracy, in dispatch, in bulk, and in regularity, a time may come when upon the accomplishment of every social event a figure will at once issue forth automatically, so to speak, to take its place on the statistical registers that will be continuously communicated to the public and spread abroad pictorially by the daily press. Then, at every step, at every glance cast upon poster or newspaper, we shall be assailed, as it were, with statistical facts, with precise and condensed knowledge of all the peculiarities of actual social conditions, of commercial gains or losses, of the rise or falling off of certain political parties, of the progress or decay of a certain doctrine, etc., in exactly the same way as we are assailed when we open our eyes by the vibrations of the ether which tell us of the approach or withdrawal of such*

> *and such a so-called body and of many other things of a similar nature"* (The Laws of Imitation, p. 167-168).

Is this the prose of someone who despises quantitative science? If it true, as Tarde never tired of objecting to his younger colleague, Durkheim, that the theory of "society" was an artifact of rudimentary statistics, then the consequence for the present are obvious: what would happen to the respective programs of Tarde and Durkheim if social scientists began to have access, a century later, for reasons totally unexpected to both, to types of data that would allow them to follow, without any interruption, with the same tools, and in the same optically coherent space, those "imitative rays" that encompass individual innovations as well their aggregates? It is on this point that we discover why Tarde appears so fresh. The interest he triggers is not about a curious failure of social theory to become scientific, a quaint and queer qualitative view of the social. The most interesting part of Tarde is his lucid expectation of the type of information that should be gathered for a science of the social.

It is indeed striking that at this very moment, the fast expanding fields of "data visualization", "computational social science" or "biological networks"[xxxiii] are tracing, before our eyes, just the sort of data Tarde would have acclaimed. If the sociology of science, because of the traceability inherent in the scientific references, would have been the model for disentangling the "hes" and "shes" from the "they" for Tarde, then what we are witnessing, thanks to the digital medium, is a fabulous extension of this principle of traceability. It has been put in motion for not only to scientific statements, but also for opinions, rumors, political disputes, individual acts of buying and bidding, social affiliations, movements in space, telephone calls, and so on. What has previously been possible for only scientific activity —that we could have our cake (the aggregates) and eat it too (the individual contributors)— is now possible for most events leaving digital traces, archived in digital databanks, thanks, let's say, to Google and associates.

It is quite amusing to imagine Tarde directing his statistical bureau, nurturing so many doubts about the quality of the data he was handing out to the Ministry of Justice (and also to Marcel Mauss who was helping his uncle to write his book, *Suicide,* in which Tarde was trashed every two footnotes…), while dreaming, at the same time, of the many interesting quantitative instruments he had no way of obtaining: the "gloriometer" for following reputation (so easily accessible now with page rankings); conversation for understanding economic transactions (now the object of so many tools following buzz and viral marketing);[xxxiv] "phonometers" like those invented by Abbé Rousselot[xxxv] in order to follow the smallest inflexions of the native speakers (now accessible through the automated study of vast corpus of documents).

When Tarde claimed that statistics would one day be as easy to read as newspapers, he could not have anticipated that the newspapers themselves would be so transformed by digitalization that they would merge into the new domain of data visualization. This is a clear case of a social scientist being one century ahead of his time because he had anticipated a quality of connection and traceability necessary for good statistics which was totally unavailable in 1900. A century later, networks and traces are triggering the excitement of social and natural scientists

everywhere.[xxxvi] Here again, we note that the same scholars no longer make any distinction between the natural and the social domains to which they apply the same notion of networks: "Everything is a society", including ants, bacteria, cells, scientific paradigms or markets.

What Tarde could not have anticipated, however, are the added bonuses of the digital world that now provides an embodiment for his theory, at last: the notion of *navigation* where we are able to physically (well, virtually) navigate on our screens from the individual data points to the aggregates *and back*. In other words, the aggregate has lost the privilege it maintained for one century. Through the ease with which we can navigate a datascape, we manage to interrupt the transubstantiation of the aggregate into a law, a structure, a model and complicate the way through which one monad may come to summarize the "whole". But he "whole" is now nothing more than a provisional visualization which can be modified and reversed at will, by moving back to the individual components, and then looking for yet other tools to regroup the same elements into alternative assemblages.[xxxvii]

To be sure, the many tools we now have on our screens are still primitive (and many network based images are often no more readable than tea leaves at the bottom of a cup). But that's not the essential point. The point is that the whole has lost its privileged status: we can produce out of the same data points, as many aggregates as we see fit, while reverting back at any time, to the individual components.[xxxviii] This is precisely the sort of movement that was anticipated by Tarde's social theory although he had no tool to explicate his vision, other than his prose. While he was attempting to direct attention towards the "imitative ray" *in and of itself*, in order to displace the individual element *as well* as the structural whole, it has been altogether too easy for sociologists, starting with Durkheim, to corner him into dead end discussions about the micro versus the macro, the psychological versus the sociological, or the individualistic versus the holistic. In an unfair twist, it has been those who had only rudimentary tools, who have appeared more scientific than the one who was envisioning a much more refined and accurate type of data. Digital navigation through point-to-point datascapes might, a century later, vindicate Tarde's insights.

The overarching advantage of this type of quantification is worth underscoring: because "everything is a society" there is no clear divide between the biological and the social. For the first time in the history of science, the same data may look just as familiar to those who come from the "natural" sciences as to those who come from the "interpretative" ones. At the very least, reading Tarde might help social scientists to seize upon the opportunity provided by new digital media much faster than they might otherwise have done. The insights in his work can assist us in abandoning the impossible task of reconciling an old social theory, born out of discontinuous data, with the research terrain we now have readily available, at a click of a mouse.

---

* This paper has been written with the support of the European Program MACOSPOL (www.macospol.com). I thank Dominique Boullier, Emmanuel Didier, Louise Salmon & especially Isabelle Stengers for their useful remarks. I benefited once again from Martha Poon's editorial skills.

[i] Latour, Bruno (2002) "Gabriel Tarde and the End of the Social", In **The Social in Question. New Bearings in the History and the Social Sciences**, (Eds, Joyce, P.) London, Routledge, pp. 117-132 ; Toews, David (2003) "The New Tarde: Sociology after the End of the Social", *Theory, Culture and Society,* **20,** 81-98.

[ii] It is the very definition of the individual being that is in question for Tarde, see below.

[iii] Chapter one: "The Debate Tarde/Durkheim", pp.xx

[iv] See the excellent point made in Montebello, Pierre (2003) **L'autre métaphysique. Essai sur Ravaisson, Tarde, Nietzsche et Bergson**, Paris, Desclée de Brouwer, on those two difficult and central notions of Tarde, especially pp. 122-127.

[v] **Monadologie et Sociologie** translation Terry N. Clarke (1969) **Gabriel Tarde. On Communication and Social Influence. Selected Papers. Edited by Terry N. Clark**, Chicago, University of Chicago Press.

[vi] This was a great attraction at the turn of the century, especially when it was used to visualize the "criminal type" by superimposing images of criminals in the police archives! Gamboni, Dario (2005) « Composing the Body Politic. Composite Images and Political Representations 1651-2004 », In **Making Things Public. The Atmospheres of Democracy**, (Eds, Latour, B. and P. Weibel) Cambridge, Mass, MIT Press.

[vii] On the key concept of possession, see Debaise, Didier (2008) "Une métaphysique des possessions. Puissances et sociétés chez Gabriel Tarde", *Revue de Métaphysique et de Morale***,** 60, 8, 447-460.

[viii] : "Il en résulte que la contagion imitative de cette corporation antisociale [les brigands] ne reste pas tout entière renfermée dans son propre sein, où elle se traduit par le mutuel endurcissement, mais qu'elle rayonne en partie au dehors parmi les déclassés qu'elle classe, parmi les oisifs qu'elle occupe, parmi les décavés de tout genre qu'elle enfièvre des perspectives d'un nouveau jeu, le plus riche en émotions. Voilà la vraie source du mal." (**Criminalité comparée**, p. 52) cité in Didier Emmanuel "Tarde et le mouvement statistique" (document de travail, CESDIP, juin 2007).

[ix] For Tarde, the production of data by the administrations and the institutions is always foregrounded which makes him, once again, an important precursor of science studies. For him, the sciences –natural, social or cameral— are added to the world they study. This is especially true in the case of criminology, Tarde, Gabriel (2004) **La Criminalité comparée**, Paris, Les Empêcheurs. In the case of criminal records, he had a first hand knowledge of the ways they work (see below).

[x] In **Laws of imitation** Tarde claims that the best way to detect those imitative rays is in archeology since only there -when the living beings have disappeared and you are left with long series of artifacts- do you see in the purest and most abstract light what has been imitated by the long disappeared humans.

[xi] Tarde, Gabriel (1962) **The laws of imitation**. Translated from the 2d French ed. by Elsie Clews Parsons. With an introd. by Franklin H. Giddings**.**, Gloucester, Mass, P. Smith.

[xii] This is the critique made by Sperber, Dan (1996) **La contagion des idées**, Paris, Editions Odile Jacob. No doubt that Tarde would have been fascinated nonetheless by the discovery of mirror neurons, Rizzolatti, Giacomo, Sinigaglia Corrado and Raiola Marilène (2008) **Les neurones miroirs**, Paris, Odile Jacob..

[xiii] Tarde does for social theory what Pasteur had done in epidemiology: in the same way as bacteriology allows one to move from a regional theory of miasmas to a point to point and person to person theory of contagion through a specific vector (cholera bacillus, Koch' bacillus, etc), Tarde moves from an aggregated cloud of collective qualities to a highly specific point to point, person to person "contagion" of ideas each of them having its own peculiar effectivity.

[xiv] This is what allowed me to consider Tarde as the real inventor of ANT Latour, Bruno (2005) **Reassembling the Social. An Introduction to Actor-Network Theory**, Oxford, Oxford University Press.

[xv] What makes a society in Tarde has been the special concern of Debaise, op. cit.

[xvi] I am following here Didier, Emmanuel "Tarde et le mouvement statistique" (document de travail, CESDIP, juin 2007) and (2009) **En quoi consiste l'Amérique? Les statistiques, le New Deal et la démocratie**, Paris, La Découverte.

[xvii] For a broad view of the many different ways social sciences have developed to grasp the collective, see Desrosières, Alain (2002) **The Politics of Large Numbers: A History of Statistical Reasoning (translated by Camille Naish)**, Cambridge, Mass, Cambridge University Press.

[xviii] A "Mémoire sur l'organisation de la statistique criminelle en France", 1893. Most of his work is now available in Tarde, Gabriel (2004) **La Criminalité comparée**, Paris, Les Empêcheurs.

[xix] Latour, Bruno (2009) **Law in the Making. An Ethnography of the Conseil d'Etat** (translated by Marina Brilman and Alain Pottage), London, Polity Press..

[xx] See the same argument in Milet, Jean, "Introduction", in Tarde Gabriel, *Les Transformations du droit*, Paris, Berg International, 1994, pp. 7-9. I thank Louise Salmon for this reference. Her thesis on the history of Tarde's milieu will contain many important material on this link between the practice of law and Tarde's social theory.

[xxi] He even extended this diffidence to the laws of nature : "materialists have to invoke, as complement of their erratic and blinds atoms, universal laws or the unique formula to which all those laws could be reduced, a sort of mystical commandment to whom all beings would obey and which would emanate from no being whatsoever, sort of ineffable and unintelligible verb which, without having ever been uttered by anyone, would nonetheless be listened to always and everywhere." p. 56 **Monadologie et sociologie.**

[xxii] Strum, Shirley and Fedigan Linda (Eds.), (2000) **Primate Encounters**, Chicago, University of Chicago Press ; Cheney, Dorothy L. and Seyfarth Robert M. (1990) **How Monkeys See the World. Inside the Mind of Another Species**, Chicago, University of Chicago Press.

[xxiii] Stewart, Eric J., Madden Richard, Paul Gregory and Taddei François (2004) "Aging and Death in an Organism That Reproduces by Morphologically Symmetric Division", *PLoS Biol,* **3,** doi:10.1371/journal.pbio.0030045.

[xxiv] But on which he had already contributed in one of its earliest articles "La psychologie ou économie politique", ***Revue Philosophique***, tome xii, 1881, pp. 232-250; 401-418.

[xxv] **Psychologie économique** is published in 1904 ; see Latour, Bruno and Lépinay Vincent (2008) **L'économie science des intérêts passionnés - introduction à l'anthropologie économique de Gabriel Tarde**, Paris, La Découverte, (English translation: (2009) **The Science of Passionate Interests. An Introduction to Gabriel Tarde's Economic Anthropology**, Prickly Paradigm Press, Chicago (2009), most of the important passages are accessible on the following site : http://www.bruno-latour.fr/). See also the special issue on Tarde's economics: ***Economy and Society***, Volume36, Number4, November2007.

[xxvi] See Tarde, Gabriel (1999) **La logique sociale**, Paris, Les Empêcheurs de penser en rond which is entirely devoted to an alternative quantitative and yet non formalist socio-logic.

[xxvii] The same argument is made by the pragmatists, see Dewey, John (1927 1954) **The Public and Its Problems**, Athens, Ohio University Press, especially the second chapter which deduces the very notion of an "individual" from a faulty definition of the State. It is interesting to note that the domination of the notion of structure on social thought is so strong that Tarde, as well as the pragmatists, have been constantly misunderstood.

[xxviii] Hence Tarde's interest in the phenomenon that economists of innovation and historians of technology call "lock in", "standardization" or "entrenchment".

[xxix] Witness the radical critique of providentialism Tarde pursues throughout the whole of **Psychologie économique**. This critique allows him to criticize the notion of a social animal as well as that of the laissez-faire free marketers… Latour and Lépinay op. cit. p.xx.

[xxx] Tarde's first paper on the question from 1880 has a very revealing title: "La croyance et le désir, la possibilité de leur mesure", ***Revue philosophique***, tome X, pp. 150-180, 264-283, republished in **Essais et mélanges sociologiques**, Maloine 1895. "No intellectual effort will make it possible to conceive of an animal, or a monocellular organism, which, being sensitive, would not also be endowed with belief and desire, that is, will not associate and dissociate, collect and reject its impressions, its sensations whatever they are, with more or less intensity. M. Delboeuf explains very well that even an infusoria is able to utter this mute judgment: I am hot". **Essais et Mélanges,** p. 185 candad

[xxxi] Even though the word "performative" is hotly debated (see Emmanuel Didier "Do Statistics "perform" the Economy?" in MacKenzie, Donald, Muniesa Fabian and Siu Lucia (Ed.), (2007) **Do Economists Make Markets? On the Performativity of Economics**, Princeton, Princeton University Press) it is still the best concept to define science studies' interpretation of the reflexive nature of formalisms.

[xxxii] Pages 75-132 in Tarde, Gabriel (1903) **The Laws of Imitation** (translated by Else Clews Parsons with an introduction by Franlin H. Giddings), New York, Henry Holt and Company.

[xxxiii] Lazer, David and alii (2009) "Computational Social Science", *Science,* **323,** 721-723; Wimsatt, William C. (2007) **Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality**, Cambridge, Mass, Harvard University Press.

[xxxiv] Emanuel Rosen (2009) **The Anatomy of Buzz Revisited: Real-life lessons in Word-of-Mouth Marketing**, Broadway Business.

[xxxv] See Andy Barry's chapter in this volume.

[xxxvi] Barabasi, Albert-Laszlo (2003) **Linked: How Everything Is Connected to Everything Else and What It Means**, Plume, New York ; Benkler, Yochai (2006) **The Wealth of Networks. How Social Production Transforms Market and Freedom**, New Haven, Yale University Press.

[xxxvii] For striking examples of such a navigation, see http://www.demoscience.org/ assembled by the European project MACOSPOL.

[xxxviii] Mogoutov, Andrei, Cambrosio Alberto and Mustar Philippe (2008) "Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays," *Journal of Informetrics,* **2,** 341-353.

Ontology is Overrated -- Categories, Links, and Tags

Ontology is Overrated: Categories, Links, and Tags

This piece is based on two talks I gave in the spring of 2005 -- one at the O'Reilly ETech conference in March, entitled "Ontology Is Overrated", and one at the IMCExpo in April entitled "Folksonomies & Tags: The rise of user-developed classification." The written version is a heavily edited concatenation of those two talks.

Today I want to talk about categorization, and I want to convince you that a lot of what we think we know about categorization is wrong. In particular, I want to convince you that many of the ways we're attempting to apply categorization to the electronic world are actually a bad fit, because we've adopted habits of mind that are left over from earlier strategies.

I also want to convince you that what we're seeing when we see the Web is actually a radical break with previous categorization strategies, rather than an extension of them. The second part of the talk is more speculative, because it is often the case that old systems get broken before people know what's going to take their place. (Anyone watching the music industry can see this at work today.) That's what I think is happening with categorization.

What I think is coming instead are much more organic ways of organizing information than our current categorization schemes allow, based on two units -- the link, which can point to anything, and the tag, which is a way of attaching labels to links. The strategy of tagging -- free-form labeling, without regard to categorical constraints -- seems like a recipe for disaster, but as the Web has shown us, you can extract a surprising amount of value from big messy data sets.

PART I: Classification and Its Discontents #

Q: What is Ontology? A: It Depends on What the Meaning of "Is" Is.

I need to provide some quick definitions, starting with ontology. It is a rich irony that the word "ontology", which has to do with making clear and explicit statements about entities in a particular domain, has so many conflicting definitions. I'll offer two general ones.

The main thread of ontology in the philosophical sense is the study of entities and their relations. The question ontology asks is: What kinds of things exist or can exist in the world, and what manner of relations can those things have to each other? Ontology is less concerned with what is than with what is possible.

The knowledge management and AI communities have a related definition -- they've taken the word "ontology" and applied it more directly to their problem. The sense of ontology there is something like "an explicit specification of a conceptualization."

The common thread between the two definitions is essence, "Is-ness." In a particular domain, what kinds of things can we say exist in that domain, and how can we say those things relate to each other?

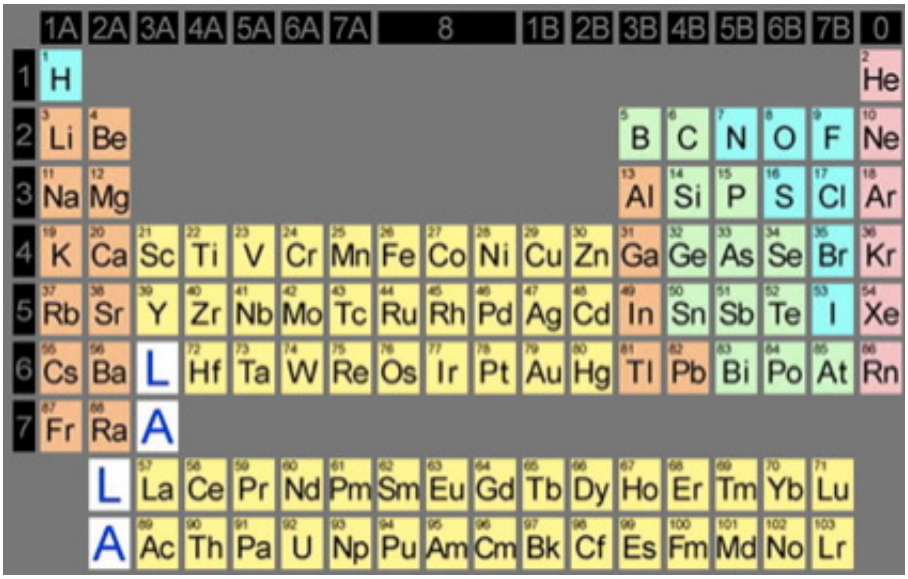
The other pair of terms I need to define are categorization and classification. These are the act of organizing a collection of entities, whether things or concepts, into related groups. Though there are some field-by-field distinctions, the terms are in the main used interchangeably.

And then there's ontological classification or categorization, which is organizing a set of entities into groups, based on their essences and possible relations. A library catalog, for example, assumes that for any new book, its logical place already exists within the system, even before the book was published. That strategy of designing categories to cover possible cases in advance is what I'm primarily concerned with, because it is both widely used and badly overrated in terms of its value in the digital world.

Now, anyone who deals with categorization for a living will tell you they can never get a perfect system. In working classification systems, success is not "Did we get the ideal arrangement?" but rather "How close did we come, and on what measures?" The idea of a perfect scheme is simply a Platonic ideal. However, I want to argue that even the ontological *ideal* is a mistake. Even using theoretical perfection as a measure of practical success leads to misapplication of resources.

Now, to the problems of classification.

Cleaving Nature at the Joints #



[The Periodic Table of the Elements]

The periodic table of the elements is my vote for "Best. Classification. Evar." It turns out that by organizing elements by the number of protons in the nucleus, you get all of this fantastic value, both descriptive and predictive value. And because what you're doing is organizing *things*, the periodic table is as close to making assertions about essence as it is physically possible to get. This is a really powerful scheme, almost perfect. Almost.

All the way over in the right-hand column, the pink column, are noble gases. Now noble gas is an odd category, because helium is no more a gas than mercury is a liquid. Helium is not fundamentally a gas, it's just a gas at most temperatures, but the people studying it at the time didn't know that, because they weren't able to make it cold enough to see that helium, like everything else, has different states of matter. Lacking the right measurements, they assumed that gaseousness was an essential aspect -- literally, part of the essence -- of those elements.

Even in a nearly perfect categorization scheme, there are these kinds of context errors, where people are placing something that is merely true at room temperature, and is absolutely unrelated to essence, right in the center of the categorization. And the category 'Noble Gas' has stayed there from the day they added it, because we've all just gotten used to that anomaly as a frozen accident.

If it's impossible to create a completely coherent categorization, even when you're doing something as physically related to essence as chemistry, imagine the problems faced by anyone who's dealing with a domain where essence is even less obvious.

Which brings me to the subject of libraries.

Of Cards and Catalogs

The periodic table gets my vote for the best categorization scheme ever, but libraries have the best-known categorization schemes. The experience of the library catalog is probably what people know best as a high-order categorized view of the world, and those cataloging systems contain all kinds of odd mappings between the categories and the world they describe.

Here's the first top-level category in the Soviet library system:

> **A: Marxism–Leninism**
> A1: Classic works of Marxism–Leninism
> A3: Life and work of C.Marx, F.Engels, V.I.Lenin
> A5: Marxism–Leninism Philosophy
> A6: Marxist–Leninist Political Economics
> A7/8: Scientific Communism

Some of those categories are starting to look a little bit dated.

Or, my favorite -- this is the Dewey Decimal System's categorization for religions of the world, which is the 200 category.

> **Dewey, 200: Religion**
> 210 Natural theology
> 220 Bible
> 230 Christian theology
> 240 Christian moral & devotional theology
> 250 Christian orders & local church
> 260 Christian social theology
> 270 Christian church history
> 280 Christian sects & denominations
> 290 Other religions

How much is this not the categorization you want in the 21st century?

This kind of bias is rife in categorization systems. Here's the Library of Congress' categorization of History. These are all the top-level categories -- all of these things are presented as being co-equal.

> **D: History (general)**
>
> DA: Great Britain
> DB: Austria
> DC: France
> DD: Germany
> DE: Mediterranean
> DF: Greece
> DG: Italy
> DH: Low Countries
> DJ: Netherlands
> DK: Former Soviet Union
> DL: Scandinavia
> DP: Iberian Peninsula
> DQ: Switzerland
> **DR: Balkan Peninsula**
> **DS: Asia**
> **DT: Africa**
> DU: Oceania
> DX: Gypsies

I'd like to call your attention to the ones in bold: The Balkan Peninsula. Asia. Africa.

And just, you know, to review the geography:

[Spot the difference?]

Yet, for all the oddity of placing the Balkan Peninsula and Asia in the same level, this is harder to laugh off than the Dewey example, because it's so puzzling. The Library of Congress -- no slouches in the thinking department, founded by Thomas Jefferson -- has a staff of people who do nothing but think about categorization all day long. So what's being optimized here? It's not geography. It's not population. It's not regional GDP.

What's being optimized is number of books on the shelf. That's what the categorization scheme is categorizing. It's tempting to think that the classification schemes that libraries have optimized for in the past can be extended in an uncomplicated way into the digital world. This badly underestimates, in my view, the degree to which what libraries have historically been managing is an entirely different problem.

The musculature of the Library of Congress categorization scheme looks like it's about concepts. It is organized into non-overlapping categories that get more detailed at lower and lower levels -- any concept is supposed to fit in one category and in no other categories. But every now and again, the skeleton pokes through, and the skeleton, the supporting structure around which the system is really built, is designed to minimize seek time on shelves.

The essence of a book isn't the ideas it contains. The essence of a book is "book." Thinking that library catalogs exist to organize concepts confuses the container for the thing contained.

The categorization scheme is a response to physical constraints on storage, and to people's inability to keep the location of more than a few hundred things in their mind at once. Once you own more than a few hundred books, you have to organize them somehow. (My mother, who was a reference librarian, said she wanted to reshelve the entire University library by color, because students would come in and say "I'm looking for a sociology book. It's green...") But however you do it, the frailty of human memory and the physical fact of books make some sort of organizational scheme a requirement, and hierarchy is a good way to manage physical objects.

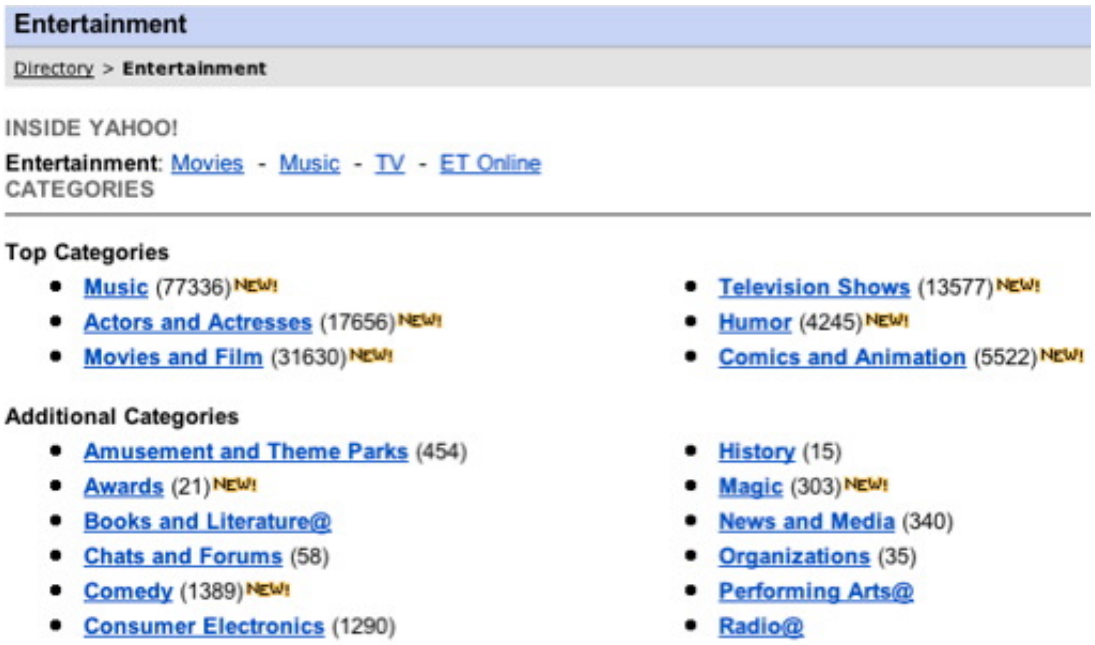
The "Balkans/Asia" kind of imbalance is simply a byproduct of physical constraints. It isn't the ideas in a book that have to be in one place -- a book can be about several things at once. It is the book itself, the physical fact of the bound object, that has to be one place, and if it's one place, it can't also be in another place. And this in turn means that a book has to be declared to be *about* some main thing. A book which is equally about two things breaks the 'be in one place' requirement, so each book needs to be declared to about one thing more than others, regardless of its actual contents.

People have been freaking out about the virtuality of data for decades, and you'd think we'd have internalized the obvious truth: there is no shelf. In the digital world, there is no physical constraint that's forcing this kind of organization on us any longer. We can do without it, and you'd think we'd have learned that lesson by now.

And yet.

A little over ten years ago, a couple of guys out of Stanford launched a service called Yahoo that offered a list of things available on the Web. It was the first really significant attempt to bring order to the Web. As the Web expanded, the Yahoo list grew into a hierarchy with categories. As the Web expanded more they realized that, to maintain the value in the directory, they were going to have to systematize, so they hired a professional ontologist, and they developed their now-familiar top-level categories, which go to subcategories, each subcategory contains links to still other subcategories, and so on. Now we have this ontologically managed list of what's out there.

Here we are in one of Yahoo's top-level categories, Entertainment.



[Yahoo's Entertainment Category]

You can see what the sub-categories of Entertainment are, whether or not there are new additions, and how many links roll up under those sub-categories. Except, in the case of Books and Literature, that sub-category doesn't tell you how many links roll up under it. Books and Literature doesn't end with a number of links, but with an "@" sign. That "@" sign is telling you that the category of Books and Literature isn't 'really' in the category Entertainment. Yahoo is saying "We've put this link here for your convenience, but that's only to take you to where Books and Literature 'really' are." To which one can only respond -- "What's real?"

Yahoo is saying "We understand better than you how the world is organized, because we are trained professionals. So if you mistakenly think that Books and Literature are entertainment, we'll put a little flag up so we can set you right, but to see those links, you have to 'go' to where they 'are'." (My fingers are going to fall off from all the air quotes.) When you go to Literature -- which is part of Humanities, not Entertainment -- you are told, similarly, that booksellers are not 'really' there. Because they are a commercial service, booksellers are 'really' in Business.

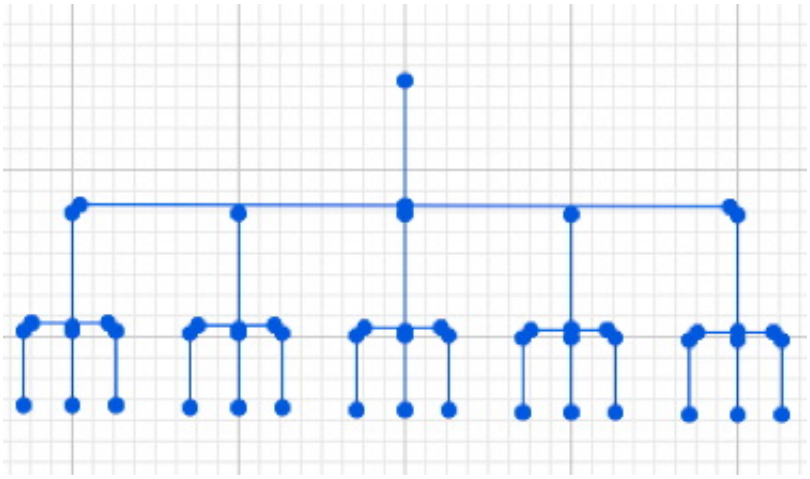
['Literature' on Yahoo]

Look what's happened here. Yahoo, faced with the possibility that they could organize things with no physical constraints, *added the shelf back*. They couldn't imagine organization without the constraints of the shelf, so they added it back. It is perfectly possible for any number of links to be in any number of places in a hierarchy, or in many hierarchies, or in no hierarchy at all. But Yahoo decided to privilege one way of organizing links over all others, because they wanted to make assertions about what is "real."

The charitable explanation for this is that they thought of this kind of a priori organization as their job, and as something their users would value. The uncharitable explanation is that they thought there was business value in determining the view the user would have to adopt to use the system. Both of those explanations may have been true at different times and in different measures, but the effect was to override the users' sense of where things ought to be, and to insist on the Yahoo view instead.

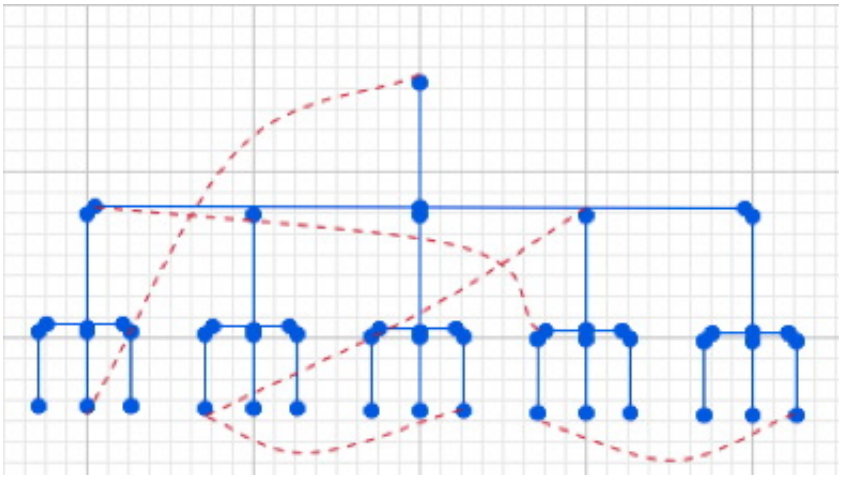
File Systems and Hierarchy #

It's easy to see how the Yahoo hierarchy maps to technological constraints as well as physical ones. The constraints in the Yahoo directory describes both a library categorization scheme and, obviously, a file system -- the file system is both a powerful tool and a powerful metaphor, and we're all so used to it, it seems natural.



[Hierarchy]

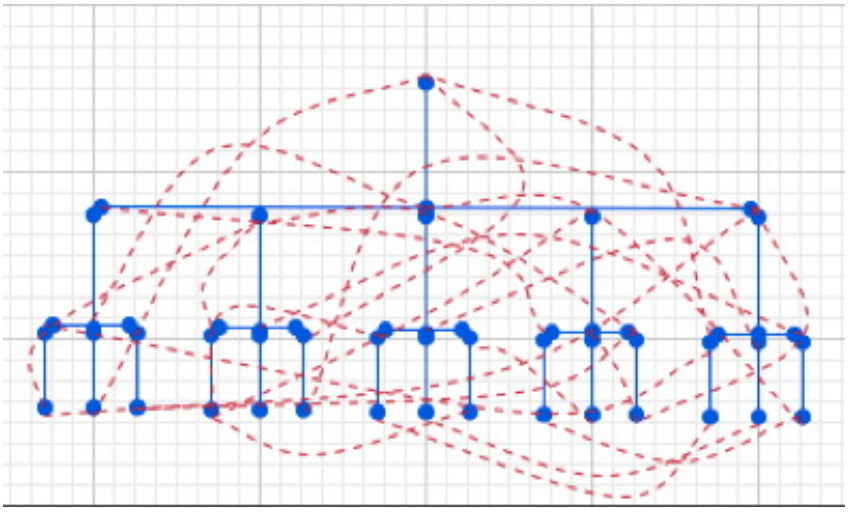
There's a top level, and subdirectories roll up under that. Subdirectories contain files or further subdirectories and so on, all the way down. Both librarians and computer scientists hit the same next idea, which is "You know, it wouldn't hurt to add a few secondary links in here" -- symbolic links, aliases, shortcuts, whatever you want to call them.



[Plus Links]

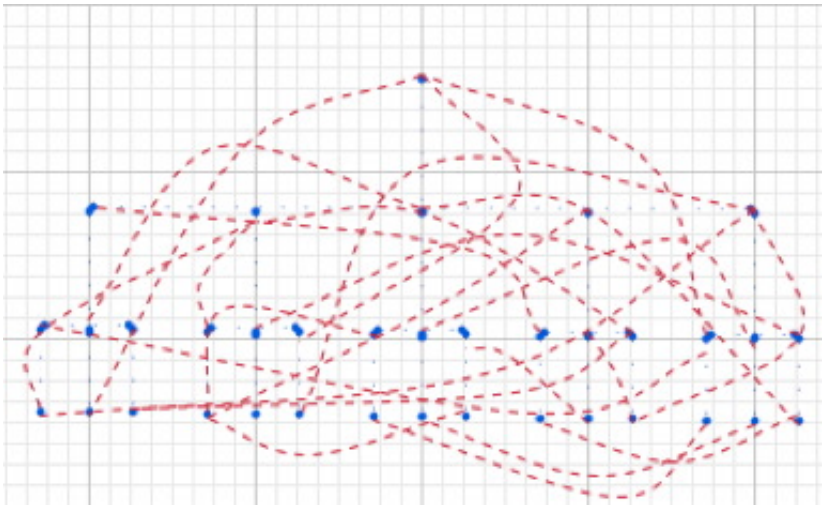
The Library of Congress has something similar in its second-order categorization -- "This book is mainly about the Balkans, but it's also about art, or it's mainly about art, but it's also about the Balkans." Most hierarchical attempts to subdivide the world use some system like this.

Then, in the early 90s, one of the things that Berners-Lee showed us is that you could have a lot of links. You don't have to have just a few links, you could have a whole lot of links.



[Plus Lots of Links]

This is where Yahoo got off the boat. They said, "Get out of here with that crazy talk. A URL can only appear in three places. That's the Yahoo rule." They did that in part because they didn't want to get spammed, since they were doing a commercial directory, so they put an upper limit on the number of symbolic links that could go into their view of the world. They missed the end of this progression, which is that, if you've got enough links, you don't need the hierarchy anymore. There is no shelf. There is no file system. The links alone are enough.



[Just Links (There Is No Filesystem)]

One reason Google was adopted so quickly when it came along is that Google understood there is no shelf, and that there is no file system. Google can decide what goes with what *after* hearing from the user, rather than trying to predict in advance what it is you need to know.

Let's say I need every Web page with the word "obstreperous" and "Minnesota" in it. You can't ask a cataloguer in advance to say "Well, that's going to be a useful category, we should encode that in advance." Instead, what the cataloguer is going to say is, "Obstreperous plus Minnesota! Forget it, we're not going to optimize for one-offs like that." Google, on the other hand, says, "Who cares? We're not going to tell the user what to do, because the link structure is more complex than we can read, except in response to a user query."

Browse versus search is a radical increase in the trust we put in link infrastructure, and in the degree

of power derived from that link structure. Browse says the people making the ontology, the people doing the categorization, have the responsibility to organize the world in advance. Given this requirement, the views of the catalogers necessarily override the user's needs and the user's view of the world. If you want something that hasn't been categorized in the way you think about it, you're out of luck.

The search paradigm says the reverse. It says nobody gets to tell you in advance what it is you need. Search says that, at the moment that you are looking for it, we will do our best to service it based on this link structure, because we believe we can build a world where we don't need the hierarchy to coexist with the link structure.

A lot of the conversation that's going on now about categorization starts at a second step -- "Since categorization is a good way to organize the world, we should..." But the first step is to ask the critical question: Is categorization a good idea? We can see, from the Yahoo versus Google example, that there are a number of cases where you get significant value out of *not* categorizing. Even Google adopted DMOZ, the open source version of the Yahoo directory, and later they downgraded its presence on the site, because almost no one was using it. When people were offered search and categorization side-by-side, fewer and fewer people were using categorization to find things.

When Does Ontological Classification Work Well?

Ontological classification works well in some places, of course. You need a card catalog if you are managing a physical library. You need a hierarchy to manage a file system. So what you want to know, when thinking about how to organize anything, is whether that kind of classification is a good strategy.

Here is a partial list of characteristics that help make it work:

Domain to be Organized

- Small corpus
- Formal categories
- Stable entities
- Restricted entities
- Clear edges

This is all the domain-specific stuff that you would like to be true if you're trying to classify cleanly. The periodic table of the elements has all of these things -- there are only a hundred or so elements; the categories are simple and derivable; protons don't change because of political circumstances; only elements can be classified, not molecules; there are no blended elements; and so on. The more of those characteristics that are true, the better a fit ontology is likely to be.

The other key question, besides the characteristics of the domain itself, is "What are the participants like?" Here are some things that, if true, help make ontology a workable classification strategy:

Participants

- Expert catalogers
- Authoritative source of judgment
- Coordinated users
- Expert users

DSM-IV, the 4th version of the psychiatrists' Diagnostic and Statistical Manual, is a classic example of an classification scheme that works because of these characteristics. DSM IV allows psychiatrists all over the US, in theory, to make the same judgment about a mental illness, when presented with the same list of symptoms. There is an authoritative source for DSM-IV, the American Psychiatric Association. The APA gets to say what symptoms add up to psychosis. They have both expert cataloguers and expert users. The amount of 'people infrastructure' that's hidden in a working system like DSM IV is a big part of what makes this sort of categorization work.

This 'people infrastructure' is very expensive, though. One of the problem users have with categories is that when we do head-to-head tests -- we describe something and then we ask users to guess how we described it -- there's a very poor match. Users have a terrifically hard time guessing how something they want will have been categorized in advance, unless they have been educated about those categories in advance as well, and the bigger the user base, the more work that user education is.

You can also turn that list around. You can say "Here are some characteristics where ontological classification doesn't work well":

Domain

- Large corpus
- No formal categories
- Unstable entities
- Unrestricted entities
- No clear edges

Participants

- Uncoordinated users
- Amateur users
- Naive catalogers
- No Authority

If you've got a large, ill-defined corpus, if you've got naive users, if your cataloguers aren't expert, if there's no one to say authoritatively what's going on, then ontology is going to be a bad strategy.

The list of factors making ontology a bad fit is, also, an almost perfect description of the Web -- largest corpus, most naive users, no global authority, and so on. The more you push in the direction of scale, spread, fluidity, flexibility, the harder it becomes to handle the expense of starting a cataloguing system and the hassle of maintaining it, to say nothing of the amount of force you have to get to exert over users to get them to drop their own world view in favor of yours.

The reason we know SUVs are a light truck instead of a car is that the Government says they're a light truck. This is voodoo categorization, where acting on the model changes the world -- when the Government says an SUV is a truck, it *is* a truck, by definition. Much of the appeal of categorization comes from this sort of voodoo, where the people doing the categorizing believe, even if only unconciously, that naming the world changes it. Unfortunately, most of the world is not actually amenable to voodoo categorization.

The reason we don't know whether or not *Buffy, The Vampire Slayer* is science fiction, for example, is because there's no one who can say definitively yes or no. In environments where there's no authority and no force that can be applied to the user, it's very difficult to support the voodoo style of organization. Merely naming the world creates no actual change, either in the world, or in the minds of potential users who don't understand the system.

Mind Reading #

One of the biggest problems with categorizing things in advance is that it forces the categorizers to take on two jobs that have historically been quite hard: mind reading, and fortune telling. It forces categorizers to guess what their users are thinking, and to make predictions about the future.

The mind-reading aspect shows up in conversations about controlled vocabularies. Whenever users are allowed to label or tag things, someone always says "Hey, I know! Let's make a thesaurus, so that if you tag something 'Mac' and I tag it 'Apple' and somebody else tags it 'OSX', we all end up looking at the same thing!" They point to the signal loss from the fact that users, although they use these three different labels, are talking about the same thing.

The assumption is that we both can and should read people's minds, that we can understand what they meant when they used a particular label, and, understanding that, we can start to restrict those labels, or at least map them easily onto one another.

This looks relatively simple with the Apple/Mac/OSX example, but when we start to expand to other groups of related words, like movies, film, and cinema, the case for the thesaurus becomes much less clear. I learned this from Brad Fitzpatrick's design for LiveJournal, which allows user to list their own interests. LiveJournal makes absolutely no attempt to enforce solidarity or a thesaurus or a minimal set of terms, no check-box, no drop-box, just free-text typing. Some people say they're interested in movies. Some people say they're interested in film. Some people say they're interested in cinema.

The cataloguers first reaction to that is, "Oh my god, that means you won't be introducing the movies people to the cinema people!" To which the obvious answer is "Good. The movie people don't *want* to hang out with the cinema people." Those terms actually encode different things, and the assertion that restricting vocabularies improves signal assumes that that there's no signal in the difference itself, and no value in protecting the user from too many matches.

When we get to really contested terms like queer/gay/homosexual, by this point, all the signal loss is in the collapse, not in the expansion. "Oh, the people talking about 'queer politics' and the people talking about 'the homosexual agenda', they're really talking about the same thing." Oh no they're

not. If you think the movies and cinema people were going to have a fight, wait til you get the queer politics and homosexual agenda people in the same room.

You can't do it. You can't collapse these categorizations without some signal loss. The problem is, because the cataloguers assume their classification should have force on the world, they underestimate the difficulty of understanding what users are thinking, and they overestimate the amount to which users will agree, either with one another or with the catalogers, about the best way to categorize. They also underestimate the loss from erasing difference of expression, and they overestimate loss from the lack of a thesaurus.

Fortune Telling

The other big problem is that predicting the future turns out to be hard, and yet any classification system meant to be stable over time puts the categorizer in the position of fortune teller.

Alert readers will be able to spot the difference between Sentence A and Sentence B.

> ```
> A: "I love you."
> B: "I will always love you."
> ```

Woe betide the person who utters Sentence B when what they mean is Sentence A. Sentence A is a statement. Sentence B is a prediction.

But this is the ontological dilemma. Consider the following statements:

> ```
> A: "This is a book about Dresden."
> B: "This is a book about Dresden,
>  and it goes in the category 'East Germany'."
> ```

That second sentence seems so obvious, but East Germany actually turned out to be an unstable category. Cities are real. They are real, physical facts. Countries are social fictions. It is much easier for a country to disappear than for a city to disappear, so when you're saying that the small thing is contained by the large thing, you're actually mixing radically different kinds of entities. We pretend that 'country' refers to a physical area the same way 'city' does, but it's not true, as we know from places like the former Yugoslavia.

There is a top-level category, you may have seen it earlier in the Library of Congress scheme, called Former Soviet Union. The best they were able to do was just tack "former" onto that entire zone that they'd previously categorized as the Soviet Union. Not because that's what they thought was true about the world, but because they don't have the staff to reshelve all the books. That's the constraint.

Part II: The Only Group That Can Categorize Everything Is Everybody

"My God. It's full of links!"

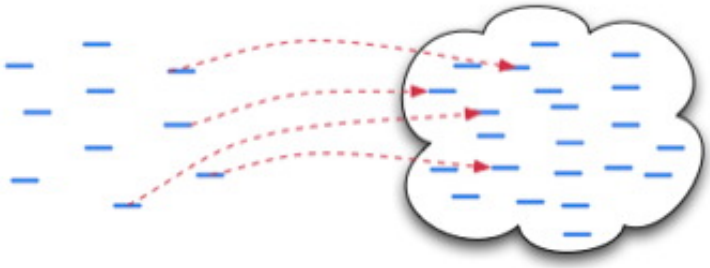
When we reexamine categorization without assuming the physical constraint either of hierarchy or

When we reexamine categorization without assuming the physical constraint either of hierarchy on disk or of hierarchy in the physical world, we get very different answers. Let's say you wanted to merge two libraries -- mine and the Library of Congress's. (You can tell it's the Library of Congress on the right, because they have a few more books than I do.)

[Two Categorized Collections of Books]

So, how do we do this? Do I have to sit down with the Librarian of Congress and say, "Well, in my world, *Python In A Nutshell* is a reference work, and I keep all of my books on creativity together." Do we have to hash out the difference between my categorization scheme and theirs before the Library of Congress is able to take my books?

No, of course we don't have to do anything of the sort. They're able to take my books in while ignoring my categories, because all my books have ISBN numbers, International Standard Book Numbers. They're not merging at the category level. They're merging at the globally unique item level. My entities, my uniquely labeled books, go into Library of Congress scheme trivially. The presence of unique labels means that merging libraries doesn't require merging categorization schemes.



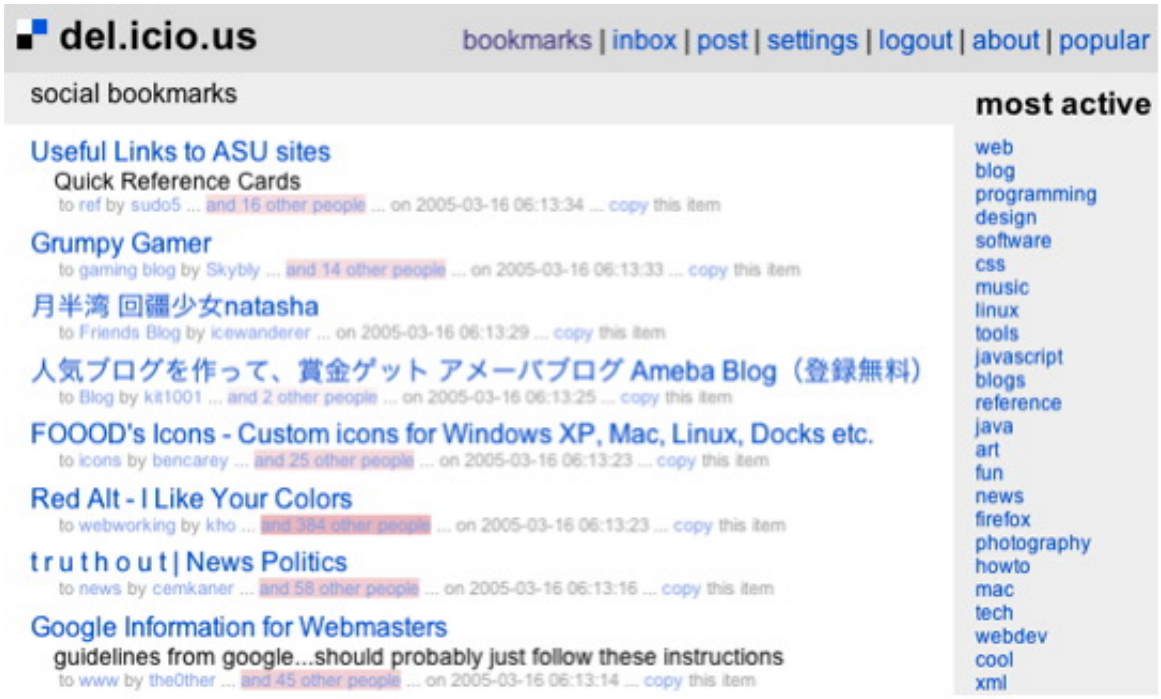
[Merge ISBNs]

Now imagine a world where *everything* can have a unique identifier. This should be easy, since that's the world we currently live in -- the URL gives us a way to create a globally unique ID for anything we need to point to. Sometimes the pointers are direct, as when a URL points to the contents of a Web page. Sometimes they are indirect, as when you use an Amazon link to point to a book. Sometimes there are layers of indirection, as when you use a URI, a uniform resource identifier, to name something whose location is indeterminate. But the basic scheme gives us ways to create a globally unique identifier for anything.

And once you can do that, anyone can label those pointers, can tag those URLs, in ways that make them more valuable, and all without requiring top-down organization schemes. And this -- an explosion in free-form labeling of links, followed by all sorts of ways of grabbing value from those labels -- is what I think is happening now.

Great Minds Don't Think Alike

Here is del.icio.us, Joshua Shachter's social bookmarking service. It's for people who are keeping track of their URLs for themselves, but who are willing to share globally a view of what they're doing, creating an aggregate view of all users' bookmarks, as well as a personal view for each user.



[Front Page of del.icio.us]

As you can see here, the characteristics of a del.icio.us entry are a link, an optional extended description, and a set of tags, which are words or phrases users attach to a link. Each user who adds a link to the system can give it a set of tags -- some do, some don't. Attached to each link on the home page are the tags, the username of the person who added it, the number of other people who have added that same link, and the time.

Tags are simply labels for URLs, selected to help the user in later retrieval of those URLs. Tags have the additional effect of grouping related URLs together. There is no fixed set of categories or officially approved choices. You can use words, acronyms, numbers, whatever makes sense to you, without regard for anyone else's needs, interests, or requirements.

The addition of a few simple labels hardly seems so momentous, but the surprise here, as so often with the Web, is the surprise of simplicity. Tags are important mainly for what they leave out. By forgoing formal classification, tags enable a huge amount of user-produced organizational value, at vanishingly small cost.

There's a useful comparison here between gopher and the Web, where gopher was better organized, better mapped to existing institutional practices, and utterly unfit to work at internet scale. The Web, by contrast, was and is a complete mess, with only one brand of pointer, the URL, and no mechanism for global organization or resources. The Web is mainly notable for two things -- the way it ignored most of the theories of hypertext and rich metadata, and how much better it works than any of the proposed alternatives. (The Yahoo/Google strategies I mentioned earlier also split along those lines.)

With those changes afoot, here are some of the things that I think are coming, as advantages of tagging systems:

- **Market Logic** - As we get used to the lack of physical constraints, as we internalize the fact that there is no shelf and there is no disk, we're moving towards market logic, where you deal with individual motivation, but group value.

  As Schachter says of del.icio.us, "Each individual categorization scheme is worth less than a professional categorization scheme. But there are many, many more of them." If you find a way to make it valuable to individuals to tag their stuff, you'll generate a lot more data about any given object than if you pay a professional to tag it once and only once. And if you can find any way to create value from combining myriad amateur classifications over time, they will come to be more valuable than professional categorization schemes, particularly with regards to robustness and cost of creation.

  The other essential value of market logic is that individual differences don't have to be homogenized. Look for the word 'queer' in almost any top-level categorization. You will not find it, even though, as an organizing principle for a large group of people, that word matters enormously. Users don't get to participate those kind of discussions around traditional categorization schemes, but with tagging, anyone is free to use the words he or she thinks are appropriate, without having to agree with anyone else about how something "should" be tagged. Market logic allows many distinct points of view to co-exist, because it allows individuals to preserve their point of view, even in the face of general disagreement.

- **User and Time are Core Attributes** - This is absolutely essential. The attitude of the Yahoo ontologist and her staff was -- "We are Yahoo We do not have biases. This is just how the world is. The world is organized into a dozen categories." You don't know who those people were, where they came from, what their background was, what their political biases might be.

  Here, because you can derive 'this is who this link is was tagged by' and 'this is when it was tagged, you can start to do inclusion and exclusion around people and time, not just tags. You can start to do grouping. You can start to do decay. "Roll up tags from just this group of users, I'd like to see what they are talking about" or "Give me all tags with this signature, but anything that's more than a week old or a year old."

  This is group tagging -- not the entire population, and not just me. It's like Unix permissions -- right now we've got tags for user and world, and this is the base on which we will be inventing group tags. We're going to start to be able to subset our categorization schemes. Instead of having massive categorizations and then specialty categorization, we're going to have a spectrum between them, based on the size and make-up of various tagging groups.

- **Signal Loss from Expression** - The signal loss in traditional categorization schemes comes from compressing things into a restricted number of categories. With tagging, when there is signal loss, it comes from people not having any commonality in talking about things. The loss is from the multiplicity of points of view, rather than from compression around a single point of

view. But in a world where enough points of view are likely to provide some commonality, the aggregate signal loss falls with scale in tagging systems, while it grows with scale in systems with single points of view.

The solution to this sort of signal loss is growth. Well-managed, well-groomed organizational schemes get worse with scale, both because the costs of supporting such schemes at large volumes are prohibitive, and, as I noted earlier, scaling over time is also a serious problem. Tagging, by contrast, gets better with scale. With a multiplicity of points of view the question isn't "Is everyone tagging any given link 'correctly'", but rather "Is anyone tagging it the way I do?" As long as at least one other person tags something they way you would, you'll find it -- using a thesaurus to force everyone's tags into tighter synchrony would actually worsen the noise you'll get with your signal. If there is no shelf, then even *imagining* that there is one right way to organize things is an error.

- **The Filtering is Done Post Hoc** - There's an analogy here with every journalist who has ever looked at the Web and said "Well, it needs an editor." The Web has an editor, it's everybody. In a world where publishing is expensive, the act of publishing is also a statement of quality -- the filter comes before the publication. In a world where publishing is cheap, putting something out there says nothing about its quality. It's what happens after it gets published that matters. If people don't point to it, other people won't read it. But the idea that the filtering is *after* the publishing is incredibly foreign to journalists.

  Similarly, the idea that the categorization is done after things are tagged is incredibly foreign to cataloguers. Much of the expense of existing catalogue systems is in trying to prevent one-off categories. With tagging, what you say is "As long as a lot of people are tagging any given link, the rare tags can be used or ignored, as the user likes. We won't even have to expend the cost to prevent people from using them. We'll just help other users ignore them if they want to."

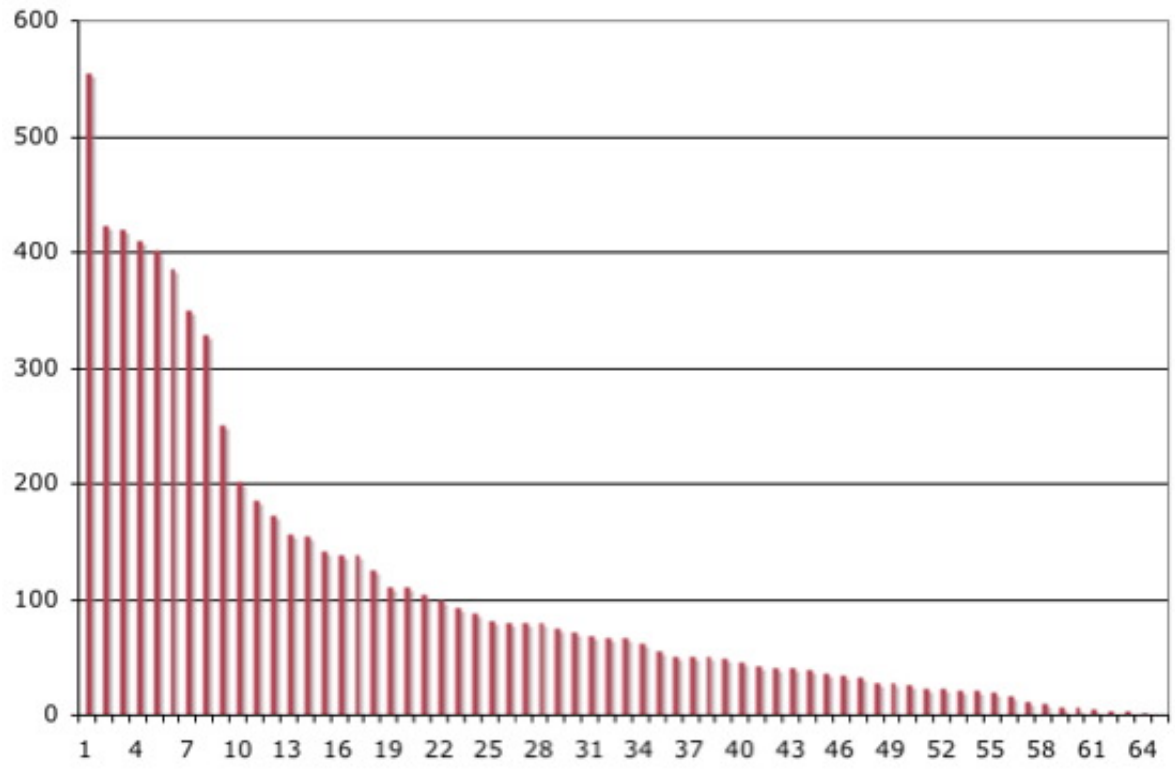
  Again, scale comes to the rescue of the system in a way that would simply break traditional cataloging schemes. The existence of an odd or unusual tag is a problem if it's the only way a given link has been tagged, or if there is no way for a user to avoid that tag. Once a link has been tagged more than once, though, users can view or ignore the odd tags as it suits them, and the decision about which tags to use comes after the links have been tagged, not before.

- **Merged from URLs, Not Categories** - You don't merge tagging schemes at the category level and then see what the contents are. As with the 'merging ISBNs' idea, you merge individual contents, because we now have URLs as unique handles. You merge from the URLs, and then try and derive something about the categorization from there. This allows for partial, incomplete, or probabilistic merges that are better fits to uncertain environments -- such as the real world -- than rigid classification schemes.
- **Merges are Probabilistic, not Binary** - Merges create partial overlap between tags, rather than defining tags as synonyms. Instead of saying that any given tag "is" or "is not" the same as another tag, del.icio.us is able to recommend related tags by saying "A lot of people who tagged this 'Mac' also tagged it 'OSX'." We move from a binary choice between saying two tags are the same or different to the Venn diagram option of "kind of is/somewhat is/sort of is/overlaps to

same or different to the Venn diagram option or "kind of is/somewhat is/sort of is/overlaps to this degree". That is a really profound change.

Tag Distributions on del.icio.us

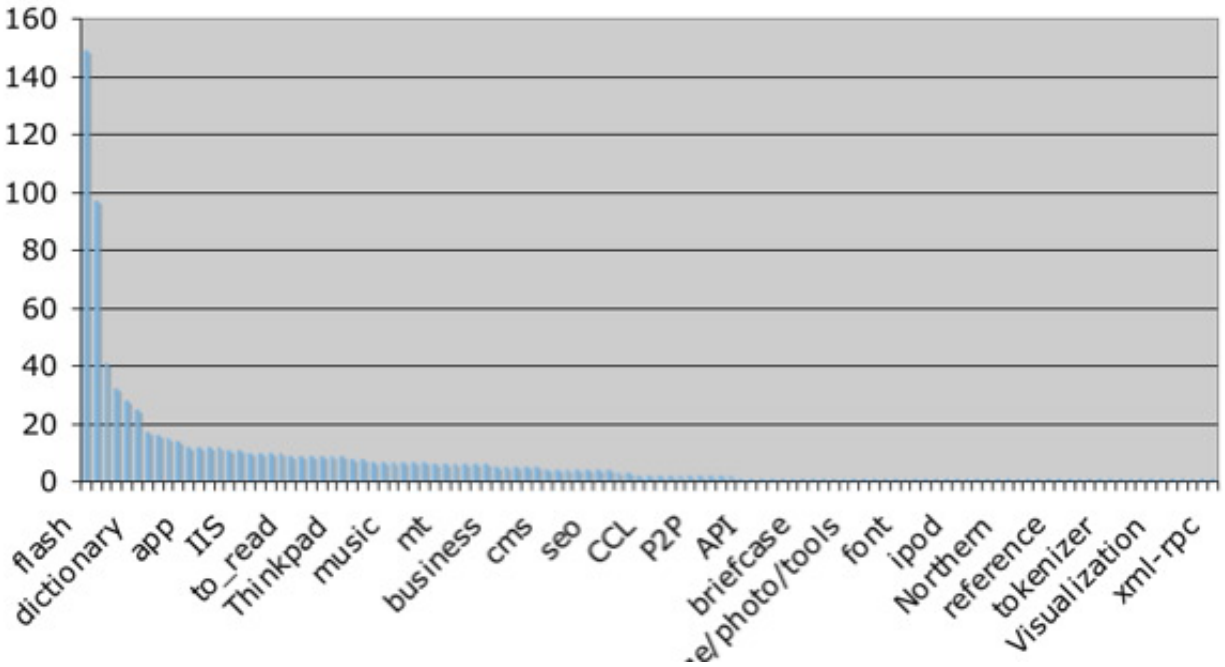
Here's something showing what I mean about the breakdown of binary categorization.



[Tags per user]

This is a chart based on a small sample of links from the del.icio.us front page, taken during a 2-hour window. The X axis is the 64 users who posted links during that period. The Y axis is the total number of discrete kinds of tags that those users have ever used in their history on del.icio.us.

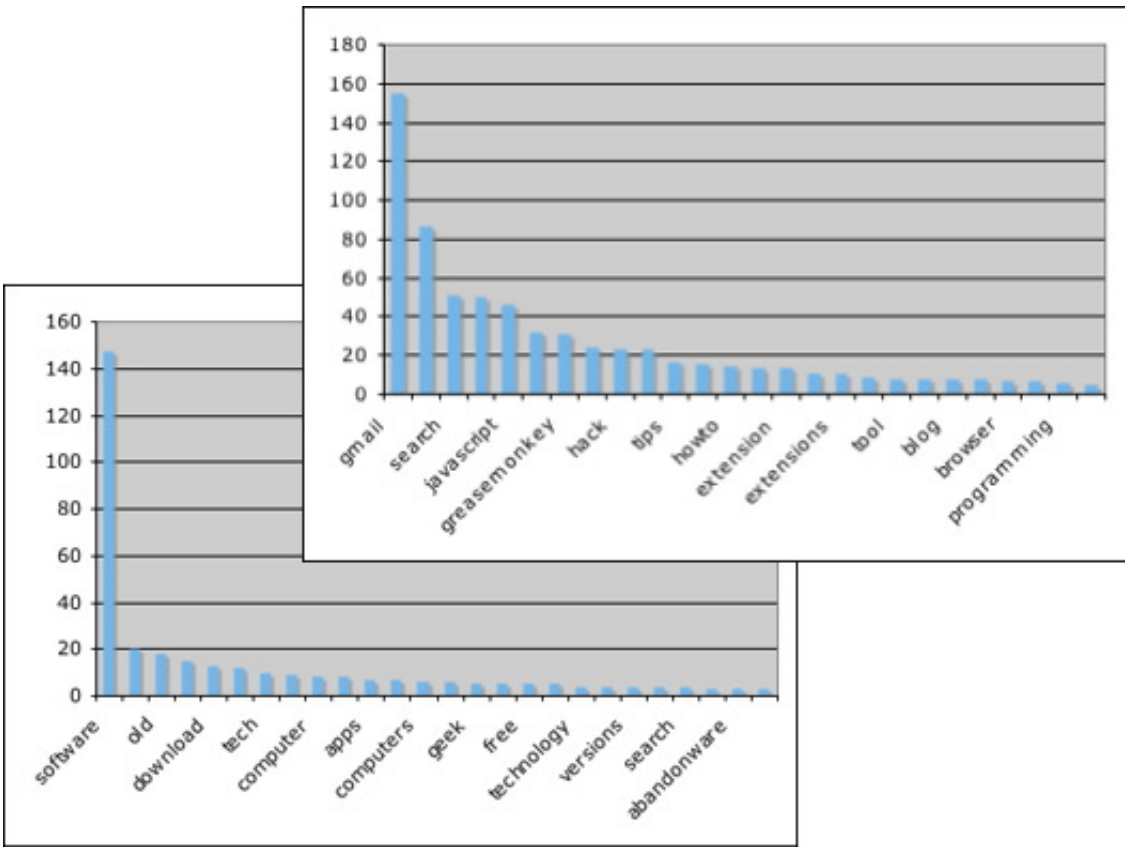
The chart shows a great variability in tagging strategies among the various users. The user all the way to the left has an enormous number of unique tags, almost 600 of them. Then there's this group of people who are not quite power taggers but who tag quite a bit, and of course to the right of them there's the characteristic long tail of people who use many fewer tags than the power taggers. (Because this is a two-hour snapshot, it has a natural bias towards frequent del.icio.us users. I'm trying to get a larger data set. My guess is the tail goes out quite a bit further than this.) But this is what organization looks like when you turn it over to the users -- many different strategies, each of which works in its own context, but which can also be merged.



[A single user's tags]

This is a single user's tags. From here, you can tell something about this person -- he or she is obviously a Flash programmer -- the commonest tag here is Flash, followed by a number of other frequently used tags mainly related to programming. Like the front page, this distribution has the organic signature. Experts don't catalog this way; experts who learn how to catalogue produce much more consistent labeling. Here, it's whatever the user thought would help them remember the link later.

You can see there's a tag "to_read". A professional cataloguer would look at this tag in horror -- "This is context-dependent and temporary." Well, so was the category "East Germany." Once you expand your time scale to include the actual life of the categorization scheme itself, you recognize that the distinction between temporary and permanent is awfully vague. There isn't in fact a binary condition of a tag that can or cannot survive any kind of long-term examination.



[Different tag 'signatures' for different URLs]

Then there's this set of graphs. This is to me in a way the most interesting and least well understood part of the del.icio.us right now -- these are two different URLs and the tags that a whole group of users applied to them. The graph at the bottom left refers to a site for downloading old versions of programs that are no longer supported. You can see here that there is broad communal consensus. 140 people tagged this Software. Then, the next commonest tag, with only 20 occurrences, is Windows, then Old, then Download, and so forth. For this URL, there's a core consensus -- this link is about software -- and after that one bit of commonality, there is a really sharp, clear fall off in tags.

The graph at the upper right, by contrast, shows the tags for a page detailing how to embed standing searches in Gmail. You can see the tags -- Gmail, Firefox, Search, Javascript, GreaseMonkey -- this is a much smearier distribution, with a much less sharp fall-off. The consensus view is that this link is about more kinds of things than the software download link is, or, rather, occupies more contexts for del.icio.us users than the software download link does.

Looking at this sort of data, we can start to say, of particular URLs, that the users tagging this URL either did or did not center around a certain core tags, with this degree of certainty, and, thanks to the time stamps, we can even start to understand how the distribution of a URLs tags changes over time. It was 5 years between the spread of the link and Google's figuring out how to use whole collections of links to create additional value. We're early in the use of tags, so we don't yet have large, long-lived data sets to look at, but they are being built up quickly, and we're just figuring out how to extract novel value from whole collections of tags.

Organization Goes Organic #

We are moving away from binary categorization -- books either are or are not entertainment -- and into this probabilistic world, where N% of users think books are entertainment. It may well be that within Yahoo, there was a big debate about whether or not books are entertainment. But they either had no way of reflecting that debate or they decided not to expose it to the users. What instead happened was it became an all-or-nothing categorization, "This is entertainment, this is not entertainment." We're moving away from that sort of absolute declaration, and towards being able to roll up this kind of value by observing how people handle it in practice.

It comes down ultimately to a question of philosophy. Does the world make sense or do we make sense of the world? If you believe the world makes sense, then anyone who tries to make sense of the world differently than you is presenting you with a situation that needs to be reconciled formally, because if you get it wrong, you're getting it wrong about the real world.

If, on the other hand, you believe that we make sense of the world, if we are, from a bunch of different points of view, applying some kind of sense to the world, then you don't privilege one top level of sense-making over the other. What you do instead is you try to find ways that the individual sense-making can roll up to something which is of value in aggregate, but you do it without an ontological goal. You do it without a goal of explicitly getting to or even closely matching some theoretically perfect view of the world.

Critically, the semantics here are in the users, not in the system. This is not a way to get computers to understand things. When del.icio.us is recommending tags to me, the system is not saying, "I know that OSX is an operating system. Therefore, I can use predicate logic to come up with recommendations -- users run software, software runs on operating systems, OSX is a type of operating system -- and then say 'Here Mr. User, you may like these links.'"

What it's doing instead is a lot simpler: "A lot of users tagging things foobar are also tagging them frobnitz. I'll tell the user foobar and frobnitz are related." It's up to the user to decide whether or not that recommendation is useful -- del.icio.us has no idea what the tags *mean*. The tag overlap is in the system, but the tag semantics are in the users. This is not a way to inject linguistic meaning into the machine.

It's all dependent on human context. This is what we're starting to see with del.icio.us, with Flickr, with systems that are allowing for and aggregating tags. The signal benefit of these systems is that they don't recreate the structured, hierarchical categorization so often forced onto us by our physical systems. Instead, we're dealing with a significant break -- by letting users tag URLs and then aggregating those tags, we're going to be able to build alternate organizational systems, systems that, like the Web itself, do a better job of letting individuals create value for one another, often without realizing it.

Thank you very much.

Thanks to Alicia Cervini for invaluable editorial help.

II

Digital Humanities

Anderson, Chris (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

Berry, David M. (2011). “The Computational Turn: Thinking About the Digital Humanities.” Culture Machine. Vol 12. http://www.culturemachine.net/index.php/cm/article/view/440/470

Manovich, Lev. "Trending: The Promises and the Challenges of Big Social Data." Debates in the Digital Humanities, edited by Matthew K. Gold. The University of Minnesota Press, forthcoming 2012. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Tooling Up for Digital Humanities: Digitization. Stanford University, 2011. http://toolingup.stanford.edu/?page_id=123

THE END OF THEORY By Chris Anderson

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age.

The Petabyte Age is different because more is different. Kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to — well, at petabytes we ran out of organizational analogies.

At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day. And Google was right.

Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them (given equal corpus data, Google can translate Klingon into Farsi as easily as it can translate French into German). And why it can match ads to content without any knowledge or assumptions about the ads or the content.

Speaking at the O'Reilly Emerging Technology Conference this past March, Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."

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.

The big target here isn't advertising, though. It's science. The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years.

Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. Consider physics: Newtonian models were crude approximations of the truth (wrong at the atomic level, but still useful). A hundred years ago, statistically based quantum mechanics offered a better picture — but quantum mechanics is yet another model, and as such it, too, is flawed, no doubt a caricature of a more complex underlying reality. The reason physics has drifted into theoretical speculation about *n*-dimensional grand unified models over the past few decades (the "beautiful story" phase of a discipline starved of data) is that we don't know how to run the experiments that would falsify the hypotheses — the energies are too high, the accelerators too expensive, and so on.

Now biology is heading in the same direction. The models we were taught in school about "dominant" and "recessive" genes steering a strictly Mendelian process have turned out to be an even greater simplification of reality than Newton's laws. The discovery of gene-protein interactions and other aspects of epigenetics has challenged the view of DNA as destiny and even introduced evidence that environment can influence inheritable traits, something once considered a genetic impossibility.

In short, the more we learn about biology, the further we find ourselves from a model that can explain it.

There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

The best practical example of this is the shotgun gene sequencing by J. Craig Venter. Enabled by high-speed sequencers and supercomputers that statistically analyze the data they produce, Venter went from sequencing individual organisms to sequencing entire ecosystems. In 2003, he started sequencing much of the ocean, retracing the voyage of Captain Cook. And in 2005 he started sequencing the air. In the process, he discovered thousands of previously unknown species of bacteria and other life-forms.

If the words "discover a new species" call to mind Darwin and drawings of finches, you may be stuck in the old way of doing science. Venter can tell you almost nothing about the species he found. He doesn't know what they look like, how they live, or much of anything else about their morphology. He doesn't even have their entire genome. All he has is a statistical blip — a unique sequence that, being unlike any other sequence in the database, must represent a new species.

This sequence may correlate with other sequences that resemble those of species we do know more about. In that case, Venter can make some guesses about the animals — that they convert sunlight into energy in a particular way, or that they descended from a common ancestor. But besides that, he has no better model of this species than Google has of your MySpace page. It's just data. By analyzing it with Google-quality computing resources, though, Venter has advanced biology more than anyone else of his generation.

This kind of thinking is poised to go mainstream. In February, the National Science Foundation announced the Cluster Exploratory, a program that funds research designed to run on a large-scale distributed computing platform developed by Google and IBM in conjunction with six pilot universities. The cluster will consist of 1,600 processors, several terabytes of memory, and hundreds of terabytes of storage, along with the software, including IBM's Tivoli and open source versions of Google File System and MapReduce.[1] Early CluE projects will include simulations of the brain and the nervous system and other biological research that lies somewhere between wetware and software.

Learning to use a "computer" of this scale may be challenging. But the opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?

Original URL:

http://www.edge.org/3rd_culture/anderson08/anderson08_index.html

THE COMPUTATIONAL TURN: THINKING ABOUT THE DIGITAL HUMANITIES

David M. Berry

Introduction

Few dispute that digital technology is fundamentally changing the way in which we engage in the research process. Indeed, it is becoming more and more evident that research is increasingly being mediated through digital technology. Many argue that this mediation is slowly beginning to change what it means to undertake research, affecting both the epistemologies and ontologies that underlie a research programme. Of course, this development is variable depending on disciplines and research agendas, with some more reliant on digital technology than others, but it is rare to find an academic today who has had no access to digital technology as part of their research activity. Library catalogues are now probably the minimum way in which an academic can access books and research articles without the use of a computer, but, with card indexes dying a slow and certain death (Baker, 1996, 2001), there remain few outputs for the non-digital scholar to undertake research in the modern university. Email, Google searches and bibliographic databases are become increasingly crucial, as more of the world libraries are scanned and placed online. Whilst some decry the loss of the skills and techniques of older research traditions, others have warmly embraced what has come to be called the digital humanities (Schreibman *et al.*, 2008; Schnapp & Presner, 2009; Presner, 2010; Hayles, 2011).

The digital humanities try to take account of the plasticity of digital forms and the way in which they point toward a new way of working with representation and mediation, what might be called the digital 'folding' of reality, whereby one is able to approach culture in a radically new way. To mediate an object, a digital or *computational* device requires that this object be translated into the digital code that it can understand. This minimal transformation is effected

through the input mechanism of a socio-technical device within which a model or image is stabilised and attended to. It is then internally transformed, depending on a number of interventions, processes or filters, and eventually displayed as a final calculation, usually in a visual form. This results in real-world situations where computation is event-driven and divided into discrete processes to undertake a particular user task. The key point is that without the possibility of *discrete* encoding there is no object for the computational device to process. However, in cutting up the world in this manner, information about the world necessarily has to be discarded in order to store a representation within the computer. In other words, a computer requires that everything is transformed from the continuous flow of our everyday reality into a grid of numbers that can be stored as a representation of reality which can then be manipulated using algorithms. These subtractive methods of understanding reality (*episteme*) produce new knowledges and methods for the control of reality (*techne*). They do so through a digital mediation, which the digital humanities are starting to take seriously as their problematic.

The digital humanities themselves have had a rather interesting history. Starting out as 'computing in the humanities', or 'humanities computing', in the early days they were often seen as a technical support to the work of the 'real' humanities scholars, who would drive the projects. This involved the application of the computer to the disciplines of the humanities, something that has been described as treating the 'machine's efficiency as a servant' rather than 'its participant enabling of criticism' (McCarty, 2009). As Hayles explains, changing to the term '"Digital Humanities" was meant to signal that the field had emerged from the low-prestige status of a support service into a genuinely intellectual endeavour with its own professional practices, rigorous standards, and exciting theoretical explorations' (Hayles, 2011). Ironically, as the projects became bigger and more complex, and as it developed computational techniques as an intrinsic part of the research process, technically proficient researchers increasingly saw the computational as part and parcel of what it meant to do research in the humanities itself. That is, computational technology has become the very condition of possibility required in order to think about many of the questions raised in the humanities today. For example, as Schnapp and Presner explain in the *Digital Humanities Manifesto 2.0*,

> The first wave of digital humanities work was quantitative, mobilizing the search and retrieval

> powers of the database, automating corpus linguistics, stacking hypercards into critical arrays. The second wave is *qualitative, interpretive, experiential, emotive, generative in character*. It harnesses digital toolkits in the service of the Humanities' core methodological strengths: attention to complexity, medium specificity, historical context, analytical depth, critique and interpretation. (2009, original emphasis)

Presner argues further that

> the first wave of Digital Humanities scholarship in the late 1990s and early 2000s tended to focus on large-scale digitization projects and the establishment of technological infrastructure, [while] the current second wave of Digital Humanities -- what can be called 'Digital Humanities 2.0' -- is deeply generative, creating the environments and tools for producing, curating, and interacting with knowledge that is 'born digital' and lives in various digital contexts. While the first wave of Digital Humanities concentrated, perhaps somewhat narrowly, on text analysis (such as classification systems, mark-up, text encoding, and scholarly editing) within established disciplines, Digital Humanities 2.0 introduces entirely new disciplinary paradigms, convergent fields, hybrid methodologies, and even new publication models that are often not derived from or limited to print culture. (2010: 6)

The question of quite how the digital humanities undertake their research, and whether the notions of first and second wave digital humanities captures the current state of different working practices and methods in the digital humanities, remains contested. Yet these can be useful analytical concepts for thinking through the changes in the digital humanities. We might, however, observe the following: first-wave digital humanities involved the building of infrastructure in the studying of humanities texts through digital repositories, text markup, etc., whereas second-wave digital humanities expands the notional limits of the archive to include digital works, and so bring to bear the humanities' own methodological toolkits to look at 'born-

digital' materials, such as electronic literature (e-lit), interactive fiction (IF), web-based artefacts, and so forth.

I would like to explore here a tentative path for a third wave of the digital humanities, concentrated around the underlying *computationality* of the forms held within a computational medium. That is, I propose to look at the *digital* component of the digital humanities in the light of its medium specificity, as a way of thinking about how medial changes produce epistemic changes. This approach draws from recent work in software studies and critical code studies, but it also thinks about the questions raised by platform studies, namely the specifics of general computability made available by specific platforms (Fuller, 2008; Manovich, 2008; Montfort & Bogost, 2009; Berry, 2011). I also want to suggest that neither first nor second-wave digital humanities really problematized what Lakatos (1980) would have called the 'hard-core' of the humanities, the unspoken assumptions and ontological foundations which support the 'normal' research that humanities scholars undertake on an everyday basis. Indeed, we could say that third-wave digital humanities points the way in which digital technology highlights the anomalies generated in a humanities research project and which leads to the questioning of the assumptions implicit in such research, e.g. close reading, canon formation, periodization, liberal humanism, etc. We are, as Presner argues, 'at the beginning of a shift in standards governing permissible problems, concepts, and explanations, and also in the midst of a transformation of the institutional and conceptual conditions of possibility for the generation, transmission, accessibility, and preservation of knowledge' (2010: 10).

To look into this issue, I want to start with an examination of the complex field of understanding culture through digital technology. Indeed, I argue that to understand the contemporary born-digital culture and the everyday practices that populate it – the focus of a digital humanities second wave – we need a corresponding focus on the computer code that is entangled with all aspects of our lives, including reflexivity about how much code is infiltrating the academy itself. As Mathew Fuller argues, 'in a sense, all intellectual work is now "software study", in that software provides its media and its context… [yet] there are very few places where the specific nature, the materiality, of software is studied except as a matter of engineering' (2006). We also need to bring to the fore the 'structure of feeling' that computer code facilitates and the way in which people use software in their research thinking and everyday

practices. This includes the increase in the acceptance and use of software in the production, consumption and critique of culture.

Thus, there is an undeniable cultural dimension to computation and the medial affordances of software. This connection again points to the importance of engaging with and understanding code: indeed, computer code can serve as an index of digital culture (imagine digital humanities mapping different programming languages to the cultural possibilities and practices that it affords, e.g. HTML to cyberculture, AJAX to social media).[1] This means that we can ask the question: what is culture after it has been 'softwarized'? (Manovich, 2008:41). Understanding digital humanities is in some sense then understanding code, and this can be a resourceful way of understanding cultural production more generally: for example, just as digital typesetting transformed the print newspaper industry, eBook and eInk technologies are likely to do so again. We thus need to take computation as the key issue that is underlying these changes across mediums, industries and economies.

Knowing knowledge

In trying to understand the digital humanities our first step might be to problematize *computationality*, so that we are able to think critically about how knowledge in the 21st century is transformed into information through computational techniques, particularly within software. It is interesting that at a time when the idea of the university is itself under serious rethinking and renegotiation, digital technologies are transforming our ability to use and understand information outside of these traditional knowledge structures. This is connected to wider challenges to the traditional narratives that served as unifying ideas for the university and, with their decline, has led to difficulty in justifying and legitimating the postmodern university vis-à-vis government funding.

Historically, the role of the university has been closely associated with the production of knowledge. For example, in 1798 Immanuel Kant outlined an argument for the nature of the university titled *The Conflict of the Faculties.* He argued that all of the university's activities should be organised by a single regulatory idea, that of the concept of reason. As Bill Readings (1996) stated:

> Reason on the one hand, provide[d] the *ratio* for all the disciplines; it [was] their organizing

> principle. On the other hand, reason [had] its own faculty, which Kant names[d] 'philosophy' but which we would now be more likely to call the 'humanities'. (Readings, 1996: 15)

Kant argued that reason and the state, knowledge and power, could be unified in the university by the production of individuals who would capable of rational thought and republican politics – students trained for the civil service and society. Kant was concerned with the question of regulative public reason, that is, with how to ensure stable, governed and governable regimes which can rule free people, in contrast to tradition represented by monarchy, the Church or a Leviathan. This required universities, as regulated knowledge-producing organisations, to be guided and overseen by the faculty of philosophy, which could ensure that the university remained rational. This was part of a response to the rise of print culture, growing literacy and the kinds of destabilising effects that this brought. Thus, without resorting to dogmatic doctrinal force or violence, one could have a form of perpetual peace by the application of one's reason.[2]

This was followed by the development of the modern university in the 19th century, instituted by the German Idealists, such as Schiller and Humboldt, who argued that there should be a more explicitly political role to the structure given by Kant. They argued for the replacement of reason with culture, as they believed that culture could serve as a 'unifying function for the university' (Readings, 1996: 15). For the German Idealists like Humboldt, culture was the sum of all knowledge that is studied, as well as the cultivation and development of one's character as a result of that study. Indeed, Humboldt proposed the founding of a new university, the University of Berlin, as a mediator between national culture and the nation-state. Under the project of 'culture', the university would be required to undertake both research and teaching, i.e., the production and dissemination of knowledge respectively. The modern idea of a university therefore allowed it to become the preeminent institution that unified ethnic tradition and statist rationality by the production of an educated cultured individual. The German Idealists proposed

> that the way to reintegrate the multiplicity of known facts into a unified cultural science is through *Bildung*, the ennoblement of character… The university produces not servants but *subjects*.

> That is the point of the pedagogy of *Bildung*, which teaches knowledge acquisition as a *process* rather than the acquisition of knowledge as a product. (Reading, 1996: 65-67)

This notion was given a literary turn by the English, in particular John Henry Newman and Mathew Arnold, who argued that literature, not culture or philosophy, should be the central discipline in the university, and also in national culture more generally.[3] Literature therefore became institutionalised within the university 'in explicitly national terms and [through] an organic vision of the possibility of a unified national culture' (Readings, 1996: 16). This became regulated through the notion of a literary canon, which was taught to students to produce literary subjects as national subjects.

Readings argues that in the postmodern university we now see the breakdown of these ideals, associated particularly with the rise of the notion of the 'university of excellence' -- which for him is a concept of the university that has no content, no referent. What I would like to suggest is that today, we are beginning to see instead the cultural importance of the digital as the unifying idea of the university. Initially this has tended to be associated with notions such as *information literacy* and *digital literacy*, betraying their debt to the previous literary conception of the university, albeit understood through vocational training and employment. However, I want to propose that, rather than learning a *practice* for the digital, which tends to be conceptualised in terms of ICT skills and competences (see for example the European Computer Driving License[4]), we should be thinking about what reading and writing actually should mean in a computational age. This is to argue for critical understanding of the *literature* of the digital, and through that develop a shared digital culture through a form of digital *Bildung*. Here I am not calling for a return to the humanities of the past, to use a phrase of Fuller (2010), 'for some humans', but rather to a liberal arts that is 'for all humans'. To use the distinction introduced by Hofstadter (1963), this is to call for the development of a digital *intellect* -- as opposed to a digital *intelligence*. Hofstadter writes:

> Intellect… is the critical, creative, and contemplative side of mind. Whereas intelligence seeks to grasp, manipulate, re-order, adjust, intellect examines, ponders, wonders, theorizes, criticizes, imagines. Intelligence will seize the immediate meaning in a situation and evaluate it.

> Intellect evaluates evaluations, and looks for the meanings of situations as a whole… Intellect [is] a unique manifestation of human dignity. (Hofstadter, 1963: 25)

The digital assemblages that are now being built not only promise great change at the level of the individual human actor. They provide destabilising amounts of knowledge and information that lack the regulating force of philosophy -- which, Kant argued, ensures that institutions remain rational. Technology enables access to the databanks of human knowledge from anywhere, disregarding and bypassing the traditional gatekeepers of knowledge in the state, the universities and the market. There no longer seems to be the professor who tells you what you should be looking up and the 'three arguments in favour of it' and the 'three arguments against it'. This introduces not only a moment of societal disorientation, with individuals and institutions flooded with information, but also offers a computational solution to this state of events in the form of computational rationalities--something that Turing (1950) described as super-critical modes of thought. Both of these forces are underpinned at a deep structural level by the conditions of possibility suggested by computer code.

As mentioned previously, computer code enables new communicative processes, and with the increasing social dimension of networked media the possibility of new and exciting forms of collaborative thinking arises. This is not the collective intelligence discussed by Levy (1999); rather, it is the promise of a collective *intellect*. The situation is reminiscent of the medieval notion of the *universitatis*, but recast in a digital form, as a society or association of actors who can think critically together, mediated through technology. It further raises the question of what new modes of collective knowledge software can enable or constitute. Can software and code take us beyond the individualising trends of blogs, comments, twitter feeds, and so forth, and make possible something truly collaborative -- something like the super-critical thinking that is generative of ideas, modes of thought, theories and new practices? There is certainly something interesting about real-time stream forms of digital memory in that they are not afforded towards the past, as history, but neither are they directed towards a form of futurity. Instead we might say they seem to now-mediate? new-mediate? life-mediate? *Jetztzeit*-mediate (Benjamin, 1992: 252-3)? In other words, they gather together the newness of a particular group of streams, a kind of collective writing, that has the potential

to be immensely creative. These are possible rich areas for research for a third-wave digital humanities that seeks to understand these potentially new forms of literature and the medium that supports them.

For the research and teaching disciplines within the university, the digital shift could represent the beginnings of a moment of 'revolutionary science', in the Kuhnian sense of a shift in the ontology of the positive sciences and the emergence of a constellation of new 'normal science' (Kuhn 1996). This would mean that the disciplines would, ontologically, have a very similar Lakatosian computational 'hard core' (Lakatos, 1980).[5] This has much wider consequences for the notion of the unification of knowledge and the idea of the university (Readings, 1996). Computer science could play a foundational role with respect to the other sciences, supporting and directing their development, even issuing 'lucid directives for their inquiry'.[6] Perhaps we are beginning to see reading and writing computer code as part of the pedagogy required to create a new subject produced by the university, a *computational* or *data-centric* subject.[7] This is, of course, not to advocate that the *existing* methods and practices of computer science become hegemonic, rather that a *humanistic* understanding of technology could be developed, which also involves an urgent inquiry into what is human about the *computational* humanities or social sciences. In a related manner, Fuller (Fuller, S., 2006) has called for a 'new sociological imagination', pointing to the historical project of the social sciences that have been committed to 'all and only humans', because they 'take all human beings to be of equal epistemic interest and moral concern' (Fuller, 2010: 242). By drawing attention to 'humanity's ontological precariousness' (244), Fuller rightly identifies that the project of humanity requires urgent thought, and, we might add, even more so in relation to the challenge of a *computationality* that threatens our understanding of what is required to be identified as human at all.

If software and code become the condition of possibility for unifying the multiple knowledges now produced in the university, then the ability to think oneself, taught by rote learning of methods, calculation, equations, readings, canons, processes, etc., might become less important. Although there might be less need for an *individual* ability to perform these mental feats or, perhaps, even recall the entire canon ourselves due to its size and scope, using technical devices, in conjunction with collaborative methods of working and studying, would enable a cognitively supported method

instead. The internalisation of particular practices that have been instilled for hundreds of years in children and students would need to be rethought, and in doing so the commonality of thinking *qua* thinking produced by this pedagogy would also change. Instead, reasoning could shift to a more conceptual or communicative method of reasoning, for example, by bringing together comparative and communicative analysis from different disciplinary perspectives, and by knowing how to use technology to achieve a usable result – a rolling process of reflexive thinking and collaborative rethinking.

Relying on technology in a more radically decentred way, depending on technical devices to fill in the blanks in our minds and to connect knowledge in new ways, would change our understanding of knowledge, wisdom and intelligence itself. It would be a radical decentring in some ways, as the Humboldtian subject filled with culture and a certain notion of rationality would no longer exist; rather, the computational subject would know where to recall culture as and when it was needed in conjunction with computationally available others, a *just-in-time* cultural subject, perhaps, to feed into a certain form of connected *computationally* supported thinking through and visualised presentation. Rather than a method of thinking with eyes and hand, we would have a method of thinking with eyes and screen.[8]

This doesn't have to be dehumanising. Latour and others have rightly identified the domestication of the human mind that took place with pen and paper (Latour, 1986). This is because computers, like pen and paper, help to stabilise meaning by cascading and visualising encoded knowledge that allows it to be continually 'drawn, written, [and] recoded' (Latour, 1986: 16). Computational techniques could give us greater powers of thinking, larger reach for our imaginations, and, possibly, allow us to reconnect to political notions of equality and redistribution based on the potential of computation to give to each according to their need and to each according to their ability. This is the point made forcefully by Fuller (2010: 262), who argues that we should look critically at the potential for inequality which is created when new technologies are introduced into society. This is not merely a problem of a 'digital divide', but a more fundamental one of how we classify those that are more 'human' than others, when access to computation and information increasingly has to pass through the market.

Towards a digital humanities?

The importance of understanding computational approaches is increasingly reflected across a number of disciplines, including the arts, humanities and social sciences, which use technologies to shift the critical ground of their concepts and theories – something that can be termed a *computational turn*.[9] This is shown in the increasing interest in the *digital humanities* (Schreibman *et al*., 2008) and *computational social science* (Lazer *et al*., 2009), as evidenced, for example, by the growth in journals, conferences, books and research funding. In the digital humanities 'critical inquiry involves the application of algorithmically facilitated search, retrieval, and critical process that… originat[es] in humanities-based work'; therefore 'exemplary tasks traditionally associated with humanities computing hold the digital representation of archival materials on a par with analysis or critical inquiry, as well as theories of analysis or critical inquiry originating in the study of those materials' (Schreibman *et al*., 2008: xxv). In social sciences, Lazer *et al*. argue that 'computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale' (2009).

Latour speculates that there is a trend in these informational cascades, which is certainly reflected in the ongoing digitalisation of arts, humanities and social science projects that tends towards 'the direction of the greater merging of figures, numbers and letters, merging greatly facilitated by their homogenous treatment as binary units in and by computers' (Latour, 1986: 16). The financial considerations are also new with these computational disciplines, as they require more money and organisation than the old individual scholar of lore did. Not only are the start-up costs correspondingly greater, usually needed to pay for the researchers, computer programmers, computer technology, software, digitisation costs, etc., but there are real questions about sustainability of digital projects, such as: 'Who will pay to maintain the digital resources?' 'Will the user forums, and user contributions, continue to be monitored and moderated if we can't afford a staff member to do so? Will the wiki get locked down at the close of funding or will we leave it to its own devices, becoming an online-free-for all?' (Terras, 2010).[10] It also raises a lot of new ethical questions for social scientists and humanists to grapple with. As argued in *Nature*,

> For a certain sort of social scientist, the traffic patterns of millions of e-mails look like manna

> from heaven. Such data sets allow them to map formal and informal networks and pecking orders, to see how interactions affect an organization's function, and to watch these elements evolve over time. They are emblematic of the vast amounts of structured information opening up new ways to study communities and societies. Such research could provide much-needed insight into some of the most pressing issues of our day, from the functioning of religious fundamentalism to the way behaviour influences epidemics… But for such research to flourish, it must engender that which it seeks to describe… Any data on human subjects inevitably raise privacy issues, and the real risks of abuse of such data are difficult to quantify, (*Nature,* 2007)

For Latour, 'sociology has been obsessed by the goal of becoming a quantitative science. Yet it has never been able to reach this goal because of what it has defined as being quantifiable within the social domain…'. Thus, he adds, '[i]t is indeed striking that at this very moment, the fast expanding fields of "data visualisation", "computational social science" or "biological networks" are tracing, before our eyes, just the sort of data' that sociologists such as Gabriel Tarde, at the turn of the 20th century, could merely speculate about (Latour, 2010: 116).

Further, it is not merely the quantification of research which was traditionally qualitative that is offered with these approaches. Rather, as Unsworth argues, we should think of these computational 'tools as offering provocations, surfacing evidence, suggesting patterns and structures, or adumbrating trends' (Unsworth, quoted in Clement *et al.*, 2008). For example, the methods of 'cultural analytics' make it possible, through the use of quantitative computational techniques, to understand and follow large-scale cultural, social and political processes for research projects – that is, it offers massive amounts of literary or visual data analysis (see Manovich and Douglas, 2009). This is a distinction that Moretti (2007) referred to as *distant* versus *close* readings of texts. As he points out, the traditional humanities focuses on a 'minimal fraction of the literary field',

> A canon of two hundred novels, for instance, sounds very large for nineteenth-century Britain

> (and is much larger than the current one), but is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows -- and close reading won't help here, a novel a day every day of the year would take a century or so... And it's not even a matter of time, but of method: a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn't a sum of individual cases: it's a collective system, that should be grasped as such, as a whole, (Moretti, 2007: 3-4)

It is difficult for the traditional arts, humanities and social sciences to completely ignore the large-scale digitalisation effort going on around them, particularly when large quantities of research money are available to create archives, tools and methods in the digital humanities and computational social sciences. However, less understood is the way in which the digital archives being created are deeply computational in structure *and* content, because the computational logic is entangled with the digital representations of physical objects, texts and 'born digital' artefacts. Computational techniques are not merely an instrument wielded by traditional methods; rather they have profound effects on all aspects of the disciplines. Not only do they introduce new methods, which tend to focus on the identification of novel patterns in the data as against the principle of narrative and understanding, they also allow the modularisation and recombination of disciplines within the university itself.

Computational approaches facilitate disciplinary hybridity that leads to a post-disciplinary university -- which can be deeply unsettling to traditional academic knowledge. Software allows for new ways of reading and writing. For example, this is what Tanya Clement says on the distant reading of Gertrude Stein's *The Making of Americans*,

> *The Making of Americans* was criticized by [those] like Malcolm Cowley who said Stein's 'experiments in grammar' made this novel 'one of the hardest books to read from beginning to end that has ever been published'…. The highly repetitive nature of the text, comprising almost 900 pages and 3174 paragraphs with only approximately 5,000 unique words, makes

> keeping tracks of lists of repetitive elements unmanageable and ultimately incomprehensible... [However] text mining allowed me to use statistical methods to chart repetition across thousands of paragraphs… facilitated my ability to read the results by allowing me to sort those results in different ways and view them within the context of the text. As a result, by visualizing clustered patterns across the text's 900 pages of repetitions… [th]is discovery provides a new key for reading the text as a circular text with two corresponding halves, which substantiates and extends the critical perspective that *Making* is neither inchoate nor chaotic, but a highly systematic and controlled text. This perspective will change how scholars read and teach *The Making of Americans*. (Clement, quoted in Clement, Steger, Unsworth, and Uszkalo, 2008)

I wouldn't want to overplay the distinction between pattern and narrative as differing modes of analysis. Indeed, patterns implicitly require narrative in order to be understood, and it can be argued that code itself consists of a narrative form that allows databases, collections and archives to function at all. Nonetheless, pattern and narrative are useful analytic terms that enable us to see the way in which the computational turn is changing the nature of knowledge in the university and, with it, the kind of computational subject that the university is beginning to produce. As Bruce Sterling argues,

> 'Humanistic heavy iron': it's taken a long time for the humanities to get into super computing, and into massive database management. They are really starting to get there now. You are going to get into a situation where even English professors are able to study every word ever written about, or for, or because of, Charles Dickens or Elizabeth Barrett Browning. That's just a different way to approach the literary corpus. I think there is a lot of potential there. (Sterling, 2010)

Indeed, there is a cultural dimension to this process and, as we become more used to computational visualisations, we will expect to see them and use them with confidence and fluency. The computational subject is a key requirement for a data-centric age,

certainly when we begin to look at case studies that demonstrate how important a computational comportment can be in order to perform certain forms of public and private activities in a world that is increasingly pervaded by computational devices. In short, *Bildung* is still a key idea in the digital university, not as a subject trained in a vocational fashion to perform instrumental labour, nor as a subject skilled in a national literary culture, but rather as a subject which can unify the information that society is now producing at increasing rates, and which understands new methods and practices of critical reading (code, data visualisation, patterns, narrative) and is open to new methods of pedagogy to facilitate it. Indeed, Presner (2010) argues that the digital humanities

> must be engaged with the broad horizon of possibilities for building upon excellence in the humanities while also transforming our research culture, our curriculum, our departmental and disciplinary structures, our tenure and promotion standards, and, most of all, the media and format of our scholarly publications. (Presner, 2010: 6)

This is a subject that is highly computationally communicative, and that is also able to access, process and visualise information and results quickly and effectively. At all levels of society, people will increasingly have to turn data and information into usable computational forms in order to understand it at all. For example, one could imagine a form of *computational* journalism that enables the public sphere function of the media to make sense of the large amount of data which governments, amongst others, are generating, perhaps through increasing use of 'charticles', or journalistic articles that combine text, image, video, computational applications and interactivity (Stickney, 2008). This is a form of 'networked' journalism that 'becomes a non-linear, multi- dimensional process' (Beckett, 2008: 65). Additionally, for people in everyday life who need the skills that enable them to negotiate an increasingly computational field – one need only think of the amount of data in regard to managing personal money, music, film, text, news, email, pensions, etc. – there will be calls for new skills of financial and technical literacy, or, more generally, a *computational literacy* or *computational pedagogy* that the digital humanities could contribute to.

Humanity and the humanities

As the advantages of the computational approach to research (and teaching) become persuasive to the positive sciences, whether history, biology, literature or any other discipline, the ontological notion of the entities they study begins to be transformed. These disciplines thus become focused on the *computationality* of the entities in their work.[11] Here, following Heidegger, I want to argue that there remains a location for the possibility of philosophy to explicitly question the ontological understanding of what the computational is in regard to these positive sciences. Computationality might then be understood as an ontotheology, creating a new ontological 'epoch' as a new historical constellation of intelligibility. The digital humanists could therefore orient themselves to questions raised when computationality is itself problematized in this way (see Liu 2011).

With the notion of ontotheology, Heidegger is following Kant's argument that intelligibility is a process of filtering and organising a complex overwhelming world by the use of 'categories', Kant's 'discursivity thesis'. Heidegger historicizes Kant's cognitive categories by arguing that there is 'succession of changing historical ontotheologies that make up the "core" of the metaphysical tradition. These ontotheologies establish "the truth concerning entities as such and as a whole", in other words, they tell us both what and how entities are – establishing both their essence and their existence' (Thomson, 2009: 149-150). Metaphysics, grasped ontotheologically, 'temporarily secures the intelligible order' by understanding it 'ontologically', from the inside out, and 'theologically', from the outside in, which allows the formation of an epoch, a 'historical constellation of intelligibility which is unified around its ontotheological understanding of the being of entities' (Thomson, 2009: 150). As Thomson argues:

> The positive sciences all study classes of entities… Heidegger… [therefore] refers to the positive sciences as 'ontic sciences'. Philosophy, on the other hand, studies the being of those classes of entities, making philosophy an 'ontological science' or, more grandly, a 'science of being' (Thomson 2003: 529).

Philosophy as a field of inquiry, one might argue, should have its 'eye on the whole', and it is this focus on 'the landscape as a whole' which

distinguishes the philosophical enterprise and which can be extremely useful in trying to understand these ontotheological developments (Sellars, 1962: 36). If code and software are to become objects of research for the humanities and social sciences, including philosophy, we will need to grasp both the ontic and ontological dimensions of computer code. Broadly speaking, then, this paper suggests that we take a philosophical approach to the subject of computer code, paying attention to the wider aspects of code and software, and connecting them to the materiality of this growing digital world. With this in mind, the question of code becomes central to understanding in the digital humanities, and serves as a condition of possibility for the many computational forms that mediate out experience of contemporary culture and society.

Endotes

[1] HTML is the HyperText Markup Language used to encode webpages. AJAX is shorthand for Asynchronous JavaScript and XML, which is a collection of client side technologies that enable an interactive and audio-visual dynamic web.

[2] I am indebted to Alan Finlayson for his comments on this section.

[3] For example in *The Idea of a University* (Newman, 1996) and *Culture and Anarchy* (Arnold, 2009).

[4] See http://www.bcs.org/server.php?show=nav.5829

[5] What Heidegger calls 'the Danger' (*die Gefahr*) is the idea that a particular ontotheology should become permanent, particularly the ontotheology associated with technology and enframing (see Heidegger 1993).

[6] See Thomson (2003: 531) for a discussion of how Heidegger understood this to be the role of philosophy.

[7] Kirschenbaum argues:

> I believe such trends will eventually affect the minutiae of academic policy. The English

> department where I teach, like most which offer the doctorate, requires students to demonstrate proficiency in at least one foreign language. Should a graduate student be allowed to substitute demonstrated proficiency in a *computer-programming language instead*? Such questions have recently arisen in my department and elsewhere; in my own case, almost a decade ago, I was granted permission to use the computer language Perl in lieu of proficiency in the second of two languages that my department required for the Ph.D. I successfully made the case that given my interest in the digital humanities, this was far more practical than revisiting my high-school Spanish. (Kirschenbaum 2009, emphasis added)

[8] This does not preclude other more revolutionary human-computer interfaces that are under development, including haptic interfaces, eye control interfaces, or even brain-wave controlled software interfaces.

[9] See http://www.thecomputationalturn.com/

[10] See the open digital humanities translation of Plato's *Protagoras* for a good example of a wiki-based project, http://openprotagoras.wikidot.com/

[11] Here I don't have the space to explore the possibilities of a transformation of the distinction between research and teaching by digital technologies, themselves a result of the Humboldtian notion of the university. We might consider that a new hybridized form of research-teaching or teaching-research might emerge, driven, in part, by the possibility of new knowledges being created and discovered within the teaching process itself. This would mean that the old distinctions of research as creative, and teaching as dissemination would have to change too.

References

Arnold, M. (2009) *Culture and Anarchy*. Oxford: Oxford University Press.

Baker, N. (1996) *The Size of Thoughts: Essays and Other Lumber*. New York: Random House.

Baker, N. (2001) *Double Fold: Libraries and the Assault on Paper*. New York: Random House.

Beckett, C. (2008) *Supermedia: Saving Journalism So It Can Save the World*. London: Wiley-Blackwell.

Benjamin, W. (1992) 'Theses on the Philosophy of History', in *Illuminations*, trans. H. Zohn. London: Fontana, 245-55.

Berry, D. M. (2011) *The Philosophy of Software: Code and Mediation in the Digital Age*. London: Palgrave Macmillan.

Clement, T., Steger, S., Unsworth, J., & Uszkalo, K. (2008) 'How Not to Read a Million Books', accessed 21 June 2010 http://www3.isrl.illinois.edu/~unsworth/hownot2read.html#sdendnote4sym

Fuller, M. (2006) 'Software Studies Workshop', accessed 13 April 2010
http://pzwart.wdka.hro.nl/mdr/Seminars2/softstudworkshop

Fuller, M. (2008) *Software Studies: A Lexicon*. Cambridge, MA: The MIT Press.

Fuller, S. (2006) *The New Sociological Imagination*. London: Sage.

Fuller, S. (2010) 'Humanity: The Always Already – or Never to be – Object of the Social Sciences?, in J. W. Bouwel (ed.), *The Social Sciences and Democracy*. London: Palgrave.

Hayles, N. K. (2011) 'How We Think: Transforming Power and Digital Technologies', in D. M. Berry (ed.), *Understanding the Digital Humanities*. London: Palgrave.

Heidegger, M. (1993) 'The Question Concerning Technology', in D. F. Krell (ed.), *Martin Heidegger: Basic Writings*. London: Routledge, 311-41.

Hofstadter, R. (1963) *Anti-Intellectualism in American Life*. USA: Vintage Books.

Kirschenbaum, M. (2009) Hello Worlds, *The Chronicle of Higher Education*, accessed 10 Dec. 2010, http://chronicle.com/article/Hello-Worlds/5476

Kuhn, T. S. (1996) *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.

Lakatos, I. (1980) *Methodology of Scientific Research Programmes*. Cambridge: Cambridge University Press.

Latour, B. (1986) 'Visualization and Cognition: Thinking with Eyes and Hands', *Knowledge and Society*, vol. 6, 1-40.

Latour, B. (2010) 'Tarde's Idea of Quantification', in M. Candea (ed.), *The Social After Gabriel Tarde: Debates and Assessments*. London: Routledge.

Lazer, D. *et al*. (2009) 'Computational Social Science', *Science*. Vol. 323, issue 5915 (6 February): 721-723.

Levy, P. (1999) *Collective Intelligence*. London: Perseus.

Liu, A. (2011) Where is Cultural Criticism in the Digital Humanities?, accessed 15 Dec. 2011 http://liu.english.ucsb.edu/where-is-cultural-criticism-in-the-digital-humanities/

Manovich, L. (2008) *Software takes Command,* accessed 3 May 2010 http://lab.softwarestudies.com/2008/11/softbook.html

Manovich, L. & Douglas, J. (2009) 'Visualizing Temporal Patterns In Visual Media: Computer Graphics as a Research Method'. Accessed 10 October 2009 http://softwarestudies.com/cultural_analytics/visualizing_temporal_patterns.pdf

McCarty, W. (2009) 'Attending from and to the Machine, accessed 18 September 2010 http://staff.cch.kcl.ac.uk/~wmccarty/essays/McCarty,%20Inaugural.pdf

Montfort, N. & Bogost, I. (2009) *Racing the Beam: The Atari Video Computer System*. Cambridge, MA: MIT Press.

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

Nature (2007) 'A matter of trust', *Nature.* 449 (11 October), 637-638.

Newman, J. H. (1996) *The Idea of a University*. Yale: Yale University Press.

Presner, T. (2010) 'Digital Humanities 2.0: A Report on Knowledge', accessed 15 October 2010 http://cnx.org/content/m34246/1.6/?format=pdf

Readings, B. (1996) *The University in Ruins*. Cambridge, MA: Harvard University Press.

Schnapp, J. & Presner, P. (2009) 'Digital Humanities Manifesto 2.0', accessed 14 October 2010 http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Schreibman, S., Siemans, R., & Unsworth, J. (2008) *A Companion to Digital Humanities*. London: Wiley-Blackwell.

Sellars, W. (1962) 'Philosophy and the Scientific Image of Man'. In: R. Colodny (ed.), *Frontiers of Science and Philosophy*. Pittsburgh: University of Pittsburgh Press, 35-78.

Sterling, B. (2010) 'Atemporality for the Creative Artist', *Wired*, accessed 1 July 2010 http://www.wired.com/beyond_the_beyond/2010/02/atemporality-for-the-creative-artist/

Stickney, D. (2008) Charticle Fever, *American Journalism Review*, accessed 18 March 2010 http://www.ajr.org/Article.asp?id=4608

Terras, M. (2010) 'Present, Not Voting: Digital Humanities in the Panopticon', accessed 10 July 2010 http://melissaterras.blogspot.com/2010/07/dh2010-plenary-present-not-voting.html

Thomson, I. (2003) 'Heidegger and the Politics of the University', *Journal of the History of Philosophy*. Vol. 41, no. 4: 515–542.

Thomson, I. (2009) 'Understanding Technology Ontotheologically, Or: the Danger and the Promise of Heidegger, an American Perspective', in Olsen, J. K. B, Selinger, E., and Riis, S. (eds.) *New Waves in Philosophy of Technology*. London: Palgrave.

Turing, A. M. (1950) 'Computing Machinery and Intelligence', *Mind*, Oct. 1950: 433-60.

Lev Manovich

Trending: The Promises and the Challenges of Big Social Data

What is “big social data”? The following description from the 2011 grant competition organized by a number of research agencies (including National Endowment for Humanities and National Science Foundation) in USA, Canada, UK, and Netherlands provides an excellent description:

“The idea behind the Digging into Data Challenge is to address how "big data" changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials used by scholars in the humanities and social sciences -- ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data or cell phone records -- what new, computationally-based research methods might we apply? As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials.”
www.diggingintodata.org (accessed March 31, 2011).

In this article I want to address some of the theoretical and practical issues raised by emerging “big data”-driven social science and humanities. My observations are based on my own experience over last three years with big data projects carried out in my lab at UCSD and Calit2 (softwarestudies.com). The issues which we will discuss include the differences between “deep data” about a few and “surface data” about the many; getting access to transactional data; and the new “data analysis divide” between data experts and the rest of us.

The emergence of social media in the middle of 2000s created opportunities to study social and cultural processes and dynamics in new ways. For the first time, we can follow imagination, opinions, ideas, and feelings of hundreds of millions of people. We can see the images and the videos they create and comment on, monitor the conversations they are engaged in, read their blog posts and tweets, navigate their maps, listen to their track lists, and follow their trajectories in physical space. And we don’t need to ask their permission to do this, since they themselves encourage us to do by making all these data public.

In the 20th century, the study of the social and the cultural relied on two types of data: “surface data” about the many (sociology, economics, political science) and “deep data” about a few (psychology, psychoanalysis, anthropology, ethnography, art history; methods such as “thick description” and “close reading”). For example, a sociologist

worked with census data that covered most of the country's citizens. However, this data was collected only every 10 year and it represented each individual only on a "macro" level, living out her/his opinions, feelings, tastes, moods, and motivations (see http://www.census.gov). In contrast, a psychologist would be engaged with a single patient for years, tracking and interpreting exactly the kind of data which census did not capture.

In the middle between these two methodologies of "surface data" and "deep data" were statistics and the concept of sampling. By carefully choosing her sample, a researcher could expand certain types of data about the few into the knowledge about the many. For example, starting in 1950s, Nielsen Company collected TV viewing data in a sample of American homes (via diaries and special devices connected to TV sets in 25,000 homes), and then used this sample data to predict TV ratings for the whole country (i.e. percentages of the population which watched particular shows). But the use of samples to learn about larger populations had many limitations.

For instance, in the example of Nelson's TV ratings, the small sample used did not tell us anything about the actual hour by hour, day to day patterns of TV viewing of every individual or every family outside of this sample. Maybe certain people watched only news the whole day; others only tuned in to concerts; others had TV on not never paid attention to it; still others happen to prefer the shows which got very low ratings by the sample group; and so on. The sample stats could not tell us anything about this. It was also possible that a particular TV program would get zero shares because nobody in the sample audience happened to watch it – and in fact, this happened more than once.

Think of what happens then you take a low-res image and make it many times bigger. For example, lets say you stat with 10x10 pixel image (100 pixels in total) and resize it to 1000x1000 (one million pixels in total). You don't get any new details – only larger pixels. This is exactly what happens when you use a small sample to predict the behavior of a much larger population. A "pixel" which originally represented one person comes to represent 1000 people who all assumed to behave in exactly the same way.

The rise of social media along with the progress in computational tools that can process massive amounts of data makes possible a fundamentally new approach for the study of human beings and society. *We no longer have to choose between data size and data depth*. We can study exact trajectories formed by billions of cultural expressions, experiences, texts, and links. The detailed knowledge and insights that before can only be reached about a few can now be reached about many – very, very many.

In 2007, Bruno Latour summarized these developments as follows: "The precise forces that mould our subjectivities and the precise characters that furnish our imaginations are all open to inquiries by the social sciences. It is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable." (Bruno Latour, Beware, your imagination leaves digital traces, Times Higher Education Literary Supplement, April 6, 2007.)

Two years earlier, in 2005, Nathan Eagle at MIT Media Lab already was thinking along the similar lines. He and his advisor Alex Pentland put up a web site called "reality mining" (reality.media.mit.edu) and wrote how the new possibilities of capturing details of peoples' daily behavior and communication via mobile phones can create Sociology in the 21st century. To put this idea into practice, they distributed Nokia phones with special software to 100 MIT students who then used these phones for 9 months – which generated approximately 60 years of "continuous data on daily human behavior."

Now, think of Google search. Google's algorithms analyze text on all web pages they can find, plus "PDF, Word documents, Excel spreadsheets, Flash SWF, plain text files," and, since 2009, Facebook and Twitter content. (More details: en.wikipedia.org/wiki/Google_Search). Currently Google does not offer any product that would allow a user to analyze patterns in all this data the way Google Trends does with search queries and Google's Ngram Viewer does with digitized books – but it is certainly technologically conceivable. Imagine being able to study the collective intellectual space of the whole planet, seeing how ideas emerge and diffuse, burst and die, how they get linked together, and so on – across the data set estimated to contain at least 14.55 billion pages (as of March 31, 2011; see worldwidewebsize.com).

Does all this sounds exiting? It certainly does. However, what maybe wrong with these arguments? Quite a few things.

I am going to discuss four objections to the arguments which I made in the first part: the collapse of deep data / surface data divide, and the new possibilities this offers for social and cultural research.

1. Only social media companies have access to really large (surface-wise) and deep (detailed) social data (especially transactional data). So if you an anthropologist who is working for Facebook, or a sociologist working for Google, you are lucky – but the rest of us are not.

Sure, you can get some of this data through APIs provided by most social media services and largest media online retailers (YouTube, Flickr, Amazon, etc.) – but these APIs do not give all data which these companies themselves are capturing about the users. Still, you can certainly do very interesting new cultural and social research by collecting data via APIs and then analyzing it – if you are good at programming and advanced statistics. Although APIs themselves are not complicated, all large-scale research projects which use the data with these APIs I have seen so far have been done by people in computer science. A good way to follow the work in this area is to look at papers presented at yearly WWW conferences. Recent papers (2009-2010) investigated how information spreads on Twitter (data: 100 million tweets), what

qualities are shared by most favored photos on Flickr (data: 2.2 million photos), and what geotagged Flickr photos tell us about people's attention (data: 35 million photos). (See www2009.org and www2010.org/www/).

It worth pointing out that even researchers working inside largest social media companies can't simply access all the data collected by different services a company. Some time ago I went to a talk by a researcher from Sprint (one of the largest US phone companies) who was analyzing the relations between geographic addresses of phone users and how frequently they called other people. He did have access to this data for all Sprint customers (around 50 million.) However, when he was asked why he did not used other data Spring collects such as instant messages and apps use, he explained that these services are operated by a different part of the company, and that the laws prohibit employees to have access to all of this data together. He pointed out that like any other company, Spring does not want to get into lawsuits for breach of privacy, pay huge fines and damage their brand image, and therefore they are super careful in terms of who gets to look at what data. You don't have to believe this, but I do. For example, do you think Google enjoys all the lawsuits about Street View? If you were running a business, would you risk losing hundreds of millions of dollars and badly damaging your company image? (Of course, Facebook may be one big exception to this.)

2. Analysis of social data will only produce reliable results if social actors are authentic in their online self-created expressions (what they post to Facebook, Twitter, blogs, etc.) and/or digital footprints they leave behind. Imagine that you wanted to study cultural imagination of people in Russia in the second part of 1930s and you only looked at newspapers, books, films, and other cultural texts - which of course all went through government censors before being approved for publication. You would conclude that indeed everybody in Russia loved Lenin and Stalin, was very happy, and was ready to sacrifice his/her life to build communism. You may say that this is unfair comparison, and it would be more appropriate to look instead at people's diaries. Yes, indeed it would be better – however if you were living in Russia in that period, and you knew that any night a black car may stop in front of you house and you would be taken away and probably shot soon thereafter, would you really commit all your true thoughts about Stalin and government to paper? Famous poet Osip Mandelstam wrote a short poem that criticized Stalin only indirectly without even naming him – and he paid for this with his life.

Today, if you live in a pretty large part of the world, you know that the government is likely to scan your electronic communications systematically (See en.wikipedia.org/wiki/Internet_censorship_by_country). In some of the countries, it also may arrest you simply for visiting a wrong web site. In these countries, you will be careful in what you are saying online. Some of us live in other countries where a statement against the government does not automatically put you in prison, and therefore people feel they can be more open. In other words, it does not matter if the government is tracking us not; what is important is what it can do with this information. (I grew up in Soviet Union in 1970s and then moved to the US, and believe me, in this

respect the difference between the two is huge. In USSR, we never made any political jokes on the phone, and only discussed politics with close friends at home.)

OK, so lets say we live in a country where we are highly unlikely to be prosecuted for occasional anti-government remarks. But still, how authentic are all the rest of our online expressions? As Ervin Goffman and other sociologists pointed out a long time ago, people always construct their public presence, carefully shaping how they present themselves to others – and social media is certainly not an exception to this. The degree of this "constructability" varies. For instance, most of us tend to do less self-censorship and editing on FB than in the profiles on dating sites, or in a job interview. Others carefully curate their profile pictures to construct an image they want to project. (If you scan your friends FB profile pictures, you are likely to find a big range). But just as we do in all other areas of our everyday life, we exercise some control all the time when we are online – what we say, what we upload, what we show as our interests, etc.

Again, this does not mean that we can't do interesting research by analyzing larger numbers of tweets, Facebook photos, YouTube videos, etc. – we just have to keep in mind that what all this data is not a transparent window into peoples' imaginations, intentions, motifs, opinions, and ideas. Its more appropriate to think of it as an interface people present to the world – i.e., a particular view which shows only some of the data of their actual lives and imaginations and which may also include other fictional data designed to project a particular image.

3. Is it really true that "We no longer have to choose between data size and data depth" as I stated? Yes and no. Imagine this hypothetical scenario. On the one side, we have ethnographers who are spending years inside a particular community. On another side, we have computer scientists who never meet people in this community but have access to their social media and digital footprints - daily spatial trajectories captured with GPS, all video recorded by surveillance cameras, online and offline conversations, uploaded media, comments, likes, etc. According to my earlier argument, both parties have "deep data" – but the advantage of computer science team is that they can capture this data about hundreds of millions of people as opposed to only small community. However, you may disagree and argue that actually only ethnographers have the "deep" data, while computer scientists have only "shallow" data. No matter how good are their data analysis ideas and algorithms, they will never arrive at the same insights and understanding of people and dynamics in the community, as ethnographers can.

Both arguments look convincing – and both maybe incorrect. It may be more accurate to say that ethnographers and computer scientists have access to *different kinds of data*. Therefore they are likely to ask different questions, notice different patterns, and arrive at different insights.

This does not mean that the new data "surface" is somehow less "deep" than the data obtained through long-term personal contact. In terms of the sheer number of "data

points," it is likely to be much deeper. However, many of these data points are quite different than the data points available to ethnographers.

For instance, if you physically present in some situation, you may notice some things which you would not notice if you watching a high-res video of the same situation. But at the same time, if you do computer analysis of this video you may find patterns you would not notice if you were on the scene physically only. (Of course, people keep coming with new techniques that combine on the scene physical presence and computer and network-assisted techniques. For an amazing example, see valleyofthekhans.org project at Calit2.)

These all sounds logical – but somehow, it goes against many of our intuitions. I can imagine people arguing that at least at present, even the most comprehensive social data about people which can be automatically captured via cameras, sensors, computer devices (phones, game consoles, etc.) can't be used to arrive at the same "deep" knowledge as having face to face interaction with these people over long periods of time.

I often encounter similar objections when I lecture about cultural analytics research in my own lab. We use digital image analysis and high-res visualization to explore cultural patterns in large sets of images and video – YouTube video, visual art, magazine covers and pages, graphic design, photographs, etc. One of the typical responses to my lectures is that computers can't lead to the same nuanced interpretation as traditional humanities methods and that they can't help understand deep meanings of artworks. My response is that we don't want to replace human experts with computers. As I will describe in the hypothetical scenario of working with one million YouTube documentary-style videos below, we can use computers to map the patterns in massive visual data sets and to select the objects that we then examine manually. While computer-assisted examination of massive cultural data sets typically reveals new patterns in this data which even best manual "close reading" would miss – and of course, even an army of humanists will not be able to carefully "close read" massive data sets in the first place – a human is still needed to make sense of these patterns.

Ultimately, completely automatic analysis of social and cultural data will not produce meaningful results today because computers' ability to understand the content texts, images, video and other media is still limited. (Recall the mistakes made by IBM Watson artificial intelligence computer when it competed on the TV quiz show Jeopardy! in early 2011).

Ideally, we want to combine human ability to understand and interpret - which computers can't completely match yet - and computers' ability to analyze massive data sets using algorithms we create. Lets say, for example, that you want to study documentary-type YouTube videos created by users in country X during the period Y, and you were able to determine that the relevant data set contains 1 million videos. So what do you do next? Computational analysis would be perfect as the next step to map the overall "data landscape": identify most typical and most unique videos,

automatically cluster all videos into a number of categories; find all videos that follow the same strategies, etc. At the end of this analytical stage, you may be able to reduce the set of one million videos to 100 videos which represent it in a more comprehensive way than if you simply used a standard sampling procedure. For instance, you reduced set may contain both most typical and most unique videos in various categories. Now that you have a manageable number of videos, you can actually start watching them. If you find some video to be particularly interesting, you can then ask computer to fetch more videos which have similar characteristics, so you can look at all of them. At any point in the analysis, you can go back and forth between particular videos, groups of videos and the whole collection of one million videos, experimenting with new categories and groupings. And just as Google analytics allows you to select any subset of data and look at its patterns over time (number of viewed pages) and space (where do visitors come from), you will be able to select any subset of the videos and see various patterns across this subsets.

This is my vision of how we can study large cultural data sets – whether these are billions of videos on YouTube or billions of photos on Flickr, or smaller samples of semi-professional or professional creative productions such as 100 million images on deviantart.com, or 250,000 design portfolios on coroflot.com. Since 2007, our lab has been working on visualization methods which would enable such research methodology and which would be particularly suitable for visual data. (For examples, visit cultural analytics page at softwarestudies.com, and in particular one-million manga pages project.)

4. Lets say a user has software that combines large-scale automatic data analysis and interactive visualization (we are working to integrate various tools which we designed in our lab to create such a system). If user has also skills to examine individual artifacts and the openness to ask new questions, the software will help her/him to take research in many new exiting directions. However, there are also many kinds of interesting questions that require expertise in computer science, statistics, and data mining – something which social and humanities researchers typically don't have.

The explosion of data and the emergence of computational data analysis as the key scientific and economic approach in contemporary societies create the new kinds of divisions. In big data society, people and organizations are divided into three categories: 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. The first group includes pretty much everybody in the world who is using the web and/or mobile phones; the second group is smaller; and the third group is much smaller still.

At Google, computer scientists are working on the algorithms that scan a web page a user is on currently and select which ads to display. At YouTube, computer scientist's work on algorithms that automatically show a list of other videos deemed to be relevant to one you are currently watching. At BlogPulse, computer scientists work on algorithms that allow companies to use sentiment analysis to study the feelings that millions of

people express about their products in blog posts. At certain Hollywood movie studios, computer scientists work on algorithms that predict popularity of forthcoming movies by analyzing tweets about them (it works). In each case, the data and the algorithms can also reveal really interesting things about human cultural behavior in general – but this is not what the companies who are employing these computer scientists are interested in. Instead, the analytics are used for specific business ends. (For more examples, see What People Want (and How to Predict It)).

So what about the rest of us? Today we are given a variety of sophisticated and free software tools to select the content of interest to us from this massive and constantly expanding universe of professional media offerings and user-generated media. These tools include search engines, RSS feeds, and recommendation systems. But while they can help you find what to read, view, listen to, play, remix, share, comment on, and contribute to, in general they are not designed for carrying systematic social and cultural research along the lines of "cultural analytics" scenario I described earlier.

While a number of free data analysis and visualization tools have become available on the web during last few years (Many Eyes, Tableau, Google docs, etc.), they are not useful unless you have access to large social datasets. Some commercial web tools allow anybody to analyze certain kinds of trends in certain data sets they are coupled with in some limited ways (or at least, they wet our appetites by showing what is possible). I am thinking of already mentioned Google Ngram Viewer, Trends, Insights for Search, Blogpulse, and also YouTube Trends Dashboard, Social Radar, Klout**.** (Searching for "social media analytics" or "twitter analytics" brings up lists of dozens of other tools.)

For example, Google Ngram Viewer plots relative frequencies of words or phrases you input across a few million English language books published over last 400 years and digitized by Google (data sets in other languages are also available). You can use it to reveal all kinds of interesting cultural patterns. Here are some of my favorite combinations of words and phrases to use as input: "data, knowledge"; "engineer, designer"; "industrial design, graphic design." In another example, YouTube Trends Dashboard allows you to compare most viewed videos across different geographic locations and age groups.

Still, what you can with these tools today is quite limited. One of the reasons for this is that companies make money by analyzing patterns in the data they collect about our online and physical behavior, and target their offerings, ads, sales events, and promotions accordingly; in other cases, they sell this data to other companies. Therefore they don't' want to give consumers direct access to all this data. (According to an estimate by ComScore, in the end of 2007 five large web companies were recording "at least 336 billion transmission events in a month.")

If a consumer wants to analyze patterns in the data which constitutes/reflects her/his economic relations with a company, here the situation is different. The companies often provide the consumers with professional level analysis of this data - financial activities

(for example, my bank web site shows a detailed breakdown of my spending categories), their web sites and blogs (Google Analytics), or their online ad campaigns (Google AdWords).

Another relevant trend in to let a user compare her/his data against the statistical summaries of data about others. For instance, Google Analytics shows the performance of my web site against all web sites of similar type, while many fitness devices and sites allow you to compare your performance against the summarized performance of other users. However, in each case, the companies do not open the actual data, but only provide the summaries.

Outside of commercial sphere, we do see a gradual opening up of the data collected by government agencies. For USA examples, check www.data.gov, HealthData.gov, and radar.oreilly.com/gov2/. As Alex Howard notes in Making Open Government Data Visualizations That Matter (3.13.2011), "Every month, more open government data is available online. Local governments are becoming data suppliers." Note, however, that this data is typically statistical summaries, as opposed to transactional data (the traces of people online behavior) or their media collected by social media companies.

The limited access to massive amounts of transactional social data that is being collected by companies is one of the reasons why today large contemporary data-driven social science and large contemporary data-driven humanities are not easy to do in practice. (For examples of digitized cultural archives available at the moment, see the list of repositories that agreed to make their data available to Digging Into Data competitors.) Another key reason is the large gap between what can be done with the right software tools, right data, and no knowledge of computer science and advanced statistics - and what can only be done if you do have this knowledge.

For example, lets imagine that you were given full access to the digitized books used in Ngram Viewer (or maybe you created your own large data set by assembling texts from Project Guttenberg, or another source) and you want software to construct graphs which show changing frequencies of topics over time, as opposed to individual words. If you want to do this, you better have knowledge of text mining or computational linguistics. (A search for "topic analysis" on Google Scholar returned over 2000 articles in these fields.)

Or lets say that you interested in how social media facilitates information diffusion, and you want to use Twitter data for your study. In this case, you can obtain the data using Twitter API, or third party services that collect this data and make it available for free or for a fee. But again, you better have the right background to make use of this data. The software itself is free and readily available – R, Weka, Gate, etc. - but you need to have the right training (at least some classes in computer science and statistics) and practical experience to get meaningful results.

Here is an example of what can be done by people with the right expertise. In 2010 four researchers from Computer Science department at KAIST (South Korea's leading

university for technology) published a paper entitled What is Twitter, a social network or a news media? Using Twitter API, they were able to study the entire Twittersphere as of 2009: 41.7 million user profiles, 1.47 billion social relations, 106 million tweets. Among their discoveries: over 85% of trending topics are “headline news or persistent news in nature.” (For more examples of the analysis of “social flows”, check the papers presented at IEEE International Conference on Social Computing 2010.)

So where does this all leaves us? Is it true that “surface is the new depth” – in a sense that the quantities of “deep” data that in the past was obtainable about a few can now be automatically obtained about many? Theoretically, the answer is yes (as long as we keep in mind that the two kinds of deep data have different content.) Practically, there are a number of obstacles before this can become a reality. I tried to describe a few of these obstacles, but there are also others I did not analyze. (Obviously, a very big one is privacy – would you trust academic researchers to have all your communication and behavior data automatically captured?) However, with what we already can use today (social media companies APIs, Infochimps.com data marketplace and data commons, Gnip social media aggregator, free archives such Project Guttenberg, Internet Archive, etc.), the possibilities are endless – if you know programming and data analytics, and also are open to asking new questions about human beings, their social life and their cultural expressions and experiences.

I have no doubt that eventually we will see many more humanities and social science researchers who are equally good at most abstract theoretical arguments as well the latest data analysis algorithms which they can implement themselves, as opposed to relying on computer scientists. And one of the trends, which will get us there is digital humanities. Put this phrase into Google Insights for Search and you will see a line that is systematically keep going up.

Digging Into Data organized by NEH Office of Digital Humanities are particularly important in this respect. They publicize the idea of “big data”-driven research, and they also fund actual projects, which bring together humanists, social scientists and computer science.

As 2011 competition announcement says, “Let's get digging.”

---

* I am grateful to UCSD faculty member James Fowler for an inspiring conversation a few years ago about the collapse of depth/surface distinction. See his work at jhfowler.ucsd.edu.

Tooling Up for Digital Humanities

Digitization

As archives increasingly go digital, what do humanities scholars need to know in order to optimize their use of these new resources? This section offers an introduction to online archives, focusing on digitization, text mark-up and metadata.

1: Making Documents Digital

As archives increasingly go digital, what do humanities scholars need to know in order to optimize their use of these new resources?

First and foremost, the format of those resources matters. A digital archive represents the work and decisions of many librarians and archivists, and understanding those decisions and the forms they take can help researchers access and manipulate data more effectively. Certain types of digital analysis can only be done on data that exists in certain formats, for instance. Although scholars need not understand the digital archiving process in enormous depth, they must understand on a basic level how a particular online archive is structured.

The first step in the archival process of transforming a source from an analog (non-digital) format to an electronic format usually takes the form of scanning the document in order to create a digitized image of the text (a PDF, for instance). This initial imprint is like taking a photograph, even if the imprint is of a page of text. A great deal of archival material is available in this “images of words” format. Although useful for someone who wants to read read the document, this image alone does little to help researchers find, access, or manipulate that text.

INDEPENDENT IN ALL THINGS—NEUTRAL IN NOTHING.

JEFFERSON, TEXAS, FRIDAY, JUNE 30, 1911

LETTERS FROM CAPT. C. T. TODD

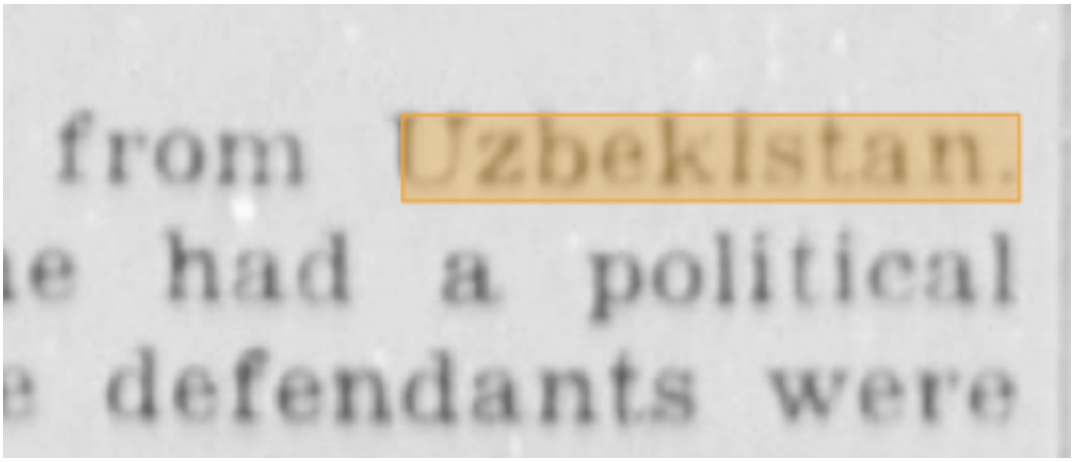
Campaign Notes.

The moral forces of Texas ought to make prohibition the chief concern every day till July 22. It is the one reform that has every other moral reform wrapped up in it as the saloon has every evil in it.

We must take nothing for granted. Normally Texas is at least 50,000 dry. But we must go after them and poll them. In every county and precinct a systematic campaign ought to be made to reach every voter, and see that he is not misled.

among the people, mainly in the country. These men are adepts in hipocracy, and skilled in all the dark and devious ways of the saloon. Many of them profess to be prohibitionists at heart;

Documents become more usable through a process called “extraction,” in which computer software performs Optical Character Recognition (OCR). This process scans the image of text and attempts to recognize characters and words, which it then stores as a separate layer of text based on the scanned image. This layer can be thought of as being superimposed on the image – while the human eye understands a picture of the word “Uzbekistan,” the computer uses the translated, invisible layer to recognize “Uzbekistan” as a series of characters that form a word.



This process is largely automated, although the software can be checked and supplemented by a human reader. A fully-automated process is commonly known as a “dirty OCR” because it will often have many typos and errors (letter combinations like “rn” might be transcribed as “m,” for instance.). The initial OCR text translation is usually around 70 or 80 percent accurate, and often needs human review. It is important to recognize that when performing searches within OCR’d documents the search is not necessarily comprehensive. If a word was not properly recognized or translated by the process, a computer search will not find it (even if a human reader will correctly recognize the word). From here, the image and “dirty OCR” can be reviewed by a human reader to improve its translation accuracy.

2: Metadata and Text Markup

Metadata comprises the “data about data” – information associated with archival material that lists key attributes, such as its author, date, publisher, or general subject. Metadata is far from new – the back of a book’s title page listing publication information is a straightforward example of metadata in non-digital form. Attaching this “data about data” to archival material is one of the most crucial steps in making that material findable and, consequently, usable by humanities scholars.

Metadata often falls under a broader process of text markup, whereby additional information is grafted onto the “raw” text of a document. There are different ways in which text can be marked up, but the most common in the archival world is Extensible Markup Language (XML). XML is a standardized set of rules for attaching information to text in order to make it readable by machines; like any language, it has its own syntax and conventions. XML works largely by wrapping chunks of text (words, sentences,

paragraphs, etc.) in tags that describe what is between them. Tags can also be nested within one another for greater flexibility. Take the below example:

```
<painting>
 <caption>This is Raphael's "Foligno" Madonna, painted in
 <date>1511</date>–<date>1512</date>.
 </caption>
</painting>
```

Everything between the “painting” opening tag (<painting>) and the “painting” closing tag (</painting>) has to do with, unsurprisingly, a painting. The <caption> open and close tags surround an enclosed text indicating it as a caption, and the <date> tags specify which words in the caption refer to dates.

An archive with well-refined markup content allows users to search for specific terms (author, title, subject) within and across many different documents. When you perform a search in an archival database it might return information based on the metadata contained in the headers of each text. In the Early Americas Digital Archive, for example, users can search by genre (prose, poetry, drama), format (chronicle, diary, etc.), mode (satire, pastoral, etc.), historical period (in 50-year intervals), geographic location (New England, New Spain, Virginia, etc.), among others. A user can search for “Georgic” “Poetry” about the “Caribbean” published “1750-1800.” These searches are built on the framework of metadata and text mark-up. The plain text of a document is inert: without marking it up, a computer would not be able to locate and retrieve this additional information.

Metadata and text mark-up has traditionally been generated by human labor in the form of decisions made by archivists about how to categorize and describe a source. These decisions are oftentimes more complex than they would initially appear: for instance, should Charlotte Bronte’s Jane Eyre be classified as a Bildungsroman or a late Gothic novel? Often older items will come with existing information and classifications, but they still require modifications. Glen Worthey, head of Stanford’s Humanities Digital Information Service, reminds us that “We interpret and map older forms of data into newer forms, but we not only need to map the old data but also put it into a common form so that the information works in a database.”

More specifically, in order for mark-up to be effective across institutions and archives there needs to be standards for how text should be marked up. One of the major groups working to develop and maintain standards for digitized texts is the Text Encoding Initiative (TEI). TEI is a consortium of academic, institutions, research groups, and individual scholars from around the world. Among the databases that use TEI standards are the Perseus Project, the Women Writers Project, the Early Americas Digital Archive, and the SWORD Project. The Online Archive of California (OAC) is another database consortium that coordinates metadata standards and provides free public access to detailed descriptions of primary resource collections maintained by more than 150 contributing institutions. In standardizing rules and syntax, initiatives such as TEI and OAC try to ensure that searches become interoperable from one system to the next.

It is important to remember that the storage of information is not neutral. As Dan Cohenargues, "Scholars who structure historical documents with markup languages such as XML make choices—often quite good choices, but choices none the less—about which elements of a document are most important." Many initiatives such as Google Books are attempting to automate the process in order to enhance millions of digitized sources. This leads to its own set of problems, as subtle distinctions can be lost. As Worthey says, "It's dangerous for a humanities scholar to entrust too much to a programmer or mathematician." The process of digitizing documents is not entirely a "sausage factory," however. With a basic understanding of how online archives are created and organized, scholars will have a better sense of what they're actually looking at and the quality of their sources. As more and more information goes digital, grasping the structure behind this information will become increasingly critical for scholarship and research.

3: Further Reading

Julie Meloni, "A Pleasant Little Chat About XML" in Profhacker, 6 October 2009, http://chronicle.com/blogPost/A-Pleasant-Little-Chat-abou/22746/

Dan Cohen, "Is Google Good for History?" http://www.dancohen.org/2010/01/07/is-google-good-for-history/

Early America's Digital Archive – Introduction. http://www.mith2.umd.edu/eada/intro.php

Interview with Glen Worthey. Stanford University. August 12, 2010.

III

Big Data Spaces

Queries
Mohebbi et. al. (2011). Google Correlate Whitepaper. http://correlate.googlelabs.com/whitepaper.pdf

Streams
Berry, David (2011). Philosophy of Software. London: Palgrave Macmillan. Excerpt.

Platforms
Caplan, Paul (2011). "Software Tunnels Through the Rags 'n Refuse: Object Oriented Software Studies and Platform Politics". Presented at Platform Politics conference in Cambridge, 13 May 2011. http://theinternationale.com/blog/2011/05/software-tunnels-through-the-rags-n-refuse/

Commentspaces
Shah and Yazdani nia (2011). "Politics 2.0 with Facebook – Collecting and Analyzing Public Comments on Facebook for Studying Political Discourses." The Journal of Information Technology and Politics Annual Conference. http://scholarworks.umass.edu/jitpc2011/3/

Fora
Bernstein et al (2011). "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community." Association for the Advancement of Artificial Intelligence. http://projects.csail.mit.edu/chanthropology/4chan.pdf

Location-based services
Cheng, Zhiyuan et. al (2011). "Exploring Millions of Footprints in Location Sharing Services." Fifth International AAAI Conference on Weblogs and Social Media. 17-21 July 2011, Barcelona, Spain. http://students.cse.tamu.edu/kyumin/papers/cheng11icwsm.pdf

Google Correlate Whitepaper

Matt Mohebbi, Dan Vanderkam, Julia Kodysh,

Rob Schonberger, Hyunyoung Choi & Sanjiv Kumar

Draft Date: June 9, 2011

Trends in online web search query data have been shown useful in providing models of real world phenomena. However, many of these results rely on the careful choice of queries that prior knowledge suggests should correspond with the phenomenon. Here, we present an online, automated method for query selection that does not require such prior knowledge. Instead, given a temporal or spatial pattern of interest, we determine which queries best mimic the data. These search queries can then serve to build an estimate of the true value of the phenomenon. We present the application of this method to produce accurate models of influenza activity and home refinance rate in the United States. We additionally show that spatial patterns in real world activity and temporal patterns in web search query activity can both surface interesting and useful correlations.

Background

Web search activity has previously been shown useful for providing estimates of real-world activity in a variety of contexts, with the most common being health and economics. Examples in health include influenza[1,2,3,4,5,6,9], acute diarrhea[6], chickenpox[6], listeria[7], and salmonella[8]. Examples in economics include movie box office sales[9], computer game sales[9], music billboard ranking[9], general retail sales[10], automotive sales[10], home sales[10], travel[10], investor attention[11], and initial claims for unemployment[12].

Modeling real-world activity using web search data can provide a number of benefits. First, it can be more timely, especially when the alternative is not electronically collected. Influenza surveillance from the United States Centers for Disease Control and Prevention (CDC), Influenza Sentinel Provider Surveillance Network (ILINet) has a delay of one to two weeks[1]. For economic indicators like unemployment, this delay is measured in months[10]. In contrast, search data can "predict the present" since it is available as the target activity happens[10]. Second, query data has good temporal and spatial resolution. If an indicator of interest is incomplete (missing time periods or regions, coarser temporal or spatial resolution, etc.), query data can sometimes be used to fill in the gaps. For example, influenza rate data from ILINet is only published by the CDC at the national and regional level and is not published for the off season[13], but models based on query data can be used to provide estimates year-round and at a state and sometimes even city level, provided there is sufficient search activity at that level[1,14,15]. Third, there can be considerable expenses incurred in collecting data for traditional indicators. Finally, while Internet users do not represent a random sample of the United States population, this population has become increasingly less biased over time and now represents 77% of the adult population[16]. In the 18-29 subgroup, this number is almost 90%. This is in contrast to traditional landline phone surveys which must either under-represent this age group or blend in cell-phone survey data at considerable difficulty and expense[17].

Three Google tools have been released previously to enable access to aggregated online web search query data. Google Trends and Google Insights for Search are both real-time systems which provide temporal and spatial activity for a given query. However, they are both unable to automatically surface queries which correspond with a particular pattern of activity. Google Flu Trends provides estimates of Influenza-like Illness (ILI) activity in the United States, using models based on query data. These queries are selected from millions of possible candidates through an automated process[1]. Due to the computational requirements of this process, a batch-based distributed computing framework[18] was employed to distribute the task across hundreds of machines.

Google Correlate builds on this previous work. Google Correlate is a generalization of Flu Trends that allows for automated query selection across millions of candidate queries for any temporal or spatial pattern of interest. Similar to Trends and Insights for Search, Google Correlate is an online system and can surface its results in real time.

Data Summary

Using anonymized logs of Google web search queries submitted from January 2003 to present, we computed two different databases for Google Correlate:

us-weekly: temporal only: weekly time series data for the United States at a national level.

us-states: spatial only: state-by-state series data for the United States summed across all time.

Each database contains tens of millions of queries. For additional details, please see the Data section below.

Methods Summary

The objective of Google Correlate is to surface the queries in the database whose spatial or temporal pattern is most highly correlated (R^2) with a target pattern. Google Correlate employs a novel approximate nearest neighbor (ANN) algorithm over millions of candidate queries in an online search tree to produce results similar to the batch-based approach employed by Google Flu Trends but in a fraction of a second. For additional details, please see the Methods section below.

Flu Trends

Google Flu Trends produces estimates of ILI activity in the United States using query data. The Flu Trends modeling process is composed of two steps: variable selection and model building. Google Correlate can perform the variable selection and provide the associated time series data as a CSV download to enable the construction of a model using the selected queries. In this section we provide a test of the quality and computational power of Google Correlate, demonstrating that this automated system can be used to build a new Flu Trends model for the United States with comparable performance, but in a fraction of the time used to build the original Flu Trends model.

The baseline for this comparison the original regional Google Flu Trends model[1]. For these models, query selection was performed on the regional level, and a single set of queries was chosen to optimize the results across all regions. The values of the query time series were summed into a single input variable per region, and a model was fitted from the data across all nine regions. This model was built using weekly training data between 9/28/2003 and 3/11/2007 inclusive, and evaluated by computing the correlation between the resulting predictive estimates and the corresponding regional weekly truth data over the holdout period between 3/18/2007

to 5/11/2008.

While we sought to make a close comparison between the results of the Google Flu Trends methodology and modeling of ILI activity using Google Correlate, there are several differences between the methods employed. First, we worked with a different resolution for query selection. Since Google Correlate provides only national query time series data, we can only perform query selection on the national level. After the national-level query selection, we sum the query time series into a single explanatory variable and fit a linear model to the nine census regions. Second, we used a different cross-validation technique for variable selection in Google Correlate from the one used in Flu Trends.

We used Google Correlate to perform query selection by uploading ILI activity data from the CDC over the training time period. This weekly time series is at the national level and represents the rate of ILI-related doctors office visits per 100,000 visits. We summed the time series of all 100 queries returned by Google Correlate into a single explanatory variable. We then fit a linear model to the nine census regions and generated regional estimates for the holdout time period.

Training window correlation (R^2)

	Mean	*Min*	*Max*
Google Flu Trends	*0.90*	*0.80*	*0.96*
Google Correlate	*0.87*	*0.70*	*0.97*

n = 9 regions

Holdout window correlation (R^2)

	Mean	*Min*	*Max*
Google Flu Trends	*0.97*	*0.92*	*0.99*
Google Correlate	*0.96*	*0.88*	*0.98*

n = 9 regions

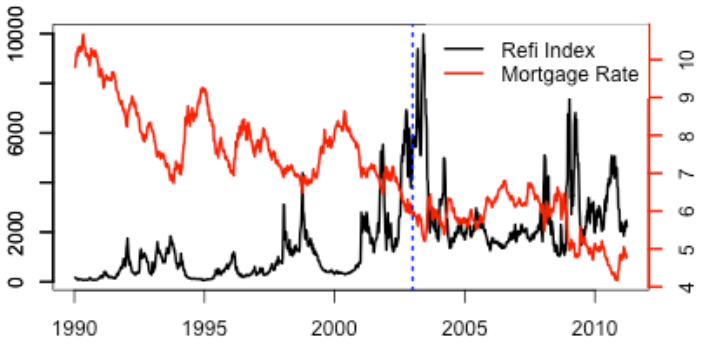
We see that the Google Correlate-based model slightly underperforms the Flu Trends model for the hold out time, with average correlation across all nine regions of 0.97 for Flu Trends and 0.96 for Correlate. This difference could be due, in part, to the difference in resolution of the query selection process. The time required to create the model with Google Correlate was a fraction of that required for the original Flu Trends model.

Refinance

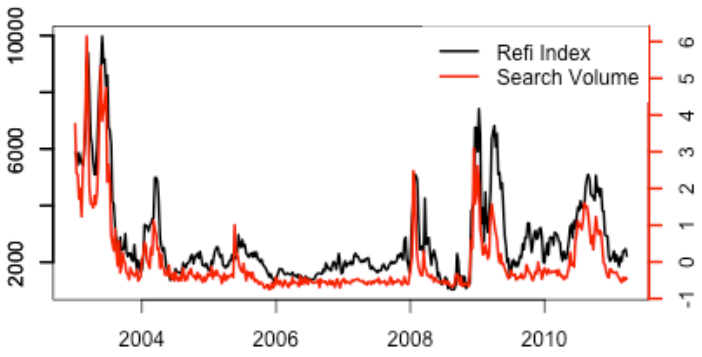
Every week, Mortgage Bankers Association of America (MBA) compiles all mortgage application to refinance an existing mortgage into a refinance index. The MBA's loan application survey covers more than half of all United States residential mortgage loan applications and is considered by many to be the best gauge of mortgage refinancing activity.

Consumers refinance a home for a number of reasons, including to switch to a lower mortgage interest rate, to change the mortgage length, to tap into their home equity and to switch mortgage type. In 2003, the refinancing activity peaked due to record low interest rate and the real estate boom. Despite the lower mortgage interest rate in 2010, the level of refinancing was not as high as in 2003 due to the housing recession and the subprime credit crisis.

We examined the top 100 most correlated queries with the refinance index time series from January 2003 to August 2010 and extended the window week by week until the end of March 2011. Fifty percent of the selected queries were refinance-related, including *refinancing calculator*, *refinancing closing costs*, and *refinance comparison*. Mortgage rate related queries such as *lowest mortgage rates* and *no cost mortgage* accounted for about 35% of queries selected. Even though queries for mortgage rates are related to refinancing, it is not always about refinancing and thus the signal could be mixed.



Refi Index vs. Mortgage Rate



Refi Index vs. Search Volume of *refinancing calculator*

Using these queries, we applied the same method from Choi and Varian[10] and compared two alternative models with baseline model with a moving window from August 2010 to March 2011. Let y_t be the time series of the refinance index, Refi_t be the summed query time series for queries returned by Google Correlate containing "refinance" or "refinancing", and Finance_t be the summed query time series for all 100 queries returned by Google Correlate.

Baseline Model: $y_t = \alpha + \phi y_{t-1} + e_t$
Alternative Model 1: $y_t = \alpha + \phi y_{t-1} + \beta \text{Refi}_t + e_t$
Alternative Model 2: $y_t = \alpha + \phi y_{t-1} + \beta \text{Finance}_t + e_t$

The model fit is significantly improved and prediction error is decreased for the two alternatives. The out of sample mean absolute error (MAE) with rolling window for the 31 weeks is decreased by 7.04% for Alternative Model 1 and the MAE for Alternative Model 2 is increased by 9.12%.

		ϕ	β	R^2	MAE
Baseline Model	Est	0.945		0.9011	8.42
	s.e.	0.015			
Alternative Model 1	Est	0.473	16.7	0.9363	7.61
	s.e.	0.033	1.1		
Alternative Model 2	Est	0.481	9.1	0.9366	7.82
	s.e.	0.032	0.6		

Ribosome

A ribosome is a component inside living cells. Using the *us-weekly* database, the query *ribosome* surfaces the following highly-correlated ($R^2 > 0.96$) queries:

1. *mitochondria*
2. *cell wall*
3. *chloroplasts*
4. *chromatin*
5. *plant cells*
6. *vacuole*
7. *chloroplast*
8. *nuclear membrane*
9. *reticulum*
10. *cell function*

The time series for these queries feature upticks in the Fall and Spring, sharp drops during Thanksgiving and Christmas and a long trough in the summer. This mirrors the school year in the United States and suggests that the queries are being driven by biology classes.

It is worth noting that all of these top terms relate to biology. Other school topics (e.g. the Canterbury Tales) are also studied early in the school semester and yet this time series is not correlate nearly as well. It's both surprising and impressive that the phenomenon of biology study appears to be uniquely characterized by its temporal pattern. This can be seen with other queries, for example *eigenvector*, but to a smaller extent.

Latitude

Using a *us-states* data series containing the latitude for each state in the United States, we find the following highly-correlated queries were surfaced ($R^2 > 0.84$):

1. *sad light therapy*
2. *defroster*
3. *seasonal affective disorder lights*
4. *10000 lux*
5. *sun lamp*
6. *track length*
7. *floor heating*
8. *fleece hat*
9. *irish water spaniel*
10. *hydronic*

The "sad" in *sad light therapy* is likely the acronym for seasonal affective disorder, which also seems to describe the relationship between queries *sad light therapy*, *seasonal affective disorder lights*, *10000 lux* and *sun lamp*. These top results surfaced by Google Correlate imply that latitude in the United States can be modeled using the spatial patterns in SAD-related queries. This is consistent with studies on the correlation of SAD prevalence and latitude in North America[19].

Disclaimers

This system is not intended to serve as a replacement for traditional data collection mechanisms. While the queries selected by Google Correlate for a specific target series exhibit strong correlations with the target series over many years, this correspondence may not hold in the future due to changes in user behavior which are unrelated to the target behavior. For example, the correlation of a drug whose time series historically tracked well the activity of a disease, could significantly be changed by a recall of the drug.

Additionally, the underlying cause of search behavior can never be known. Users submitting influenza-like illness (ILI) queries are not necessarily experiencing ILI-symptoms. And similarly, non-ILI related queries which are highly correlated with an ILI series do not necessarily increase or decrease the likelihood of contracting influenza.

Query data does not represent a random sample of the population. While over three quarters of United States adults use the Internet, several subgroups are underrepresented. This could lead to sampling error depending on the modeling performed.

Google Correlate requires indicators with unique spatial or temporal patterns. Indicators with little variation or with very regular variation are unlikely to surface meaningful results. Indicators with unique variation may still not surface results due to a lack of information-seeking behavior for the indicator.

Acknowledgements

The authors would like to thank Doug Beeferman and Jeremy Ginsberg for providing early inspiration for Google Correlate. We'd also like to thank Hal Varian for his valuable feedback on Google Correlate and Jean-Baptiste Michel for his useful comments on this manuscript. Finally, we'd like to thank Craig

Nevill-Manning and Corinna Cortes for their guidance and support.

Privacy

At Google, we recognize that privacy is important. None of the data in Google Correlate can be associated with a particular individual. The data contains no information about the identity, IP address, or specific physical location of any user. Furthermore, any original web search logs older than nine months are anonymized in accordance with Google's Privacy Policy[20].

Data

Google Correlate contains two different databases of Google web search queries. The first contains contains weekly time series for the United States at a national resolution (*us-weekly*). The second contains state-by-state series for the United States summed across all time (*us-states*). Both datasets are one-dimensional, with *us-weekly* having a time dimension but no space dimension and *us-states* having a space dimension but no time dimension. Both dataset contain tens of millions of series.

To help smooth query data across similar underlying user behavior, n-grams of the queries are used as series identifiers. This approach is similar to Google Trends and Insights for Search but is in contrast to Flu Trends where only lowercasing was performed on the queries.

The following example illustrates how n-grams are extracted from the query 'cold and flu symptoms'.

cold *
cold and
cold and flu *
cold and flu symptoms *
and *
and flu
and flu symptoms
flu
flu symptoms *
symptoms *

This list is filtered to contain only n-grams which appear often and in many states. The n-grams marked with an asterisk are kept when this filter is applied using the *us-weekly* dataset. Each of these filtered n-grams has a corresponding time series stored in the database, and for each instance of 'cold and flu symptoms' in the web search logs, each resulting n-gram receives a count. Filtering is done for privacy reasons but since rare queries are sporadic in nature, they are unlikely to be useful for modeling of long term phenomena. Distracting queries such as misspellings and those containing adult sexual content are also excluded.

The series in both datasets are normalized by dividing by the total count for all queries in that week (us-weekly) or state (*us-states*). The normalization controls for the year over year growth in all Internet search use (us-weekly) and state-by-state variation in Internet usage (*us-states*). Finally, each time series is standardized to have a mean value of zero and a variance of one, so that queries can be easily compared.

Methods

In our Approximate Nearest Neighbor (ANN) system, we achieve a good balance of precision and speed by using a two-pass hash-based system. In the first pass, we compute an approximate distance from the target series to a hash of each series in our database. In the second pass, we compute the exact distance function on the top results returned from the first pass.

Each query is described as a series in a high-dimensional space. For instance, for us-weekly, we use normalized weekly counts from January 2003 to present to represent each query in a 400+ dimensional space. For *us-states*, each query is represented as a 51-dimensional vector (50 states and the District of Columbia). Since the number of queries in the database is in the tens of millions, computing the exact correlation between the target series and each database series is costly. To make search feasible at a large scale, we employ an ANN system that allows fast and efficient search in high-dimensional spaces.

Traditional tree-based nearest neighbors search methods are not appropriate for Google Correlate due to the high dimensionality which results in sparsenes. Most of these methods reduce to brute force linear search with such data. For Google Correlate, we used a novel asymmetric hashing technique which uses the concept of projected quantization[21] to reduce the search complexity. The core idea behind projected quantization is to exploit the clustered nature of the data, typically observed with various real-world applications. At the training time, the database query series are projected in to a set of lower dimensional spaces.

Each set of projections is further quantized using a clustering method such as K-means. K-means is appropriate when the distance between two series is given by Euclidean distance. Since Pearson correlation can be easily converted into Euclidean distance by normalizing each series to be a standard Gaussian (mean of zero, variance of one) followed by a simple scaling (for details, see appendix), K-means clustering gives good quantization performance with the Google Correlate data. Next, each series in the database is represented by the center of the corresponding cluster.

This gives a very compact representation of the query series. For instance, if 256 clusters are generated, each query series can be represented via a unique ID from 0 to 255. This requires only 8 bits to represent a vector. This process is repeated for each set of projections. In the above example, if there are m sets of projections, it yields an 8m bit representation for each vector.

During the online search, given the target series, the most correlated database series are retrieved by asymmetric matching. The key concept in asymmetric matching is that the target query is not quantized but kept as the original series. It is compared against the quantized version of each database series. For instance, in our example, each database series is represented as an 8m bit code. While matching, this code is expanded by replacing each of the 8 bits by the corresponding K-means center obtained at training time, and Euclidean distance is computed between the target series and the expanded database series. The sum of the Euclidean distances between the target series and the database series in m subspaces represents the approximate distance between the two. Approximate distance between target series and the database series is used to rank all the database series. Since the number of centers is usually small, matching of the target series against all the database series can be done very quickly.

To further improve the precision, we take the top one thousand series from the database returned by our approximate search system (the first pass) and reorder those by doing exact correlation computation (the second pass). By combining asymmetric hashes and reordering, the system is able to achieve more than 99% precision for the top result at about 100 requests per second on O(100) machines, which is orders of magnitude faster than exact search.

References

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012-1014.

2. Eysenbach G (2006) Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc: 244-248.

3. Hulth A, Rydevik G, Linde A (2009) Web queries as a source for syndromic surveillance. PLoS One 4: e4378-e4378.

4. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, et al. (2004) Analysis of Web access logs for surveillance of influenza. Stud Health Technol Inform 107: 1202-1206.

5. Polgreen PM, Chen Y, Pennock DM, Nelson FD (2008) Using internet searches for influenza surveillance. Clin Infect Dis 47: 1443-1448.

6. Pelat C, Turbelin Cm, Bar-Hen A, Flahault A, Valleron A-J (2009) More diseases tracked by using Google Trends. Emerg Infect Dis 15: 1327-1328.

7. http://ecmaj.ca/cgi/content/full/180/8/829

8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/

9. http://www.pnas.org/content/107/41/17486.full.pdf

10. http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf

11. http://www.nd.edu/~zda/Google.pdf

12. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/papers/initialclaimsUS.pdf

13. http://www.cdc.gov/flu/weekly

14. http://www.eht-journal.net/index.php/ehtj/article/view/7183/8094

15. http://www.nature.com/nature/journal/v457/n7232/extref/nature07634-s1.pdf

16. http://www.pewinternet.org/Static-Pages/Trend-Data/Whos-Online.aspx

17. http://pewresearch.org/pubs/515/polling-cell-only-problem

18. Dean, J. & Ghemawat, S. Mapreduce: Simplified data processing on large clusters. OSDI: Sixth Symposium on Operating System Design and Implementation (2004)

19. http://cbn.eldoc.ub.rug.nl/FILES/root/1999/JAffectDisordMersch/1999JAffectDisordMersch.pdf

20. http://www.google.com/privacypolicy.html

21. A. Gersho and R. M. Gray, Vector Quantization and Signal Compression, Springer, 1991.

6
Real-Time Streams

The growth of the Internet has been astonishing, both in terms of its breadth of geographic cover, but also the staggering number of digital objects that have been made to populate the various webpages, databases, and archives that run on the servers. This has traditionally been a rather static affair, however, there is evidence that we are beginning to see a change in the way in which we use the web, and also how the web uses us. This is known as the growth of the so-called 'real-time web' and represents the introduction of a technical system that operates in real-time in terms of multiple sources of data fed through millions of data streams into computers, mobiles, and technical devices more generally. Utilising Web 2.0 technologies, and the mobility of new technical devices and their locative functionality, they can provide useful data to the user on the move. Additionally, these devices are not mere 'consumers' of the data provided, they also generate data themselves, about their location, their status and their usage. Further, they provide data on data, sending this back to servers on private data stream channels to be aggregated and analysed. That is,

1. The web is transitioning from mere interactivity to a more dynamic, real-time web where read-write functions are heading towards balanced synchronicity. The real-time web… is the next logical step in the Internet's evolution.
2. The complete disaggregation of the web in parallel with the slow decline of the destination web.
3. More and more people are publishing more and more "social objects" and sharing them online. That data deluge is creating a new kind of search opportunity (Malik 2009).

The way we have traditionally thought about the Internet has been in terms of pages, but we are about to see this changing to the concept of 'streams'. In essence, the change represents a move from a notion of *information retrieval*, where a user would attend to a particular machine to extract data as and when it was required, to an *ecology of data streams* that forms an intensive information-rich computational environment. This notion of living within streams of data is predicated on the use of technical devices that allow us to manage and rely on the streaming feeds. Thus,

> Once again, the Internet is shifting before our eyes. Information is increasingly being distributed and presented in real-time streams instead of dedicated Web pages. The shift is palpable, even if it is only in its early stages… The stream is winding its way throughout the Web and organizing it by nowness (Schonfeld 2009).

The real-time stream is not just an empirical object; it also serves as a technological *imaginary*, and as such points the direction of travel for new computational devices and experiences. In the real-time stream, it is argued that the user will be constantly bombarded with data from a thousand different places, all in real-time, and that without the complementary technology to manage and comprehend the data she would drown in information overload. Importantly, the user is expected to desire the real-time stream, both to be in it, to follow it, and to participate in it, and where the user opts out, the technical devices are being developed to manage this too. Information management becomes an overriding concern in order to keep some form of relationship with the flow of data that doesn't halt the flow, but rather allows the user to step into and out of a number of different streams in an intuitive and natural way. This is because the web becomes,

> A stream. A real time, flowing, dynamic stream of information — that we as users and participants can dip in and out of and whether we participate in them or simply observe we are […] a part of this flow. Stowe Boyd talks about this as the web as flow: "the first glimmers of a web that isn't about pages and browsers" (Borthwick 2009).

These streams are computationally real-time and it is this aspect that is important because they deliver liveness, or 'nowness' to the users and contributors. Many technologists argue that we are currently undergoing a transition from a 'slow web to a fast-moving stream... And as this happens

we are shifting our attention from the past to the present, and our "now" is getting shorter' (Spivak 2009). Today, we live and work among a multitude of data streams of varying lengths, modulations, qualities, quantities and granularities. The new streams constitute a new kind of public, one that is ephemeral and constantly changing, but which modulates and represents a kind of reflexive aggregate of what we might think of as a stream-based publicness – which we might call *riparian-publicity.* Here, I use riparian to refer to the act of watching the flow of the stream go by. But as, Kierkegaard, writing about the rise of the mass media argued:

> The public is not a people, a generation, one's era, not a community, an association, nor these particular persons, for all these are only what they are by virtue of what is concrete. *Not a single one of those who belong to the public has an essential engagement with anything (*Kierkegaard, quoted in Dreyfus 2001b: 77, italics added).

Here too, the riparian user is strangely connected, yet simultaneously disconnected, to the data streams that are running past at speeds which are difficult to keep up with. To be a member of the riparian public one must develop the ability to recognise patterns, to discern narratives, and to aggregate the data flows. Or to use cognitive support technologies and software to do so. The riparian citizen is continually watching the flow of data, or delegating this 'watching' to a technical device or agent to do so on their behalf. It will require new computational abilities for them to make sense of their lives, to do their work, and to interact with both other people and the technologies that make up the datascape of the real-time web. These abilities have to be provided by new technical devices that give the user the ability may therefore to manage this new data-centric world. In a sense, one could think of the real-time streams as distributed *narratives* which, although fragmentary, are running across and through multiple media, in a similar way to that Salman Rushdie evocatively described in *Haroun and the sea of stories*:

> Haroun looked into the water and saw that it was made up of a thousand thousand thousand and one different currents, each one a different color, weaving in and out of one another like a liquid tapestry of breathtaking complexity; and [the Water Genie] explained that these were the Streams of Story, that each colored strand represented and contained a single tale. Different parts of the Ocean contained different sorts of stories, and as all the stories that had ever been told and many that were still in the process of being invented could be found here, the Ocean of the Streams of Story was in fact the

> biggest library in the universe. And because the stories were held here in fluid form, they retained the ability to change, to become new versions of themselves, to join up with other stories and so become yet other stories; so that unlike a library of books, the Ocean of the Streams of Story was much more than a storeroom of yarns. It was not dead but alive (Salman Rushdie, *Haroun and the sea of stories*, quoted in Rumsey 2009).

Of course, the user becomes a source of data too, essentially a real-time stream themselves, feeding their own narrative data stream into the cloud, which is itself analysed, aggregated, and fed back to the user and other users as patterns of data. This real-time computational feedback mechanism will create many new possibilities for computational products and services able to leverage the masses of data in interesting and useful ways. Indeed, we might begin to connect these practices of computational intensification with a wider computational economy which is facilitated by technology, which Kittler (1997) calls the technical a priori. These technologies may provide a riparian habitus for the kinds of subjectivity that thrives within a fast moving data-centric environment, and through a process of concretization shape the possibility of thought and action available. As Hayles (1999) states:

> Modern humans are capable of more sophisticated cognition than cavemen not because moderns are smarter… but because they have constructed smarter environments in which to work (Hayles 1999: 289).

Here, computational technology becomes instrumental to the processes of investment that individuals make into their lives, whereby success and intelligence is expressly linked to a technological process that makes these individual computational 'streams' more productive. The stream is also linked to the creation of a complex temporality through an assemblage of computational processes, through, for example the storage and recall of time-series data, a 'global' market-place and cycles of investment, dividends and company reporting requirements. These create wider oscillations which provide an informatised environment that is constantly changing but yet provides predictive patterns from seemingly random distributions of data.

The question now arises as to the form of subjectivity that is both postulated and in a sense required for the *computational* subject. In this final chapter, I want to think through the question of the subject as a computational stream, that is both a recipient of real-time data streams

as a consumer and user of data and information, but also what a stream-like consciousness might experience. After spending the majority of the book thinking through the question of code and software through the optic of the computational, I now want to turn to the question of the computational subject. To do this, I want to look at the work of Jean-François Lyotard, a French philosopher and literary theorist, especially his ideas expressed in *Postmodern Fables*. Here, Lyotard introduces 'fifteen notes on postmodern aestheticization' (Lyotard 1999: vii). In these essays, he attempts to analyze the workings of the capitalist market through culture. His method shifts to a new 'subterranean practice' in which he moves from a commitment to the future anterior or the 'what will have been', (i.e. through experimentation by proceeding in such a way that the methods only emerge through the 'playing of the game' or after the event (see Beer and Gane 2004)), to a radical politics, or aesthetics, of disruption, using the fable as an exploratory approach. As Gane (2003) explains:

> The fable plays with the boundaries between fiction and reality, and in the process disturbs the narrative structures that frame and legitimate knowledge. Lyotard consequently terms fables 'realist', because they recount 'the story that makes, unmakes, and remakes reality' (Lyotard 1999: 91, quoted in Gane 2003: 444).

The fable is a narrative means of presenting a fictional or 'elusive ought', and at the end of the chapter I would like to consider what the moral of a postmodern fable of 'being a good stream' might be, much as in *Aesop's Tales* one is left with a moral at the end of the story. As Lyotard explains,

> In the fable the energy of language is spent on imagining. Therefore, it really does fabricate a reality, that of the story which it is telling; but the cognitive and technical use of reality is left pending. It is exploited reflexively, that is to say, sent back to language so that it can link up with its subject… Leaving it unsettled is what distinguishes the poetic from the practical and pragmatic (Lyotard 1993: 242).

But for now it is important to understand Lyotard's fables as part of a project of political resistance, where the poetic or mythic offers a line-of-flight through fleeting or disruptive movements. I want to use this as a means to think about computational subjectivity, that is, subjectivity that is mediated through computer-based technologies, in other words, a 'stream'-like subject. The problems introduced when our

informationalised lives become mediated through the real-time is nicely captured by Borthwick (2009) who reflects that,

> The activity streams that are emerging online are all these shards — these ambient shards of people's lives. How do we map these shards to form and retain a sense of history? Like [that] objects exist and ebb and flow with or without context. The burden to construct and make sense of all of this information flow is placed, today, mostly on people. In contrast to an authoritarian state eliminating history — today history is disappearing given a deluge of flow, a lack of tools to navigate and provide context about the past. The cacophony of the crowd erases the past and affirms the present. It started with search and now its accelerated with the 'now' web. I don't know where it leads but I almost want a remember button — like the like or favourite. Something that registers something as a memory — as a salient fact that I for one can draw out of the stream at a later time. Its strangely comforting to know everything is out there but with little sense of priority of ability to find it becomes like a mythical library — its there but we can't access it (Borthwick 2009).

This concept of the stream as a new form of computational subjectivity also represents a radical departure from the individualised calculative rationality of *homo economicus* and tends rather toward the manipulation of what Brian Massumi calls 'affective fact', that is through an attempt to mobilise and distribute the body's capacity to think, feel and understand (either through a self-disciplinary or institutional form). Thus logico-discursive reasoning is suspended and replaced with a 'primary assemblage that links together statements, images, and passions in the duration of the body' (Terranova 2007:133). A link is formed between affective and empirical facts that facilitates and mobilises the body as part of the processes of a datascape or mechanism directed towards computational processes as software avidities, for example, complex risk computation for financial trading, or ebay auctions that structure desire. Indeed, the stream's comportment towards 'technical' or computational temporality and the connection between time, speed and movement for the maximization of output/profit lends it towards a form of subjectivity suited to the financialised practices that are becoming increasingly common today. This notion of computationally supported subject was developed in the notion of the 'life-stream':

> A lifestream is a time-ordered stream of documents that functions as a diary of your electronic life; every document you creat[139] and every

> document other people send you is stored in your lifestream. The tail of your stream contains documents from the past (starting with your electronic birth certificate). Moving away from the tail and toward the present, your stream contains more recent documents — papers in progress or new electronic mail; other documents (pictures, correspondence, bills, movies, voice mail, software) are stored in between. Moving beyond the present and into the future, the stream contains documents you will need: reminders, calendar items, to-do lists… You manage your lifestream through a small number of powerful operators that allow you to transparently store information, organize information on demand, filter and monitor incoming information, create reminders and calendar items in an integrated fashion, and "compress" large numbers of documents into overviews or executive summaries (Freeman and Gelernter 1996).

This is a life reminiscent of the Husserlian 'comet', that is strongly coupled to technology which facilitates the possibility of stream-like subjectivity in the first place. Memory, history, cognition and self-presentation are all managed through computational devices that manage the real-time streams that interact with and make possible the life streams described here. These make use of the processing improvements associated with technology, together with feedback, control and rational management, which are reminiscent of cybernetic theory and the focus on information, feedback, communication, and control (Beniger 1989). It is also argued that what we see are changes in the internal structure of the human mind and body to facilitate that productivity that previously took place in the factory (Hardt and Negri 2000). This is the restructuring of a post-human subjectivity that rides on the top of a network of computationally-based technical devices. This notion of a restructured subjectivity is nicely captured by Lucas (2010) when he describes the experience of dreaming about programming,

> This morning, floating through that state between sleep and consciousness in which you can become aware of your dreams as dreams immediately before waking, I realized that I was dreaming in code again… [D]reaming about your job is one thing; dreaming inside the logic of your work is quite another… But in the kind of dream that I have been having the very movement of my mind is transformed: it has become that of my job. It is as if the repetitive thought patterns and the particular logic I employ when going about my work are becoming hardwired; are becoming the default

> logic that I use to think with. This is somewhat unnerving (Lucas 2010: 1).

This is the logic of computer code, where thinking in terms of computational processes, as processual streams, is the everyday experience of the programmer, and concordantly, is inscribed on the programmer's mind and body. The particular logic of multiple media interfaces can also produce a highly stimulated experience for the user, requiring constant interaction and multi-tasking. According to Richtel (2010),

> heavy multitasking might be leading to changes in a characteristic of the brain long thought immutable: that humans can process only a single stream of information at a time. Going back a half-century, tests had shown that the brain could barely process two streams, and could not simultaneously make decisions about them. But Mr. Ophir, a researcher at Stanford University, thought multitaskers might be rewiring themselves to handle the load… [however actually] they had trouble filtering out… irrelevant information (Richtel 2010).

These are interventions that are made possible through new media technologies, such as word-processors, project management software and intimate technologies like the iPhone, technologies that provide an environment in which thinking is both guided in a logical fashion, but also continually fragmented across the media interface. This can change the very act of writing itself, as Heim writes:

> You no longer formulate thoughts carefully before beginning to write. You think on screen. You edit more aggressively as you write, making changes without the penalty of retyping. Possible changes occur to you rapidly and frequently… The power at your fingertips tempts you to believe that faster is better, that ease means instant quality (Heim 1993: 5).

It is this constantly present form of subjectivity that is closely linked to the computational experience of technical devices described above. These, of course, are highly dependent upon the code that makes up the data processing component that enables the streams in the first place. By displacing certain activities into the technology enables rapid reflexive augmentation of the data that is in a constant feedback loop back to the user. This is the human being as a data stream in its own right, or as is more commonly termed, a user stream.

Being a good stream

Lyotard develops his notion of the 'stream' from his previous works, *The Postmodern Condition* (1984), and *The Inhuman* (1993), where he drew attention to the rapid pace of technological change and its potent possibilities to extend rationalisation and domination. In *Postmodern Fables* he is expressly interested in technology's ability to speed up the exchange of information to such an extent that critical thought itself might become suppressed under the quantity of information. For example, in the first essay called 'Marie goes to Japan', Lyotard tells the tale of Marie, a overworked academic who must travel the world in order to 'sell her culture' and in doing so, becomes a 'stream of cultural capital: a member of a new "cultural labour force" that is exploited by choice' (Gane 2000: 444). Lyotard explicitly links economic value and speed, indeed, as he explains in the note: 'capital is not *time is money*, but also *money is time*. The good stream is the one that gets there the quickest. An excellent one gets there almost right after it has left' (Lyotard 1999: 5). For the 'good little stream' of the fable, it is the ability to produce rapidly that is the key marker of success, indeed, the faster something is completed and thus increases the stream's flow, the more profitable, the more successful and the greater the level of productivity. As Lyotard remarks:

> The best thing is to anticipate its arrival, its 'realisation' before it gets there. That's money on credit. It's time stocked up, ready to spend, before real time. You gain time, you borrow it.(Lyotard 1999: 5).

This improvement in the 'efficiency' of the individual recalls Marx's distinction between absolute and relative surplus value and the importance to capitalism of improvements in both organizational structure and technological improvements to maximizing profit (Marx 2004: 429–38). So, for example in this case, writing academic papers and books, using technology in any spare moments of time, together with mobility and participation, are the key to understanding this intensive new world of cultural production. However, this production is in a sense cut off from a sense of history, what Bruce Sterling calls the 'atemporal' (2010). This is constantly generating new forms of cultural capital through networked activity in the radical present and whose success or failure is judged in reference to current continual output. This is a form of production that is built around a normative ideal of continual

work, continual streams of discrete quantifiable products that can be distributed and which feed into other shared work. As the Invisible Committee (2009) noted,

> Ideally you are yourself a little business, your own boss, your own product. Whether one is working or not, it's a question of generating contacts, abilities, networking, in short "human capital" (The Invisible Committee 2009: 50–1).

But there is not just a relationship between the quantity of time spent on the project and the resultant success; rather, it is the compression of time, the raising of productivity and efficiency that is important. It is the reduction in total time between the inputs and outputs of a process that Lyotard is drawing attention to as, following Marx, 'moments are the elements of profit' (Marx 2004:352). In the computational, the moments are not measured in working days or hours, but rather in the 'technical time' of the computer, in milliseconds or microseconds. It is here, technologies are inserted into cultural production in order to speed-up the creation of culture and its circulation. This is related to what economists call Total Factor Productivity (TFP), that is, where technological advances have lead to a continual increase in productivity, rather than a reliance on increased capital and labor inputs. Lyotard explains, 'you have to buy a *word processor*. Unbelievable, the time you can gain with it' (Lyotard 1999: 5). This brings to mind the experience of Friedrich Nietzsche who in 1882 after having bought a typewriter to help him write due to his failing vision, found that 'our writing equipment takes part in the forming of our thoughts' (Kittler 1999: 201). Indeed, 'in 1874, eight years before he decide[d] to buy a typewriter, Nietzsche ask[ed] himself whether these are still men or simply thinking, writing, and computing machines' (Kittler 1987: 116).[1]

Materialising the stream

To be computable, the stream must be inscribed, written down, or recorded, and then it can be endlessly recombined, disseminated, processed and computed.[2] The recording includes the creation of collective notions of shared attributes and qualities, in many cases institutionally located and aggregated,[3] but also a computational narrative of the subject through the datascape specifically represented through the data points they collect through their lives, either privately as geodata, twitter feeds or such like, or publicly through health records, tax records or educational

qualifications.[4] The consistencies of the computational stream are supported by aggregating systems for storing data and modes of knowledge, including material apparatuses of a technical, scientific and aesthetic nature (Guattari 1996: 116).

These link directly to some of the issues raised by the body of work that has come to be known as medium theory, including Hayles (2005), Kittler (1997) and McLuhan (2001), that tries to think through the question of *storage* through the invention of 'new materials and energies, new machines for the crystallizing time' (Guattari 1996: 117) – particularly relevant in regard to the processes of computational flows. Here, I am not thinking of the way in which material infrastructures directly condition or direct collective subjectivity, rather, the components essential 'for a given set-up to take consistency in space and time' (Guattari 1996: 117).[5] We might think about how the notion of self-interest is materialised through technical devices that construct this 'self-interest', for example, through the inscription of accounting notions of profit and loss, assets and liabilities, which of course increasingly take place either through computer code which is prescribed back upon us.

It is important to consider the question of storage with regard to the computational stream. It is also crucial that a link is made between the computational and storage, as computation requires both the processing code and the data to be inscribed somewhere. This requires a chain of signification as 'memory' to be generated which translates the stream of data into a symbolic order through code. This technical a priori is crucial to understanding what it is possible to record at all, and the medium that translates and stores the data that forms the 'memory' of the computational. Here, we can think of computation requiring a network of writing which creates computable numbers that are divided into discrete countable finite elements. In other words, computational data is artifactualised and stored within a material symbolisation. This computational network requires a material channel through which the media of computation are carried, but as Kittler (1997) notes, it is a characteristic of every material channel that beyond, and against, the information it carries, it produces noise and nonsense. We have the assemblage of a network which builds the material components into an alliance of actors and which is a referential totality for the meaning that is carried over it, and past its borders, policed by human and non-human actors, we have what Doel (2009) calls *excess* and Latour (2005) calls *plasma*.

Here, then, we see the movement or translation between the temporal generation of the discrete elements of the stream and the computational storage through what Kittler calls *time axis manipulation*. This is the storing of time as *space*, and allows the linear flow to be recorded

and then reordered. The shifting from chronological time to the spatial representation means that things can be replayed and even reversed, this is the discretisation of the continuous flow of time. Without it, the complexity of financial markets would be impossible and the functions and methods applied to it, through for example the creation of new abstract classes of investment such as Credit Default Swaps (CDSs), Collateralised Debt Obligations (CDOs) and Asset Backed Securities (ABSs) would be extremely difficult, if not impossible, to create and trade.

This implicit datastream across all devices leads to an enormous amount of data being collected and held in corporate databanks and huge data centres. As Borthwick (2009) noted, bit.ly, an Url shortening site, had collected 200 gigabytes of click data by 2009, including: usage data, location data, and so forth, about the users of the site which the users would not be aware had been collected. This is the idea of a 'dataspace', richly endowed with content which is dereferentialised and equally accessible by being located within a database and which makes the presence of data seem addictive and overwhelming. As Borgman notes,

> The glamorous fog of cyberspace varies in thickness. It's denser when we sit in front of the computer than when we are face to face with a person. It's thinner for the driven and the ambitious than for the sullen and the addicted. But when it is thick, it's disorienting in a new and distinctive way. The problem is not that we can't find what we are looking for, but that we are not sure what to look for in the first place. Whatever we have summoned to appear before us is crowded by what else is ready to be called up. When everything is easily available, nothing is commandingly present (Borgman 2010).

This notion of the computational dataspace is explicitly linked to the construction of the stream-like subject and raises many important questions and challenges to the liberal humanist model of the individual. Most notably in their bounded rationality – here the information and processing to understanding is off-loaded to the machine – but also in the very idea of a central core of human individuality. It also returns us to the question of digital *bildung* and how we structure the kind of education necessary in a computationally real-time world. For example, although,

> most streams today are explicitly created by users, either by creating content, making a friend, saving a favorite etc. For every explicit action of a user, there are probably 100+ implicit datapoints from usage; whether that is a page visit, a scroll, a video/shopping abandon etc (danrua, comment in Borthwick 2009).

However, we must not lose sight of the materiality of these computational forms which is inscribed within a material substrate. This is a computationally generated digital world that is limited by certain material affordances in the use of technologies such as processor capacity (i.e. computers do not have an unlimited amount of time nor do they have infinite storage space). Computation creates a technical form of time through the conservation, accumulation and sedimentation of past stream data, what we might call its memories or its past, which is then rearticulated in light of unfolding new data computation. This is what Heidegger (1988: 260) called the technical measurement of time, the attempt to determine the undetermined through the recording of the past through an apparatus of inscription. Without preservation, there is no stream, as it would be a mere atomistic point in time. Without the storage and recall of data there is no computational possibility for the construction and action of a focal attention by the stream. Together, these devices form complex assemblages that entangle the user/stream into a particular memory-temporality which creates the conditions for particular kinds of agency.

Heidegger considered 'authentic' time to be time in relation to death, as finitude and mortality. In the time of the computational stream, however, time is found in the inauthentic time of measurement, the attempt to determine the 'undetermined' through technical devices. So, for example, in the case of the computational, there is only the abstract notion of time as reported through the continual ticks of the data-streams and the charts and visualisations that represent the time-series datascape to the viewer.

The notion of subjectivity that is embedded in the socio-technical networks of computational systems points towards a deathless existence, even as, paradoxically, the market continually relies on the anxiety represented through sickness and disease, poverty and old age, to activate and fuel desire. These technologies operate to create and sustain a market which introduces a time of indeterminacy and choice into the stream of flows even as they stimulate affective responses within the stream calling for forms of action. For financialised streams, for example, the ticks of financial data are linked to the body as the profit or loss of securities and entangled with desires, necessity and ontological security. For the user stream, it is a constant flow of everyday activity represented as a chaotic uncoordinated stream of events logged to a microblogging site.

These streams are undoubtedly creating huge storage issues for the companies that will later seek to mine this collection of streamed data. For them, the problems are manifested in the building of massive

computational data centres in locations around the world. Trying to capture the ready-to-hand world of everyday life generates such a large flow of data that can easily overwhelm these systems, witness the intermittent downtime of services like Twitter, which are also forced to regulate the flow of data into their networks through API feeds. Connected to this storage medium are the processing practices that are applied to render the stream of computational data as a source of action. These allow the analysis and visualisation of computational patterns over time, and allow the discernment of trends and traces that are left as markers within the data. There is a growing and important literature on the issue of data visualisation in general (see Pryke 2006; Manovich 2008), and financial markets in particular (Beunza and Stark 2004; Beunza and Muniesa 2005; Knorr Cetina and Bruegger 2002), here, I can only note the importance of the visual mediation of this data and its highly aestheticised content but clearly with the amount of data available the skills of a visual rhetoric will become increasingly important to render the patterns in the data meaningful.

Using financialisation as an example of a type of computational subjectivity, we might link the movement of a calculative rationality to that of an affective distributed rationality, geared towards the consumption of a financialised range of goods and services. I mark and develop the notion of the stream in the section below through a discussion of the notion of financialisation and a tentative cartography of the subjectivity associated with it, which I connect to the 'degradation of the individuals capacity for *understanding* their own circumstances, and their ability to make any effective use of whatever *correct understandings* they might achieve' (Terranova 2007: 132, original emphasis) – here particularly through the dichotomy of pattern/randomness (and here I want to connect randomness to a notion of plenitude). I want to think about the way in which life itself becomes understood as a 'life-stream' through the application of memory systems designed to support a highly informatised and visualised computational economy. That is, I want to understand the stream as a 'propagation of organised functional properties across a set of malleable media' (Hutchins 1996: 312). Connected to this are notions of calculability and processing, which relate back to the creation of technical devices that facilitate the user's ability to make sense of the movements in markets, data, and culture and more particularly, to respond to changes in risk and uncertainty. Users treat their lives as one would a market portfolio, constantly editing the contents through buying and selling, creating new narratives through the inclusion or exclusion of certain types of product or data stream.

Financial streams

Financialisation is an analytical term used to describe the processes of finance capital, including the institutions, norms, practices and discourses that are connected with it. It is thus a useful means to unpack the way in which claims to an information society or knowledge economy are bound up with particular situated approaches to organising the economy, society and politics. Financialisation has implicit within it, certain ways of acting, certain ways of being and certain ways of seeing that are connected to a particular comportment to the world, one that is highly attenuated to notions of leverage, profit and loss and so forth. Moreover, financialisation implies that the rational actor of economic theory is transformed from the calculative rationality of the protestant work ethic to an actor that is guided not only by rational self-interest but also a propensity to understand and take highly-leveraged and complex risks. That is, to move beyond Weber's description of the religious basis of capitalism as 'exhort[ing] all Christians to gain all they can, and to save all they can; that is, in effect, to grow rich' (Wesley quoted in Weber 2002: 119, emphasis removed). Where Weber described monetary acquisition as saving linked to an ethical norm supplied by protestant faith – that is, an understanding as labour and saving as a calling – with financialisation we see quite the opposite with a move towards the use of debt financing to fund investment and consumption to the extent that its lack of ethical grounding arguably leads inexorably towards endogenous financial instability through Speculative and Ponzi modes of investing (Minsky 1992). In a different register, Belfrage (2008: 277) glosses financialisation as 'emerging out of conditions which force people to weigh up the market performance of their financial assets when making everyday decisions between saving and consuming'. Financialisation is, nonetheless, an essentially contested concept, and as Randy Martin (2002) explains:

> Financialisation, like those other recently minted conceptual coins postmodernism and globalization, gets stretched and pulled in myriad directions. Part of the complexity of these terms is that they stand simultaneously as subject and object of analysis—something to be explained and a way of making sense out of what is going on around us.

Here, I follow the work in the sociology of markets to understand financialisation as the uneven process of formation of a socio-technical

network that is used to stabilise a certain kind of calculative cognitive-support, that mediates the self and the world through financial practices, categories, standards and tests (Callon 1998). More importantly, I want to link the processes of financialisation to the creation of rapidly changing data streams of financial information. Rather than restricting the notion to the purely discursive or economic, I want to tentatively explore the idea that financialisation itself is the establishment of valuation networks; that is, the construction of circuits of finance which render abstract financial objects commensurable and exchangeable, in which actors, both human and non-human, are enrolled. This is in distinction to cognitive psychology that sees the ability for actors to calculate as being either rendered within a form of mental calculation which cognitive anthropology, has shown to be far too demanding, and also in distinction to cultural approaches which see the calculative competence through social structures or cultural forms and which is unable to explain the shift from one form of calculative agency to another (Callon 1998: 4–5). Further, I want to challenge the notion of a linear process of financial transformation – 'financialisation', and instead highlight the way in which there is an assemblage of '*financial mediation*' itself marked by a series of tensions, counter-tendencies and modulations.

A financialised assemblage is connected together through the use of equipment or financial computational devices (what Deleuze would call *agencements*) whose aim is to maintain an anticipatory readiness about the world and an attenuated perception towards risk and reward which is mediated through technical affective interfaces (i.e. the computer user interface). In the first place, the computational is directly linked to quantitative statistical processing of massive amounts of time-series data and its visualisation or representation. Additionally, however, the affective dimension seems to me to be extremely important in understanding the way in which recent shifts in financial markets towards the democratisation of access have been intensified through the realignment of desire with the possibilities offered through monetary returns from finance capital – what Bloom (n.d) has called 'computational fantasies' – and it is a subject I'll return to below. But none of these practices of intensification could have been possible without information technology, which acts as a means of propagation but also a means of structuring perception – or better, of 'focusing' attention in the sense of an extended mind. Finance itself has a 'feel to it' which is generated via the computer interface or through the marketing and packaging that 'wraps' the underlying financial product.

This affective dimension to finance is also interactive, providing a model of action that situates the user (or investor) in a relation of continual interaction with their portfolio. The important point here is that you do not need to have a 'whole' human being who has intentionality and therefore makes rational decisions about the market, or has feelings and is responsible for their actions and so on. Rather, you can obtain a complete human being by composing it out of composite assemblages which is a provisional achievement, through the use of computer cognitive support (what Latour (2005) neatly calls 'plug-ins') and we might think of as software interfaces or technical devices. For example, these are the share-trading systems that initially pre-format the user as a generic market investor. But to be an active investor requires the use of particular techniques and strategies in the market supported through extra software interfaces that offer guidance on 'reading' the market (see for example the websites: The Motley Fool, or Interactive Investor). One example of this is that of Swedish pension reform, where individual pension investment is part of a process amenable to 'nudges' by technical devices that help guide the individual through up to 1000 different investment funds (Thaler and Sunstein 2009).

These can be understood as structuring templates that act as devices to give you the capacity to calculate, that is, cognitive abilities that do not have to reside in 'you' but can be distributed throughout the investment interface. It is important to note, however, that the extent of the 'nudge' that the system can provide can range from the libertarian paternalism of defaults and formatting advocated by Thaler and Sunstein (2009) to posthuman distributed aids to cognition, or even collective notions of cognition, as described by Hutchins (1996). An example is the portfolio manager software offered by a number of companies online, which purport to not only hold the investment portfolio, but rather to stimulate you to invest, trade and have a way of being-towards the market which is active (Interactive Investor is a notable web-based example). This can be achieved through email alerts set to certain time-series prices, automatic trading systems and constant feedback to the user via mobile technologies (see the Stocks.app application on the Apple iPhone, for a mobile example). Wherever the investor is, they are able to call up the portfolio and judge their asset worth as defined by the external forces of the financial markets but crucially simplified and visualised through the graphical capabilities of the mobile device (see tdameritrade, Etrade, iStockManager, m.scottrade.com, etc.). Sometimes, in a radical break with the notion of judgement being the seat of humanity and contra Weizenbaum (1984), the software can

even judge the success of the investment strategy through a number of algorithmic heuristics, something the investor may not even have the calculative or cognitive ability to challenge.

In the case of financial markets, software has completely changed the nature of stock and commodity markets creating 24 hour market trading and enabling the creation of complex derivative products and services, often beyond the understanding of the traders themselves. For example, high frequency trading (HFT) is at the cutting edge for trading on financial markets, the basic idea of HFT is to use clever algorithms and super-fast computers to detect and exploit market movements. To avoid signalling their intentions to the market, institutional investors trade large orders in small blocks—often in lots of 100 to 500 shares – and within specified price ranges.

> High-frequency traders attempt to uncover how much an investor is willing to pay (or sell for) by sending out a stream of probing quotes that are swiftly cancelled until they elicit a response. The traders then buy or short the targeted stock ahead of the investor, offering it to them a fraction of a second later for a tiny profit (*The Economist* 2009).

These changes in the practices of stock market trading reflect the implementation of high technology networks and software, indeed,

> HFT is a type of algorithmic trading that uses high-end computers, low-latency networks, and cutting-edge analytics software to execute split-second trades. Unlike long-term investing, the strategy is to hold the position for extremely short periods of time, the idea being to make micro-profits from large volumes of trades. In the US, it is estimated that 70 percent of the trade volume is executed in the HFT arena (HTCWire 2010).

This technology came to public attention on 6 May 2010 when the Dow plunged nearly 1,000 points in just a few minutes, a 9.2 per cent drop, and christened the 'flash crash'. Half a trillion dollars worth of value was erased from the market and then returned again. Due to the work of software engineer Jeffrey Donovan, it became clear that HFT systems were shooting trades into the markets in order to create arbitrage opportunities. By analysing the millisecond data stream logs of the exchange and reverse-engineering the code, he was able to see the tell-tale signs of algorithmic trading in cycles of 380 quotes a second that led to 84,000 quotes for

300 stocks being made in under 20 seconds, which set off a complex chain of reactions in the market and the resultant slump (HTCWire 2010).

Financial companies are rolling out new experimental technologies continually to give them an edge in the market place; one example is the so-called 'Dark Pools' (also known as 'Dark Liquidity'). These off-market trade matching systems work on matching trades on crossing networks which give the trader the advantage of opaqueness in trading activities, such as when trying to sell large tranches of shares (Bogoslaw 2007). Dark pools are 'a private or alternative trading system that allows participants to transact without displaying quotes publicly. Orders are anonymously matched and not reported to any entity, even the regulators' (Shunmugam 2010). Additionally, technologies such as 'dark algorithms' give firms the ability to search multiple dark pools to find hidden liquidity.

Software that acts in this cognitive support capacity can therefore be said to become a condition of possibility for a device-dependent, co-constructed subjectivity. This Guattari (1996: 114) calls a 'processual' subjectivity that 'defines its own co-ordinates and is self-consistent' but remains 'inscribed in external referential coordinates guaranteeing that they are used extensively and that their meaning is precisely circumscribed' (Guattari 1996: 116). The subject, then, is circumscribed by the technologies which mediate its relationships with finance capital, such that the field of experience is constantly shifting to reflect financial data and the movement of time. Following Lyotard, we might declare that the subject becomes a computational 'stream', in this case a stream attenuated to the risk associated with finance capital mediated through financial software.

Of course, risk itself is a pivotal category in modern finance that is stabilised through the use of technology and discourse. Risk, for Langley (2008), is distinct from uncertainty, where uncertainty is understood as non-calculable future volatilities that are beyond prediction, and risk itself is a statistical and predictive calculation of the future. Langley explains:

> There is no such thing as risk in reality... risk is a way – or rather, a set of different ways – of ordering reality, of rendering it into a calculable form it is a way of representing events in a certain form so that they might be made governable in particular ways, with particular techniques and for particular goals' (Dean quoted in Langley 2008: 481).

This is, of course, the notion of risk developed by the economist Frank Knight in his 1921 book *Risk, Uncertainty and Profit.* When encoded into

financialised software, risk is qualified, rendered and abstracted in a calculative space which interfaces to the investor through devices that seek to present the world through what Taleb (2007) calls Gaussian risk. This is risk that is presented without its limitations as a model made transparent, and that falls short of fully containing the complexity and uncertainty of life. Risk itself becomes mediated through software and becomes a processual output of normative values which are themselves the result of computational processes usually hidden within the lines of computer code. For example, software renders the display of financial portfolio information in a very stylised, simplified form, often with colour codings and increasingly with rich graphics.[6] Not only do few market participants fully understand risk as a statistical category, but the familiar bell-shaped curve of Gaussian distributions displayed on mobile screens, encourages a kind of 'domesticated' approach to risk that makes it appear familiarised and benign. Indeed, it is this misunderstanding of risk that Taleb (2007) blames for the huge leveraged asset bubble in 2007–2009 and the resultant financial crisis.[7] Indeed, only recently AXA S.A., the French financial services giant, was forced to reveal that 'that it had made a "coding error" that affected returns in its various portfolios in ways that had yet to be determined', but which could have resulted in substantial losses, and that '[i]t was an "inadvertent mistake" entered into one of AXA Rosenberg's main "risk models" by a computer programmer in April 2007' (Sommer 2010). Three years of a computer programming bug on a portfolio which at its height was worth $62 billion, demonstrates the profound effects that computer code can have, indeed, the portfolio, at the time of writing, is worth $41 billion after many investors have begun to leave the fund due to worries about the bug's effects.

These conceptualisations and arguments are clearly an important part of the content of financialisation, but now I would like to turn to the notion of the computational 'stream' by extending Lyotard cultural understanding of the stream. This concept helps to map real existing 'territories' (such as sensory, cognitive, affective and aesthetic) in relation to computational processes. Here software is active in the creation and maintenance of a temporal dimension that supports particular kinds of subjectivity. Indeed, this links with Thrift's (n.d.) notion of our having a 'minimal conscious perception which is held in place by all manner of systems and environments and sites that extend awareness' (n.d.: 3). Here, we can think of the external management of the internal perception of time that is linked to a form of Heideggerian angst towards a future event – sickness, old age, and so forth – which provides

a new affective fuel source for capitalism. This is an anxiety maintained through a destabilising sense of the rapid passage of time, manifested through, for example, continual and inexplicable changes in commodity prices, stock valuations, asset price expansions and contractions – themselves fed as data streams to the processual subject. These are connected through a series of mechanisms to the body, and here I am thinking of a machinic notion drawn from Deleuze and Guattari (2004), or to an emotional response to the representation of the future body given through a series of visual images, such as actuarial graphs and charts (again connected to the notion of mobile spaces of risk or financialisation through devices such as the Apple iPhone). This 'full-on or full palette capitalism' (Thrift n.d.: xx) functions through the exploitation of forethought, where the aim is to produce a certain expectation and preparedness into which a desire is linked to the intensification of action.

This life stream is therefore a performative subjectivity highly attenuated to interactivity and affective response to an environment that is highly mediatised and deeply inscribed by computational datascapes. This helps to explain the kinds of active investor subjects that Governments seek to encourage through financial regulation such as annual tax renewal requirements, for example in Investment Saving Accounts (ISAs), a form of tax-free saving account in the UK, which require the accounts to be moved or reinvested every April; or in the Swedish Pension case outlined by Belfrage (2008) where the intention was to encourage over 50 per cent of pension savers to undertake continual asset management activities in relation to Swedish worker's pension portfolios invested in the Stock Market (the actual number of active traders turned out to be only 8 per cent in 2005) (Belfrage 2008: 289).[8] Clearly, the financialisation of society remains a work in progress.

So financialised code is a complex set of materialities that we need to think carefully about in turn. From the material experience of the financialised user of code, both trader and consumer, to the reading and writing of code, and then finally to the execution and experience of code as it runs on financial trading systems, we need to bring to the fore how code is a condition of possibility for a computational stream whether of financial news and data, or of a datastream cognitive support for everyday life.

Lifestreams

I now want to look at the practice of creating lifestreams, particularly through the example of Twitter. Twitter is a web-based microblogging

service that allows registered users to send short status update messages of up to 140 characters to others (Herring & Honeycut 2009: 1). From a few messages per day, known as Tweets, in 2006 the service took off in popularity in 2009 and has grown to handle over 90 million messages per day in 2010 (Twitter 2010). Twitter works by encouraging the uploading and sharing of photographs, geodata tags, updates on what you are doing and so forth, this is transformed into a real-time stream of data that is fed back onto the web and combined with the updates of other people whose user-stream you 'follow'. It is helpful to,

> think about Twitter as a rope of information — at the outset you assume you can hold on to the rope. That you can read all the posts, handle all the replies and use Twitter as a communications tool, similar to IM — then at some point, as the number of people you follow and follow you rises — your hands begin to burn. You realize you cant hold the rope you need to just let go and observe the rope (Wiener, quoted in Borthwick 2009).

Although originally considered a marginal activity, Twitter, and similar microblogging services, have risen dramatically in use throughout the last few years. Particularly as politicians and the media have caught on to the unique possibilities generated by this rapid communicational medium. Designated as solipsistic and dismissed at first by the pundits, the growth in Twitter's use has meant that it can no longer be ignored and indeed it has become a key part of any communication strategy for politics, corporations and the media more generally. Twitter has evolved rapidly from a simple messaging service, to a form of real-time rolling news reporting on political and other events, from formal political meetings to protest actions. Political examples from the UK have included Gordon Brown's foray into Twitter defending the NHS over the issue of NHS 'death panels' (Toppling and Muir 2009). More recently, we have witnessed usage by the political class in the UK across the whole of the country (The Independent 2009); Damian McBride's emails to LabourList blogger Derek Draper, which were widely 'retweeted' by Twitter users (BBC 2009); and increasing concerns over the freedom of speech implications posed by the libel action against the *Guardian* reporting a parliamentary question about Trafigura regarding its relationship to exporting materials, which was widely 'retweeted' following an injunction to stop reporting on the incident (Dunt and Stephenson 2009). There is an increasing need for a cartography of both the production and empirical content of a number of these collaborative, streamed

institutions and their recording of political events, power and interests. Institutions as diverse as Downing Street, the White House, Scotland Yard, The UK Parliament, INTERPOL, NATO, the Labour Party, and the Conservative Party have all recently instituted mechanisms for using these real-time computational services to supplement the limitations of better established communications procedures.

The conditions underpinning this shift, however, are not solely communicational. What marks these real-time stream sites is their creation by the active contributions of an epistemic community surrounding the 'owner'. These communities are typically marked by very loose ties, often no more than a 'screen-name' or even anonymous contributions to the site through updates. They also have the capacity to create a form of social contagion effect whereby ideas, media and concepts can move across these networks extremely quickly. Over the past ten years, we have witnessed an explosion of media forms made possible by the peer-to-peer technologies of the Internet (Atton 2004, Benkler 2007, Gauntlett 2009, Terranova 2004) transforming political institutions and their relationship to citizens (Coleman 2005; Chadwick 2007). As such, real-time streams presents an excellent opportunity for tracing the impact of computational real-time devices in everyday life and the way in which they capture the informal representations of issues with which contemporary communities are becoming increasingly concerned. It is possible that Twitter and other real-time streams both decentre social structures and expand the numbers involved.

Filled with constant updating, real time 'tweets', Twitter users disseminate affect, opinion-formation, and information in a very Tardian way. Twitter, and similar real-time stream services, collect data from both elites and non-elites and can be used to reconstruct knowledge of social and political events in an online real-time context. Examples include the real-time Twitter feeds following national political debates, World Cup football matches, fashion and culture events, and the presentation of prestigious prizes and awards, such as Baftas or Grammys. The attention of political, technology and media communities have been captured by the emergence of the 'real-time web' using Twitter and other services such as Facebook, Quora, Diaspora and Meebo. But as more people participate and subscribe to the services, the difficulties in negotiating a large and complex information resource becomes acute. The network effects combined with the vast amount of information flowing through the network are difficult for the user to understand.

Twitter therefore acts to facilitate a form of social communication by rapidly distributing information and knowledge across different streams.

Indeed, Twitter is made up of streams of data that constitute a 'now web [that is] open, distributed, often appropriated, sometimes filtered, sometimes curated but often raw' (Borthwick 2009). But it is the technology that makes up Twitter that is a surprising: a simple light-weight protocol that enables the fast flow of short messages,

> The core of Twitter is a simple transport for the flow of data — the media associated with the post is not placed inline — so Twitter doesn't need to assert rights over it. Example — if I post a picture within Facebook, Facebook asserts ownership rights over that picture, they can reuse that picture as they see fit. If I leave Facebook they still have rights to use the image I posted. In contrast if I post a picture within Twitter the picture is hosted on which ever service I decided to use. What appears in Twitter is a simple link to that image. I as the creator of that image can decide whether I want those rights to be broad or narrow (Borthwick 2009).

Increasingly, we are also seeing the emergence of new types of 'geo' stream, such as location,[9] which give information about where the user is in terms of GPS co-ordinates, together with mixed media streams that include a variety of media forms such as photos, videos and music. Location based services, such as Facebook Places, FourSquare and Gowalla, enable a user to capture GPS information in real-time, updating this as a data stream recording places, activities and life events to the Internet. It is even argued that we are seeing the emergence of a new communication layer for the web based on micro-messages and sophisticated search. As Borthwick explains, '[i]f Facebook is the well organised, pre planned town, Twitter is more like new urban-ism — its organic and the paths are formed by the users' (Borthwick 2009). But this is not just a communications channel, it is also a distributed memory system, storing huge quantities of information on individuals, organisations and objects more generally. The things that are 'collected' and updated by users into these streams is remarkable, for example one user: (i) 'collect[s] sugar levels everyday (like 6 times per day). This helps me to "understand" my metabolism, my diet and my stress levels'; (ii) 'calorie expenditure and effort during my workouts'; (iii) 'blood glucose level every 5 minutes through a continuous glucose monitor stuck in my gut'; (iv) '[and] track my sexlife at bedposted.com (duration, intensity, positions)' (quoted on FlowingData 2010). This is what Kevin Kelly has revealingly called the quantified self (Kelly 2010). This raises serious privacy issues, but also the cultural and social implications of living life in such a public

way, mediated through the code that is enabling and supporting these services.

These new real-time streams and their relationships to both individuals, organisations, culture and society, let alone the state and politics, are still an emergent sphere of research. Many questions remain unanswered, not the least of which is who owns these huge data reservoirs and how will this data be used in the future. Indeed, Twitter recently turned over every tweet in its archive to the Library of Congress and now all tweets are archived automatically,

> every public tweet, ever, since Twitter's inception in March 2006, will be archived digitally at the Library of Congress. That's a LOT of tweets, by the way: Twitter processes more than 50 million tweets every day, with the total numbering in the billions (Adams 2010).

These streams are fascinating on a number of different levels, for example questions remain over the way in which national identity might be mediated through these computational forms in terms of an imagined community composed of twitter streams that aggregates institutions, people and even places.[10] Real-time streams offer some exciting potential in terms of cultural streams and movements as aggregations of data streams, real-time State representation through state institutions in a constellation of streams, and even national aggregates. Whether new political subjectivities are enabled through these streams, one is sure that the data will be captured and analysed as the capacity of these life-stream systems mature. This will be increasingly revealing for real-time polls, opinion formation Tardian analysis of social aggregates, and, of course governments and multinational corporations eager to monitor and manipulate the creators of these streams.

So far, I have mapped a number of different strands which coalesce around notions of aesthetics, affectivity, risk and processual subjectivity. Most importantly, I think I have tried to outline the value of Lyotard notion of the stream as a concept for developing our understanding of the computational subjectivity. I have only been able to outline some of the key areas of enquiry which I think are relevant to this, and there is clearly much work to be done in understanding the relationship between forms of computational temporality, subject positions and technological mediation and materiality. Additionally the highly visualised form of data representation that is increasingly used to express data in a qualitative form, together with the computational relationship with self raised by reflexive use of life-streams also raise important questions.

In the final section, I want to shift focus and consider the wider implications of thinking-streams, computer code and software.

Subterranean streams

Perhaps the first principle that one might consider with respect to computational devices is that they appear to encourage a search for simple solutions and answers to problems, and therefore a backlash against complexity (especially when they become screenic – in common with other mediums such as television, film and print). The solutions become mediated through technological proposals which themselves rely on computational notions such as computability, distributed processing, intensively recursive dynamics and computationally correct narrative strategies. The language of nature, politics, culture, society and economics becomes infused with computability to the extent that data flows outside of human consciousness and that in order to understand and act upon them, additional computational strategies are required (this is indeed the paradigm suggested through digital humanities and cultural analytics frameworks). This points towards an intensity of fast moving technological culture that privileges data streams over meaning, that is, an explosion of knowing-that rather than knowing-how – and here we might note the current political fascination with Twitter and similar social networking sites.

This could lead to a situation in which the user is unable perceive the distinction between 'knowing-how' and 'knowing-that' relying on the mediation of complexity and rapidity of real-time streams through technology. This Heidegger would presumably describe this as dasein no longer being able to make its own being an issue for itself. Indeed, this may even point to a homogeneity of being in the digital 'age', as we become a being whose existence is mediated by identical computational processes. This would have grave implications for a distributed fragmentary subject relying on computational devices that are radically uncertain and opaque. Indeed, if these computational devices are the adhesives which fix the postmodern self into a patterned flow of consciousness (or even merely visualised data), an ontological insecurity might be the default state of the subject when confronted with a society (or association) in which unreadiness-to-hand is the norm for our being-in-the-world.

To return to the question from Sellars and reframe it: it still remains difficult to reconcile the homogeneity of the manifest image with the non-homogeneity of the scientific one, but we have to additionally address the

unreadiness-to-hand of the computational image which offers the possibility of *partial reconciliation* through uncertain affordances. Additionally, the computational image in mediating a world of information, computation and process might inevitably transform the manifest image of meaning and complexity by disconnecting the possibility of familiarity from the referential totality and the subsequent reclassification of the personhood of dasein.[11] As we are inserting the computational image into the structure of everyday things, and therefore into the structure of our everyday life and its knowing-how, the deeper implications remain unclear and raise the need for a deeper understanding of the centrifugal force of the computational image. If the manifest world is the world in which dasein, 'came to be aware of [itself] as [being]-in-the world', in other words, where dasein encountered him/herself as dasein (Sellars 1962: 38), then the eclipse or colonisation by equipment that remains unready-to-hand and that fragments and destabilises the possibility of a referential totality would suggest that the manifest image, in so far as it pertains to man or woman, is now potentially a 'false' image and this falsity threatens dasein as it is, in an important sense, the being which no longer has this image of itself.[12] For poststructuralist writers such as Foucault, talking about certain structural conditions of possibility,

> If... [they]... were to disappear as they appeared, if some event of which we can at the moment do no more than sense the possibility... were to cause them to crumble, as the ground of Classical thought did, at the end of the eighteenth century, then one can certainly wager that man would be erased, like a face drawn in sand at the edge of the sea (Foucault 2002: 422).

This would represent the final act in a historical process of reclassification of entities from persons to objects – potentially, dasein becoming an entity amongst entities, an stream amongst streams – with challenging political and cultural implications for our ability to trace the boundary between the human and non-human.[13] This, of course, returns us to the questions raised at the beginning of the book regarding humanity's ontological precariousness. In allowing the computational to absorb our cognitive abilities, off-loading the required critical faculties that we presently consider crucial for the definition of a life examined, we pay a heavy price, both in terms of the inability of computational methods to offer any way of engaging with questions of being, but also in the unreadiness-to-hand that computational devices offer as a fragmentary mediation of the world. This is where the importance of digital *Bildung*

becomes crucial, as a means of ensuring the continued capability of dasein to use intellect to examine, theorise, criticise and imagine. It may also raise the possibility of a new form of resistance for a dasein that is always at the limit of emancipation as the being that is constantly dealing with equipment that is radically unready-to-hand. It is, to attempt to consider the way in which computation enables what Turing called the 'super-critical mind', one that is apt at generating more ideas than it received, rather than the sub-critical mind (Latour 2004: 248):

> The majority of [human minds] appear to be "sub-critical"… An idea presented to such a mind will on average give rise to less than one idea in reply. A smallish proportion are super-critical. An idea presented to such a mind may give rise to a whole "theory" consisting of secondary, tertiary, and more remote ideas (Turing 1950: 454).

The future envisaged by the corporations, like Google, that want to tell you what you *should* be doing next (Jenkins 2010), presents knowledge as 'knowing that', which they call 'augmented humanity', I consider this as a model of humanity that is a-critical. Instead, we should be paying attention to how computation can act as a *gathering* to promote generative modes of thinking, both individually and collectively, through super-critical modes of thinking created through practices taught and developed through this notion of digital *Bildung*. This would, as Latour explains, 'require all entities, including computers, cease to be objects defined simply by their inputs and outputs and become again things, mediating, assembling, gathering' (Latour 2004: 248).

In this book, I have attempted to outline a groundwork for understanding, in the broadest possible sense, how 'one know one's way around' in a world that is increasingly reliant on computational equipment, but more maps are needed. Computationality tends towards an understanding of the world which, whilst incredibly powerful and potentially emancipatory, cannot but limit the possibilities of thought to those laid within the code and software which runs on the tracks of silicon that thread their way around technical devices (sub-criticality). Understanding software is a key cultural requirement in a world that is pervaded by technology, and as Vico argued, as something made by humans, software is something that can and should be understood by humans. Indeed, this remains a project that is still to be fully mapped and has important consequences for the fragmentary way-of-being which continues to be desired throughout the socio-technical technicity that makes up the computational image.

In the spirit of Lyotard's expression of an aesthetics of disruption, however, I want to end the book with an elusive ought. This is an ought that is informed by a reading of *Aesop's Tales* through Michel Serres and his notion of the parasite (Serres 2007). The parasite is used not as a moral category, but in connection with an actor's strategic activities to understand and manipulate the properties of a network. Here, the parasite acts as interference, as processes that combine and mix together domains, for Serres it is this recombinant property of circulation networks rather than their general underlying patterns that is crucial to understand them. He explains:

> A human group is organized with one-way relations, where one eats the other and where the second cannot benefit at all from the first… The flow goes one way, never the other. I call this semiconduction, this value, this single arrow, this relation without a reversal of direction, 'parasitic' (Serres 2007: 5).
>
> The introduction of a parasite into the system immediately provokes a difference, a disquilibrium. Immediately, the system changes; time has begun (Serres 2007: 182).

For example, parasitic economic activities manipulate goods already available and subvert them from their original function. They are embedded in such a way as to make their removal either impossible or too expensive – reminiscent of the phrase 'too big to fail'. Finance capital and the equipment it deploys to assemble the markets that sustain it therefore acts to counteract the way in which investors look for liquidity and their ability to invest where they cannot get 'stuck', and from which they can withdraw at the smallest sign of trouble. Through parasitic technologies, users are constantly enticed back into the market, where they themselves intend to eat at the benefit of another. Here, the notion of the stream is intensified through the action of time within computational networks, literally the 'ticks' of network time which reflect the actions of millions of streams within the network and which cascade through the data streams that are threaded through the networks and chains of causality. As Serres argues:

> To parasite means to eat next to. Let us begin with this literal meaning. The country rat is invited by his colleague from town, who offers him supper. One would think that what is essential is their relation of resemblance or difference. But that is not enough; it never was. The relation of the guest is no longer simple. Giving or receiving, on the

> rug or on the tablecloth, goes through a black box. I don't know what happens there, but it functions as an automatic corrector. There is no exchange, nor will there be one. Abuse appears before use. Gifted in some fashion, the one eating next to, soon eating at the expense of, always eating the same thing, the host, and this eternal host gives over and over, constantly, till he breaks, even till death, drugged, enchanted, fascinated. The host is not prey, for he offers and continues to give (Serres 2007: 7).

Aesop's fable ends with the country mouse returning home declaring that it is better to be able to enjoy what you have in peace, than live in fear with more. But, Serres (2007) also gestures towards an alternative parasitic understanding in his retelling of the fable. The question of who this subject 'eats next to', is perhaps reflected in the way in which streams pass through other streams, consumed and consuming, but also in the recorded moments and experiences of subjects who remediate their everyday lives. This computational circulation, mediated through real-time streams, offers speculative possibilities for exploring what we might call parasitic subjectivity. Within corporations, huge memory banks are now stockpiling these lives in digital bits, and computationally aggregating, transforming and circulating streams of data – literally generating the standing reserve of the digital age. Lyotard's (1999: 5) comment to the streams that flow through our postmodern cultural economies seems as untimely as ever: 'true streams are subterranean, they stream slowly beneath the ground, they make headwaters and springs. You can't know where they'll surface. And their speed is unknown. I would like to be an underground cavity full of black, cold, and still water'.

Software tunnels through the rags 'n refuse

Text of talk given at the Platform Politics conference in Cambridge 13.05.110

It took New York police officer, William Barker two hours to find Homer Collyer dead in his apartment in March 1947. Barker had to crawl through a window into a second-story bedroom, burrow his way through newspaper bundles, empty cardboard boxes lashed together with rope, the frame of a baby carriage, a rake, and old umbrellas tied together, folding beds and chairs, half a sewing machine, boxes and parts of a wine press. For the next two days police continued to search the house, literally finding their way through 25,000 books, a horse's jawbone, a Steinway piano, an early X-ray machine, baby carriages, a doll carriage, rusted bicycles, old food, potato peelers, a collection of guns, glass chandeliers, bowling balls, camera equipment, the folding top of a horse-drawn carriage, a sawhorse, three dressmaking dummies, painted portraits, human organs pickled in jars, the chassis of a Model T Ford, tapestries, hundreds of yards of unused silks and fabric, clocks, 14 pianos (both grand and upright), a clavichord, two organs, banjos, violins, bugles, accordions, a gramophone and records, and countless bundles of newspapers and magazines. Oh and 130 tons of garbage.*o*

Rodinsky's room was also piled high with material. While it was not as overwhelming as the Collyers', when the door to 19 Princelet Street in London's Spitalfields was opened again in 1980 after over 11 years the redevelopers were met with material stuff: newspapers, books and papers, gramophone records, clothes and an A-Z marked with obscure journeys into the London suburbs, scraps of paper and sweet wrappers, all covered with indecipherable scribblings in many languages as well as a half-finished cup of tea and a pot of porridge still on the stove. What followed was another detective story as Rachel Lichtenstein pieced together the life and disappearance of David Rodinsky and Iain Sinclair traced his wanderings across London from the material objects he left behind.*o*

What unites these two stories is the way in which the Collyer brothers and David Rodinsky were positioned or even recreated as governmental subjects through their material objects, the rags 'n refuse they collected, hoarded or archived. They became targets of police reports, medical and mental health professionals as well as journalists, artists and writers who read their lives from their stuff and positioned them as subjects.*o*

Every twenty minutes Facebook adds more stuff to its collection:*o*

- 1 million links
- 1.4 million event invites
- 1.9 million friends requests accepted
- 2.7 million photos, 1.3 million of which are tagged
- 2.7 million messages sent
- 1.89 million status updates
- 1,6 million wall posts
- 10.2 million comments

This digital stuff is housed in at least 9 leased data centres or server farms, each around 35,000 square feet. Facebook is currently building its own centre which will be 307,000 square feet. These quaintly named 'farms' house 60,000 servers and cost in the order of $50m a year to run.*o*

Google is notoriously secretive about its hoard of data. What we do know is that it spent $757 million on its seven data centers in the third quarter of 2010 and that those centres process twenty petabytes of data a day. Google's hoard, like Facebook's includes our digital detritus – our email messages, our YouTube videos, our Picasa pictures and Blogger postings as well as 1 trillion cached webpages. Those farms also house the digital footprints we leave as we use Google's services – our logins, IP addresses, search terms and histories, maybe our creditcard details in Google checkout and records of the ads we clicked, the times and journies we made.*o*

Like the Collyers and Rodinsky, Facebook and Google hoard digital objects but unlike those real-world hoarders, the digital recluses also generate new data, new digital objects as they work. Their algorithms burrow through that data like a police patrolman or a researcher, tracing clues, forming connections, building pictures, but unlike those real-world investigators, Facebook and Google's algorithms create new data objects – connections between data files, between friends, searches and adverts, between activities and objects. And that new data is fed back into the archive, ready to be searched, found and connected again.*o*

What is important to note is that those data connections are also governmental data objects. Just as the Collyers' and Rodinsky's rags 'n refuse became pieces in constructing their subjectivity for media, law and social service systems, so the digital detritus we leave for Facebook and Google, and that they in turn generate from that rags 'n refuse, construct us as data objects and targets, 'friends' or demographics, healthcare risks or subversives. This goes beyond the issue of privacy of individual data objects to a wider field of governmentality through data trails and software-generated connections and subject positions. Even if our personal data is never released, even if we remain 'anonymous', the unhuman software patrolmen that burrow through the digital archives create a picture of us as part of a social graph or an aggregated search community. Whether these data subject positions are ever sold on to advertisers or insurance companies or subpoenaed by the state, they remain our social CV, our digital subjectivity. Whether those objects and traces are ever seen by human eyes is irrelevant, they remain data connections and data objects.*o*

As an example, Facebook is rolling out facial recognition where a photo added to the hoard will be processed, and suggested names or tags presented to the user who, on clicking on one, will of course create another data object just as she does when clicking on a Like button as Anne and Carolyn discussed yesterday. Only yesterday it was announced that Facebook will 'allow' users to tag Pages effectively tagging brands and objects. These data objects and trails, the photograph, the record of searching for, tagging or liking the photograph, can be seen as material in the sense which Jane Bennett talks of "vibrant matter", "quasi agents or forces with trajectories, propensities, or tendencies of their own". Here the magnetic traces on the data storage media that Matthew Kirschenbaum quite literally dissects in Mechanisms, whether what we commonly understand to be an object (e.g. the image file-object) or what we could call a weird object (the Like trace-object, the tag or tagging object/event), are in Bruno Latour's terms, actants, doing things in the world.*o*

I want to present some tentative thoughts towards an account of these data objects.*o*

Graham Harman's object-oriented philosophy draws on Bruno Latour and Alfred North Whitehead, positioning both as "philosophers of concrete, individual entities". The problem for Harman is that they do not go far enough. For Latour objects derive their power and presence from their relations or alliances. For Whitehead they are moments of becoming. For Harman any move away from a strict actualist focus on the object to either advocating a second realm of objects (the "eternal objects" of Whitehead) or a realm of potentiality beneath objects (the "plasma" as Latour speaks of it in Reassambling the Social) is a mistake. For Harman there are objects. That is it. Change happens in the world not as objects become and perish or enter new relations but as they connect. Connections are different than relations because whereas the latter happens outside the object – in a network, a "quasi-plenum" or realm of potential, the former is a matter of objects. There is no need in this framework for the object to perish or for the relations to be pushed to an outside. There is no need to take one's eye off the object-ball. Rather the flux or mesh of objects (the assemblages, media ecologies, networks, hyperobjects or whatever other term we use) can be addressed as a matter of the objects themselves.*o*

To bring this back to the digital hoard: The data-objects (the photos or credit card details); the data-mined objects (the Friends connection or clickthrough trail) and the datamining objects (the algorithms burrowing through and creating new data), all of these can be seen as objects circulating in and through Facebook and Google's archive-hoards. All three actualists would perhaps see those files, database entries and software agents as objects, entities in the world. Latour might see them as constituted by their relations with other actants in the network: hardware servers, other software, engineers and lawyers, company business plans and competition legislation. Whitehead might see them as a series of occasions, discrete instants of becoming and perishing, as occasions of data connection.*o*

Harman however would see them as objects that are not "exhausted by their relations to other objects", that withdraw from view and have an existence outside of their connections with other actants. This is his fourfold object of real object, real qualities, sensual objects and sensual qualities. Where Latour puts the emphasis on the network (relations) as what gives the Facebook wall photo or an algorithm its presence and its power and Whitehead would stress the transience of the Google image search, Harman would put the emphasis on these objects, as more than their relations, contexts and becomings.*o*

Some argue that Harman sees objects in isolation only, that he somehow refuses to entertain the idea that they connect. Far from it. Harman's whole philosophy is built on trying to understand how objects connect. He just wants to understand that connection as a matter of objects not of an exterior relationality. For Harman, objects encounter each other in the heart of another object. A real object encounters a sensual object in the heart of a new real object.*o*

We can draw the way that the photo on someone's Wall and the Facebook facial recognition algorithm connect at the level of objects rather than by recourse to some meta-framework of network or capitalism or globalisation. And this perhaps allows a new form of politics. We can talk of two

objects: the face-recognition algorithm and the photo uploaded to a user's Wall and Facebook's hoard. These objects could be seen as related within a field of network politics, info-capitalism and so governmentality. Alternatively they can be seen as connecting – however briefly – within the 'molten core' of an object we could call the "Facebook image data object", the key to Facebook's business plan, the site of governmental surveillance or subjectivity and the target of critical politics.*o*

In this perspective the real algorithm can never encounter the whole photo-object, it connects with only a dimension of that object, with a sensual object. Its connection in the moment of posting on the Wall does not exhaust what that photo or database-object is or does – the photo data-object has a position as object before and after, it is a site of other connections and workings and it is being constantly reconstituted as new searches are conducted around it, new datamine connections made or the data copied or moved from one server farm to another. At the same time the real photo-object encounters the facial-recognition algorithm (within the Facebook software). In fact it encounters an intentional image of the algorithm-object, a particular instantiation. It encounters a moment of running, a particular position of the code. It does not encounter the complete complex reality of that algorithm in terms of its history, the political, legal and business battles around its creation, or continuing work, its nature and other connections. It does not need to. It needs to encounter what is necessary to establish the database data point – the clue the governmental patrolman needs to establish subjectivity or Facebook needs to targets ads. This encounter happens not in some extra space or context, in Latour's plasma or Whitehead's space of becoming or even just between objects in trials of strength. It happens within another object, what we might call the "Facebook image data object", a real and specific instantiation of governmental power.*o*

In one sense this is a form of nested objects but it is important to emphasise that these are not nested in any hierarchical let alone value-laden sense. There is no sense in which objects connecting with other objects should be seen as leading to a foundational macro or micro object. This model not only refuses to leave the object but also refuses to find the single object. There is no Facebook-object or Surveillance-object or Capitalism-object that acts like 'context' or 'relation' as foundation for all connections. Nor is there some machine code-object or electrical charge-object that can stand in for a founding object or fundamental particle.*o*

Latour would at some level agree. He would of course insist that networks whether capitalism or the Twittersphere should be approached via objects or actants. The difference is that for Harman that can and should be done at the scale of objects not relations. Similarly Whitehead would perhaps argue for a focus on specific events not the abstract. But again the difference is that for Whitehead, as Harman says: 'actual entities "perpetually perish"'. They do not lie behind their accidents, qualities, and relations like dormant substrata, but are 'devoid of all indetermination'. For Harman this is ontologically problematic. For me it closes off potential for mapping objects, their connections and the spaces for exploit.*o*

The advantage of a non-relational object-oriented approach is (of course) fourfold: It not only allows an escape from macro/micro-reductionism but it also provides a way of escaping the problem of the subject. It allows us to talk of essences and technological determinism without a sneer; and finally it enables us to open up Exploits for intervention.*o*

Firstly this perspective escapes correlationism, Quentin Meillassoux's term for the tendency to see everything in terms of the human-world connection. From this perspective there is no world without the human nor human without the world. It is this separation (yet partnering) of subject and object that drags us away from focusing on objects, their connections and their working. In terms of data-objects, correlationism demands we address images, algorithms and the Facebook database in terms of the humans using or at least thinking about them. At the very least this means it becomes difficult to explore machine vision systems such as face-recognition where computers 'see', 'file' and 'analyse' with no human intervention, a situation an object-oriented approach could happily discuss in terms of a photo object connecting within face-recognition object within a surveillance-image-evidence object.*o*

Secondly, an object-oriented approach allows us an account of 'essence' that does not close off debate, connections, change and power. For Harman: "there is no avoiding a concept of essence". That does not mean that essences are eternal or natural. "To defend essence is not to conspire in a sinister plot by the Party of Reaction. It is nothing more than to insist that objects are not exhausted by the relations to other objects" he says. What we experience as essence is the outcome (or emanation as Harman calls it) of the tension between the object and its qualities. There are things about a table, a photograph or even an algorithm that are 'necessary' for it to be that table, photo or software that works. But these qualities are not identical with the object. They do not exhaust it. This is significant because it means we can talk of seemingly insubstantial data-objects such as searches or click throughs as things. We can say: "yes there is a data-mined object" and then trace its connections within objects. We can use that essence as a space for Exploit. An object-oriented essence is a starting point not an end.*o*

Even more controversially perhaps, this rescuing of 'essence' allows a similar embracing of 'technological determinism'. As Geoffrey Winthrop-Young puts it: "to label someone a technodeterminist is a bit like saying that he enjoys strangling cute puppies". An irreductionist, object-oriented reading of essence however allows us to say: "yes technology determines". The issue become how that determination is drawn in terms of causality or what DeLanda calls "catalysis" for instance. But leaving that debate to one side, the issue is that again an object-centred approach can explore determinations as connections within objects rather than as reflections of something more basic, foundational or powerful. It allows us to proudly and openly say that the connection between an image-file-object and the Facebook algorithm (within the Facebook image-object) does things.*o*

Finally, the escape from the subject, from the context, the relation, the continuum and the occasion – the focus on the object – allows a space for what Alexander Galloway and Eugene Thacker call the Exploit and it is here where an object-oriented approach to software and media meets politics. An approach to the computational/governmental space based on objects not networks or relations, changes the focus of struggle and change.*o*

For Galloway and Thacker, struggle operates at the level of objects – in their case protocol. Struggle "must not be anthropomorphic (the gesture, the strike); it must be unhuman (the swarm, the flood)". A virus does not fight a system, it overwhelms it. That struggle must be seen not as resistance but as "hypertrophy".*o*

Viruses or distributed denial of service (DDOS) attacks do not resist software they push it until it breaks. They clog up the server with too many requests, overloads, spam. But a DDOS attack can be seen as working not by simply overwhelming a network but by reconnecting objects (the http protocol, server requests, customers details etc) within the target object – in the case of recent Anonymous actions for instance, objects such as the PayPal or Amazon S3 object. Here an object-oriented approach of seeing and working with objects connecting within objects, rather than a field of relations, open up political potential.*o*

In a more constructive example perhaps, a focus on the code object not the whole Internet allowed the connecting of objects within the Apache server (object). This software object can be seen, and used as a model, as a reconfiguration of objects whereby new possibilities for server-client relations were released. The hackers who brought objects together as they created the (open source) code for the Apache server were working with and through objects in the creation of a new object. This Is object oriented programming as object oriented politics. A more recent example would be the Open Source Distributed Micro-blogging service thimbl.net, an attempt to connect software and social objects in a new configuration. The whole ecosystem of APIs is an equally interesting object-oriented space. Perhaps even sites and services that allow non coders to connect and mashup data and data objects whether that is Yahoo Pipes or WordPress, or even in it's own way Facebook itself are object spaces.*o*

An object-oriented approach allows one to see all the objects in play at the same scale in the computational/governmental topology. Here the photos I upload, the protocols that encode them, the data trails I leave, the proprietary iPad I create them on as well as the algorithms that position them and me – the whole governmental mix, are objects connecting within objects. The aim is not to trace relations external to those objects but connections within them. To move from understanding objects in terms of their relations is not to deny connections. Rather it is to place those connections – those governmental tunnels through the rags 'n refuse – front and centre, because they are issues of objects not issues of plasma or potential. Object-oriented approaches to the governmental mesh of the hoard allow us to deal with the unhuman objects of media and to address the connections that are made and can be made.*o*

To return finally to the Facebook and Google hoard-archives and the unhuman patrolmen who burrow through our rags 'n refuse, generating governmental positions as they go, an object-oriented approach to the Exploit offers new hope. Remaining true to a focus on objects and a flat ontology, rejecting relations as necessary to objects, it becomes possible to see how the data objects we willingly or unwillingly assign to Web 2.0 hoards are connected within those archives with others within governmental objects – the search-record object, the surveillance-object, the friend-object. These can be the target of Exploit. These are what can be reconfigured or realigned through new connections developed by new algorithms or software objects. The hoards may not be ours, the patrolmen burrowing through them may not be us, but that doesn't mean we can't find new ways through the rubbish.*o*

The read-along with pictures version is available here (1.8MB pdf) and if you're really keen you can see the video below... the advantage being you can see and hear my fellow panelists including

1-1-2011

Politics 2.0 with Facebook – Collecting and Analyzing Public Comments on Facebook for Studying Political Discourses

Chirag Shah
Rutgers University - New Brunswick/Piscataway, chirags@rutgers.edu

Tayebeh Yazdani nia
Rutgers University - New Brunswick/Piscataway, tyazdani@eden.rutgers.edu

First Submission: 01/28/2011

Revised Submission: 05/06/2011

Accepted:

RUNNING HEAD: POLITICS 2.0 WITH FACEBOOK

Politics 2.0 with Facebook – Collecting and Analyzing Public Comments on Facebook for Studying Political Discourses

Chirag Shah and Tayebeh Yazdani nia

School of Communication & Information

Rutgers, The State University of New Jersey

New Brunswick NJ 08901 USA

chirags@rutgers.edu, tyazdani@eden.rutgers.edu

Abstract

Analyzing publicly available content on various social media sites such as YouTube and Twitter, as well as social network sites such as Facebook, has become an increasingly popular method for studying socio-political issues. Such public-contributed content, primarily available as comments, let people express their opinions and sentiments on a given topic, news-story, or post, while allowing social and political scientists to extend their analysis of a political discourse to social sphere. We recognize the importance of Facebook in such analysis and present several approaches and observations of collecting and analyzing public comments from it. In particular, we demonstrate what it takes to do this manually, what we could learn from it, and how we can automate this process using a Facebook Harvester tool we have developed. In addition, we show how a hybrid approach can be formed giving us quick and easy data collection, and meaningful data analysis with substantially less effort than a manual approach. We believe these methods and tools will be highly valuable for political scientists in studying various political discourses as they take place in the Web 2.0 world.

Keywords

Facebook; Social media data extraction; Comments analysis; Political discourse.

Introduction

The public messages exchanged by social network site members, sometimes called comments or wall postings are a new type of text-based communication. These messages are unusual in that they are public – either world-visible or visible to all of a members' friends – and can be permanently associated with the identity of the poster, more directly and publicly so than listserv

postings (Thelwall, 2009). Due to the nature of these postings, that is, being public and available for harvesting,[1] they make an excellent choice for a social or political scientist to capture and study them as a proxy for people's voices for a given political discourse.

Thelwall (2009) argued that the widespread use of social network sites (SNS) in many countries (Boyd and Ellison, 2007) makes them an important object of study, and also gives an opportunity to investigate informal interpersonal communication on a larger scale than previously possible. In another study, Thelwall (2010) looked at the role of emotion in SNSs and whether emotion is typically reciprocated, and whether Friends express and/or receive similar levels of emotional expression to each other. His findings indicate statistically significant evidence for a weak correlation between the strength of positive emotion exchanged between Friends and received by Friends. This has larger implications on understanding how information propagates from person to person and source to source using SNS, and what they mean to the receiver of information (Shah, 2010).

In this workbench note, we take the problem of collecting and analyzing public comments from Facebook, one of the most popular SNSs, and demonstrate the pros and cons of various approaches. Specifically, we walk the reader through a series of steps that one may have to take for manually studying a political topic on Facebook, demonstrate a fully-automated process, and then a hybrid approach for collecting and analyzing public comments from Facebook. This note

[1] According to Facebook Statement of Rights and Responsibility: 2.4, "When you publish content or information using the 'everyone' setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information, and to associate it with you (i.e., your name and profile picture)."

provides a brief overview of the relevant literature, describes our methods, and their implications for studying political discourses. The system used for automatic and hybrid approaches is one of the crucial and sustaining contributions of this note.

Background

In the growing body of literature on SNSs, several articles have been published focusing on Facebook in particular (e.g., Mayer and Puller, 2008). These studies examine a diverse array of topics, from social capital (Ellison et al., 2007), to information disclosure (Gross and Acquisti, 2005), to temporal patterns in messaging (Golder et al., 2007).

More recently the Facebook data team (2010) looked at the usage of words in different "word categories" in status updates. This led them to discover some patterns in how people use status updates differently, and how their friends interact with different status updates.

Some research in this area has focused on the uses and gratifications of Facebook as Joinson (2008) does. Lampe et al. (2007) meanwhile explored the relationship between profile structure (namely, which fields are completed) and number of friends. Also, Ellison et al. (2007) examined the relationship between use of Facebook, and the formation and maintenance of social capital. They claimed that the site is tightly integrated into the daily media practices of its users; the typical user spends about 20 minutes a day on the site, and two-thirds of users log in at least once a day (Cassidy, 2006). Much of the existing academic research on Facebook has focused on identity presentation and privacy concerns (e.g., Gross & Acquisti, 2005; Stutzman, 2006). Looking at the amount of information Facebook participants provide about themselves, the

relatively open nature of the information, and the lack of privacy controls enacted by the users, Gross and Acquisti (2005) argued that users may be putting themselves at risk both offline (e.g., stalking) and online (e.g., identify theft). Other early Facebook research examined student perceptions of instructor presence and self-disclosure (Hewitt & Forte, 2006; Mazer, Murphy, & Simonds, 2007), temporal patterns of use (Golder, Wilkinson, & Huberman, 2007), and the relationship between profile structure and friendship articulation (Lampe, Ellison, & Steinfield, 2007).

Ringel et al. (2010) explored the phenomenon of using social network status messages to ask questions. They conducted a survey of 624 people, asking them to share the questions they have asked and answered of their online social networks. They presented detailed data on the frequency of this type of question asking, the types of questions asked, and respondents' motivations for asking their social networks rather than using more traditional search tools like Web search engines. They reported on the perceived speed and quality of the answers received, as well as what motivates people to respond to questions seen in their friends' status messages.

When studying the existing literature on Facebook, it becomes clear that there is a lack of research on extracting and analyzing public comments from Facebook. At the same time, collecting and studying such data could be highly valuable method for researchers looking at various political discourses. This motivated us to take on the task of looking at the comments that appear on various Facebook pages with a focus on studying people's comments and the discussions that follow. In other words, we are interested in the 'meaningful' discussions and debates that take place on Facebook and want to analyze these messages to be able to make sense of them.

Method

We started by asking ourselves – "how would one collect and analyze Facebook comments without any specialized support?" The answer to this is given in the following subsection as a step-wise procedure. We then describe a tool that we have developed to automatically collect a large number of Facebook comments and status updates. Finally, we show how we could combine this automated process for data collection to manual analysis and form a hybrid approach.

Manual data collection and analysis

In order to understand the process of analyzing public comments on Facebook, we visited numerous Facebook pages, and collected and analyzed data (comments) manually. Following are some of our experiences and observations that resulted from this process.

- The layout of most of Facebook pages is set up in a way that the administrator of the page posts a news piece or a bold statement, and that creates a discussions thread that can go on for a while, sometimes days after the original posting. For example, *Reform Immigration FOR America* (http://www.facebook.com/reformimmigrationforamerica) is a page that is dedicated to the issue of immigration. Their mission statement as it appears on their info page is: *"the U.S. immigration system no longer works. Fixing it presents a daunting challenge, but action must be taken sooner rather than later. The time is NOW to do the right thing and fight for practical solutions that benefit all of us and are rooted in the restoration of the rule of law, earned citizenship, united families, and fair treatment of workers."* We chose a wall post with a statement that was a quote from a senator who had spoken on the senate floor. We found 97 comments on this post that we

copied and pasted to a word file in a matter of seconds. Once all the comments were captured, we started going through each one and analyzing them based on relevancy, sentiments, objectivity, and the quality of messages. The objective of this analysis was to find "useful" comments that help us understand people's reactions to the wall post. Assessing each of the collected comments for these criteria took approximately 30-40 minutes.

- We noticed that in contrast to sports or entertainment pages on Facebook, political pages generate comments that are to the point and for the most part relevant to the given topic at hand. Comments on other pages are sometimes unrelated and completely irrelevant to the page topic. Pages that tackle social/political topics such as healthcare reform or child obesity also generate meaningful discussions that we are able to analyze thoroughly. For example, the popular music group Coldplay has a Facebook page[2] that is used for reaching out to their fans and informing them of their activities such as tour dates, etc. The layout is very much similar to the other Facebook pages. There are announcements about the band, and each announcement generates hundreds of comments. These comments are overwhelmingly centered around the fans' passion and devotion to the group. For instance, a link about a Christmas show in Liverpool[3] generated 1,310 comments and 18,960 people Liked it. The comments were transferred to a word file and analyzed using the same criteria mentioned before, which took nearly two hours. The majority of the comments were in English and although there were no *real* discussions

[2] http://www.facebook.com/coldplay/

[3] http://www.facebook.com/coldplay/posts/182958395063566

included, the great many of messages contained strong positive sentiments about the mentioned show.

- A good example for the social/political page is *Join the Coffee Party Movement* page on Facebook[4], which carries the same layout as the other pages. One of the links on this page that we examined was a statement made by the administrator, which has generated a great discussion thread, and 67 people had commented and 764 had Liked. By analyzing the messages (about 40 minutes), we found that a great majority of comments were relevant to the subject of statement and also the majority could be assessed based on sentiment. The sentiments were a mixture of negative/positive and the discussions were for the most part *meaningful*. There were agreements and strong disagreements with the statement and each created replies by the other commentators.

- When it comes to Facebook pages for corporations such as Starbucks, we found that there were fewer discussions and more open-ended opinionated comments that center on Starbucks as a brand. By looking at about 200 comments from this page (about 1 hour), we can see that the majority of commentators were expressing their love or loyalty for the brand and sharing stories that centered on Starbucks. We found no *meaningful* discussion threads, but only sentiments that were overwhelmingly positive towards the brand.[5]

[4] http://www.facebook.com/coffeeparty/posts/144485285605017

[5] Note that the sentiment analysis was done by a single individual, which may have biased such subjective judgment. For a more complete analysis, one may want to involve multiple coders in the process.

Given that we wanted to analyze a comment based on its objectivity, sentiment, relevance and other criteria, it would be simply hard to do this only via the automatic process. For example, if we are to determine whether a comment holds sentiments, we must read it to be able to decide whether it is negative/positive. Sometimes a comment is issued in a sarcastic tone and holds the opposite sentiment of what it appears to show. This can only be detected through manual data analysis. The same argument can be made for objectivity and relevance. The most significant disadvantage of this approach, however, is the great amount of time it takes to evaluate large collections of data. Several of our analyses took about one hour for less than 100 comments. Given that many interesting and important political topics generate thousands of comments, and that these comments keep coming constantly, it becomes prohibitively expensive to study many of these political discourses using public comments. The following subsection demonstrates how we could collect a large amount of data from Facebook and start analyzing it with very little effort.

Automatic data collection and analysis

We have developed a Facebook Harvester to quickly and effectively collect a large amount of data from a Facebook page. This data includes the status updates as well as the wall postings. The harvester uses newly introduced Facebook Open Graph APIs[6]. Figures 1 to 3 show the working of this web-based harvester using screenshots.

[6] http://developers.facebook.com/docs/opengraph

Figure 1: Main menu for the Facebook Harvester.

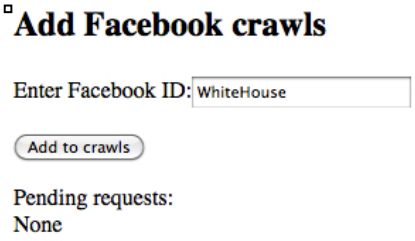


Figure 2: Starting a new harvesting process requires a Facebook page ID.

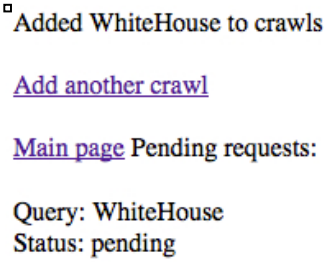


Figure 3: The harvesting process is now running in the background.

The result of running this harvesting process was thousands of messages (status updates and wall comments) within a few minutes. The data collected with these processes is stored in structured

format using MySQL. One could easily export this data in other structured formats, such as comma-separated values (CSV), or XML for further analysis. One could also run SQL queries on the MySQL database directly to filter, sort, and analyze the data.

We ran a similar harvesting process for a climate change group's Facebook page available at http://www.facebook.com/pages/Climate-Change/. The results of the data collection are shown in Figure 4 as a partial snapshot obtained using Sequel Pro.[7] We have also developed a front-end web-based interface to display the collected data from harvesting processes (Figure 5).

[7] http://www.sequelpro.com/

published	likes	comments_count	message	type
2010-05-29T16:03:55+0000	2	3	@ petite buenconsejo yes, i might have a re-write of our bataan camp bak...	status
2010-05-29T15:49:49+0000	0	3	Sayang di ka nakasama. baka meron pang susunod sama ka na. - Tito Andrei	status
2010-05-29T15:43:42+0000	2	0	Congrats sa successful clinic sa Bataan! -Dani â¥	status
2010-05-20T04:51:45+0000	0	0	To all the members of Climate Change, our rehearsals before the gig is at ...	status
2010-05-19T15:35:48+0000	3	0	See you tomorrow! Thursday, May 20. ;) SaGuijo Cafe + Bar. Makati :)	status
2010-05-17T13:41:46+0000	1	0	To all member's Of Climate Change please be reminded of our rehearsals t...	status
2010-05-03T15:41:01+0000	0	2	Paging Mary, please practice Nana song and Anesthesia for recording this week	status
2010-04-24T11:55:34+0000	0	0	Agree ko Pastor, I will talk to Master para matapus na natin to, para maka...	status
2010-04-24T07:15:20+0000	3	3	Napansin ko ang bilis dumami ng sumusuporta sa atin. Kailangan na talag...	status
2010-04-20T04:53:11+0000	1	4	Mga friends, maglaro tayo ng konti. Paki rate naman ninyo ang mga kanta ...	status
2010-04-17T18:18:52+0000	0	0	Congrats to "Climate Change" ang ganda ng tunog dahil kay master	status
2010-04-16T13:53:38+0000	3	2	Sa open air pala bukas ang gig...kasama ang rivermaya sa mga mag perfor...	status
2010-04-16T05:33:01+0000	6	0	I would like to welcome our new bassist. Gian Pineda, Welcome Gian to Cli...	status
2010-04-06T13:17:52+0000	0	0	Paging members of Climate Change Band, we have practice tomorrow sa G...	status
2011-01-04T10:49:44+0000	0	0	http://konsyltacii.com	wallpost
2011-01-04T10:48:30+0000	0	0	Prospective economics, energy management systems, new technology ma...	wallpost
2011-01-04T10:46:54+0000	0	0	Prospective economics, energy management systems, new technology ma...	wallpost
2011-01-04T05:28:46+0000	4	2	Coming Soon... >:)	wallpost
2011-01-02T04:26:38+0000	4	1	HAPPY NEW YEAR PO SA LAHAT! :)	wallpost
2011-01-01T14:06:35+0000	1	0	Climate Change will be performing tomorrow sa Gospel Jam together with ...	wallpost
2011-01-04T00:57:35+0000	4	2	Lagayan! Hehe! Happy New Year! =)	wallpost
2010-12-31T10:43:14+0000	1	0	Before the karaoke machines and fireworks go full-blast, a HAPPY NEW YE...	wallpost
2010-12-30T23:28:32+0000	0	1	klan na po labas ng album nio ?	wallpost
2010-12-30T19:26:23+0000	4	0	Happy New Year Everyone, Our call time on Gospel Jam on Sunday is 4;30pm.	wallpost
2010-12-25T01:12:53+0000	1	0	Maligayang Pasko po sa inyong lahat at maraming salamat po sa lahat ng ...	wallpost
2010-12-24T12:36:50+0000	3	0	Let's all take the time to remember that this holiday is not about receiving ...	wallpost
2010-12-24T07:09:53+0000	3	2	MERRY CHRISTMAS sa lahat ng Climate Changers and most especially sa m...	wallpost
2010-12-27T10:01:33+0000	4	3	Climate Change LIVE @ Mel & Joey	wallpost
2010-12-19T16:44:25+0000	0	0	Andrei Dionisio, Jr. on TV! â¥	wallpost
2010-12-19T14:52:07+0000	4	2	Salamat sa mga sumuporta sa guesting ni Hazel sa Mel & Joey. God bless u all!	wallpost
2010-12-19T12:45:43+0000	4	2	Mel & Joey na!!! Please watch. :-)	wallpost

Figure 4: A snapshot (back-end) of the data collected from the Climate Change Facebook page.

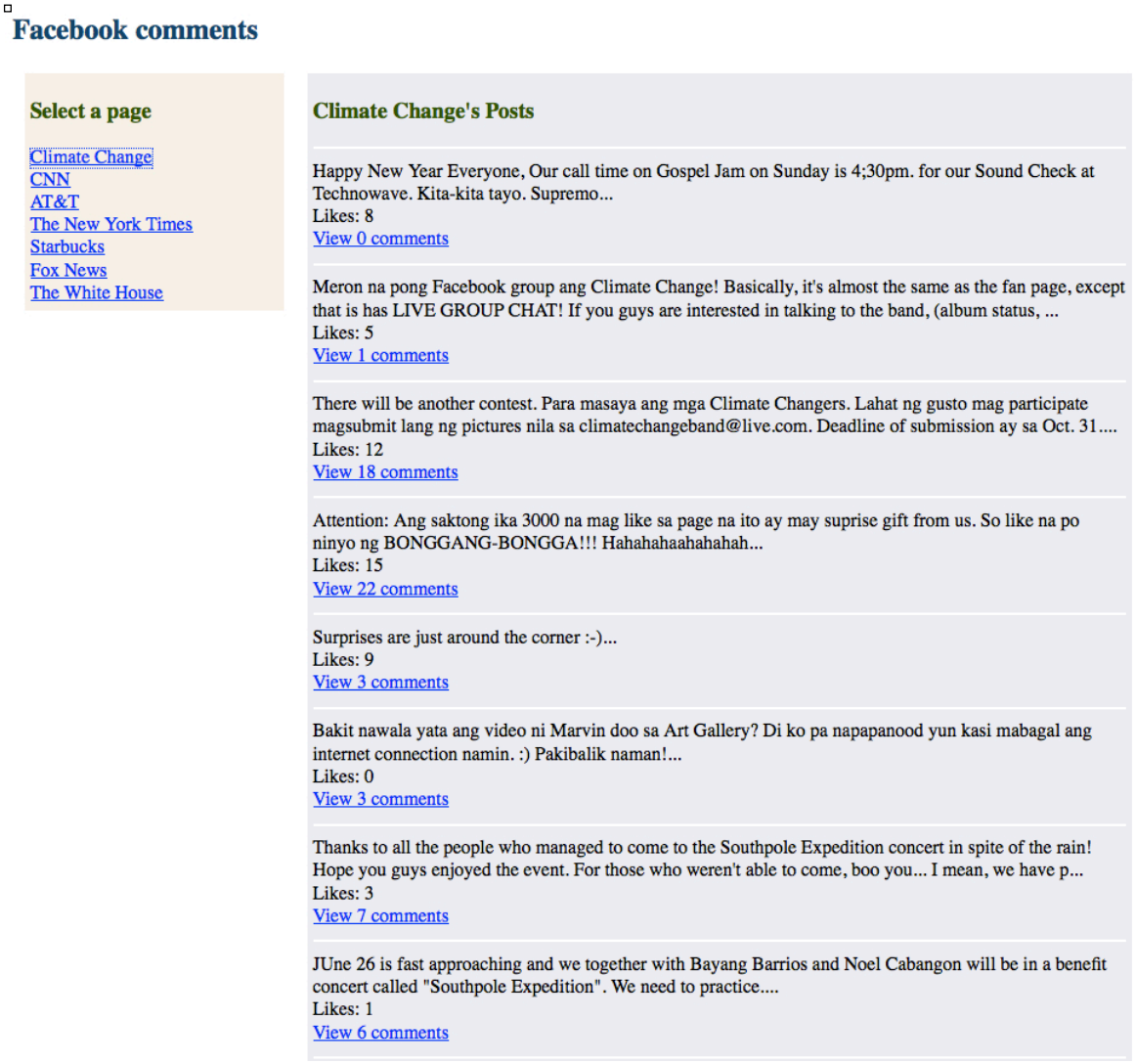


Figure 5: A snapshot (front-end) showing various harvesting jobs run and comments collected for 'Climate Change'.

Hybrid approach

At times and depending on the objective (whether it is sentiment analysis or relevance), it is both easier and more accurate to analyze the comments by simply reading them and going through them one by one. However, the manual approach may not be the most practical when it comes to

larger size data. When we are faced with thousands of comments, it can be difficult and we are looking to analyze them without the intention of investigating the details of each message, the automatic approach is our best solution.

Here we provide a hybrid approach, in which the data collection is done automatically and the analysis is facilitated with the help of sorting and filtering features of the system. Following on the same example of the previous subsection, we have collected a large number of comments from a climate change Facebook page using automated processes. This itself saved us enormous amount of time, but now we are left with thousands of these comments and it could take days to go through them.

To aid us in this process, we can use querying, sorting, and filtering using the Sequel Pro or a similar tool for database access and manipulation. This processing may differ depending on the objective of the. For instance, we may only want to look at messages that contain a certain word to see how many people are using a negative/positive term to address an issue and at the same time want to know the number of likes to those comments. Figure 6 shows a snapshot of the data with messages containing ‘climate’ word in them, and sorted by the number of comments to that message. This simple restructuring of the data was obtained by an SQL query taking only a fraction of a second to run. The obtained data is now a small subset (a few dozen messages), and more suitable and manageable for analyzing only the messages that explicitly talk about the climate, with the processing prioritized using the number of comments posted on a given message. This allows for a more thorough examination of discussions that take place on a given topic.

published	likes	comments_count	message	type
2010-10-27T07:38:57+0000	12	18	There will be another contest. Para masaya ang mga Climate Changers. La...	status
2010-06-03T03:28:31+0000	1	5	To all the members of Climate Change. we are all invited at Dani's party o...	status
2010-12-27T10:01:33+0000	4	3	Climate Change LIVE @ Mel & Joey	wallpost
2010-12-22T00:35:21+0000	4	3	Salamat po sa lahat ng patuloy na sumusuporta sa Climate Change pati na...	wallpost
2010-12-24T07:09:53+0000	3	2	MERRY CHRISTMAS sa lahat ng Climate Changers and most especially sa m...	wallpost
2010-12-18T17:39:30+0000	4	2	Malapit na ang album launching ng Climate Change...Isama po ninyo sa pa...	wallpost
2010-12-15T17:17:18+0000	4	2	Ilang tulog na lang available na ang first album ng Climate Change band. ...	wallpost
2010-11-01T15:46:13+0000	5	1	Meron na pong Facebook group ang Climate Change! Basically, it's almost ...	status
2010-05-20T04:51:45+0000	0	0	To all the members of Climate Change, our rehearsals before the gig is at ...	status
2010-05-17T13:41:46+0000	1	0	To all member's Of Climate Change please be reminded of our rehearsals t...	status
2010-04-17T18:18:52+0000	0	0	Congrats to "Climate Change" ang ganda ng tunog dahil kay master	status
2010-04-16T05:33:01+0000	6	0	I would like to welcome our new bassist. Gian Pineda, Welcome Gian to Cli...	status
2010-04-06T13:17:52+0000	0	0	Paging members of Climate Change Band, we have practice tomorrow sa G...	status
2011-01-04T10:48:30+0000	0	0	Prospective economics, energy management systems, new technology ma...	wallpost
2011-01-04T10:46:54+0000	0	0	Prospective economics, energy management systems, new technology ma...	wallpost
2011-01-01T14:06:35+0000	1	0	Climate Change will be performing tomorrow sa Gospel Jam together with ...	wallpost
2010-12-31T10:43:14+0000	1	0	Before the karaoke machines and fireworks go full-blast, a HAPPY NEW YE...	wallpost
2010-12-24T12:36:50+0000	3	0	Let's all take the time to remember that this holiday is not about receiving ...	wallpost
2010-12-19T16:44:25+0000	0	0	Andrei Dionisio, Jr. on TV! â¥	wallpost
2010-12-15T05:46:16+0000	0	0	Promote: Guys nood kau ng Mel And Joey sa linggo .. December 19 .. mag...	wallpost
2010-12-12T05:19:08+0000	0	0	Iwagayway Mo - Climate Change	wallpost

Figure 6: Filtered data filtered for messages with 'climate' in them, and sorted using the comments counts in the ascending order.

Another way of prioritizing message processing could be by using the length of a given message. We observed that more *meaningful* messages tend to be lengthier than those without useful or interesting critique. Figure 7 shows a snapshot of the results ordered by the message length.

published	likes	comments_count	message	type	msg_length
2011-01-04T10:46:54+0000	0	0	Prospective economics, energy management systems, new technology ma...	wallpost	591
2011-01-04T10:48:30+0000	0	0	Prospective economics, energy management systems, new technology ma...	wallpost	568
2011-01-04T10:49:44+0000	0	0	http://konsyltacii.com	wallpost	444
2010-11-01T15:46:13+0000	5	1	Meron na pong Facebook group ang Climate Change! Basically, it's almost ...	status	405
2010-12-24T12:36:50+0000	3	0	Let's all take the time to remember that this holiday is not about receiving ...	wallpost	391
2010-12-18T17:39:30+0000	4	2	Malapit na ang album launching ng Climate Change...Isama po ninyo sa pa...	wallpost	334
2010-12-24T07:09:53+0000	3	2	MERRY CHRISTMAS sa lahat ng Climate Changers and most especially sa m...	wallpost	330
2010-12-31T10:43:14+0000	1	0	Before the karaoke machines and fireworks go full-blast, a HAPPY NEW YE...	wallpost	319
2010-12-19T16:44:25+0000	0	0	Andrei Dionisio, Jr. on TV! â¥	wallpost	315
2010-10-27T07:38:57+0000	12	18	There will be another contest. Para masaya ang mga Climate Changers. La...	status	309
2010-12-25T01:12:53+0000	1	0	Maligayang Pasko po sa inyong lahat at maraming salamat po sa lahat ng ...	wallpost	272
2010-06-26T16:16:28+0000	3	7	Thanks to all the people who managed to come to the Southpole Expeditio...	status	269
2010-04-20T04:53:11+0000	1	4	Mga friends, maglaro tayo ng konti. Paki rate naman ninyo ang mga kanta ...	status	225
2010-04-24T07:15:20+0000	3	3	Napansin ko ang bilis dumami ng sumusuporta sa atin. Kailangan na talag...	status	186
2010-05-20T04:51:45+0000	0	0	To all the members of Climate Change, our rehearsals before the gig is at ...	status	184
2010-12-15T05:46:16+0000	0	0	Promote: Guys nood kau ng Mel And Joey sa linggo .. December 19 .. mag...	wallpost	174
2011-01-01T14:06:35+0000	1	0	Climate Change will be performing tomorrow sa Gospel Jam together with ...	wallpost	168
2010-04-16T13:53:38+0000	3	2	Sa open air pala bukas ang gig...kasama ang rivermaya sa mga mag perfor...	status	165
2010-06-12T10:04:34+0000	1	6	JUne 26 is fast approaching and we together with Bayang Barrios and Noel ...	status	162
2010-06-03T03:28:31+0000	1	5	To all the members of Climate Change. we are all invited at Dani's party o...	status	151
2010-04-06T13:17:52+0000	0	0	Paging members of Climate Change Band, we have practice tomorrow sa G...	status	149
2010-10-27T04:33:17+0000	15	22	Attention: Ang saktong ika 3000 na mag like sa page na ito ay may supris...	status	146
2010-06-28T15:11:04+0000	0	3	Bakit nawala yata ang video ni Marvin doo sa Art Gallery? Di ko pa napapa...	status	145
2010-12-30T19:26:22+0000	8	0	Happy New Year Everyone, Our call time on Gospel Jam on Sunday is 4;30pm.	status	133

Figure 7: Data ordered by the message length.

Implications and Future Work

We demonstrated how one could go about collecting public comments data from Facebook and analyzing them manually. We pointed out general observations and specific lessons learned using several examples. It was clear that such data can be a valuable asset for studying a political discourse, but very expensive without additional technology support. We then presented Facebook Harvester, a web-based tool we have developed to collect public comments and their attributes from a Facebook page. These comments include status updates and wall posts. Furthermore, we showed how such automated data collection could be combined with simple filtering to provide us significantly less expensive analysis methods. This can be extremely helpful in studying various socio-political issues. A couple of scenarios are presented below.

- *General scenario*. The White House posts constant updates, announcements, and contents on their Facebook page,[8] which includes pictures and videos. On a given post, there are typically few hundreds to few thousands comments posted by the visitors. One could not only collect these data once, but also keep collecting them at regular interval (e.g., daily) using our harvester. Such automated data collection and a few simple filtering could allow one to monitor White House's official stand on certain issues and people's opinions on them.

- *Specific scenario*. Starbucks recently rolled out a new brand logo, which created quite a bit of stir in the loyal fans and customers. Not surprisingly, they stared posting comments on Starbucks' Facebook page expressing their opinions and sentiments. Using our system, one could easily collect Starbucks status updates and wall comments from Facebook to study these opinions and sentiments, as well as Starbucks' own reactions to these comments.

The Facebook Harvester is available for public access and use for free under a Creative-Commons license at http://www.infoextractor.org/fbh/. We are currently working on extending this tool to allow collecting data from sites other than Facebook, such as CNN.com and CNNMoney.com, which incorporate Facebook wall on their posts or news-stories for people to comment using their Facebook credentials. We also plan on including other popular sites where people post comments on socio-political issues, such as nytimes.com.

[8] http://www.facebook.com/WhiteHouse

References

Boyd, D., Ellison, N. (2007), "Social network sites: definition, history, and scholarship", *Journal of Computer-Mediated Communication*, available at: http://jcmc.indiana.edu/vol2013/issue2001/Boyd.ellison.html, Vol. 13 No.1.

Cassidy, J. (2006, May 15). Me media. *The New Yorker*, 50–59.

Ellison, N.B., Steinfield, C., Lampe, C. (2007). The benefits of Facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12 (article 1).

Ellison, N. B., Steinfield, C. and Lampe, C. Charles Steinfield, A familiar face(book): profile elements as signals in an online social network. *Proceedings of the SIGCHI conference on Human factors in computing systems,* CHI '07

Ellison et al. (2007), The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication,* 12: 1143–1168. doi: 10.1111/j.1083-6101.2007.00367.x

Facebook Data Team (2010). What's on your mind? Retrieved from: http://www.facebook.com/notes/facebook-data-team/whats-on-your-mind/477517358858

Gross, R. and Acquisti, A. (2005). Information revelation and privacy in online social networks, *Proceedings of WPES'05. ACM* Alexandria, VA (2005), pp. 71–80.

Golder, S.A. ,Wilkinson D. and Huberman, B.A. (2007). Rhythms of social interaction:

messaging within a massive online network. In: C. Steinfield, B. Pentland, M. Ackerman and N. Contractor, Editors, *Proceedings of Third International Conference on Communities and Technology*, Springer, London, pp. 41–66.

Hewitt, A., & Forte, A. (2006, November). *Crossing boundaries: Identity management and student/faculty relationships on the Facebook.* Paper presented at CSCW, Banff, Alberta, Canada.

Joinson**,** A. N. Looking at, looking up or keeping up with people?: motives and use of facebook*, Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* , CHI '08.

Mayer, A. and Puller, S.L. (2008). The old boy (and girl) network: social network formation on university campuses, *Journal of Public Economics* 92, pp. 329–347.

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.

Ringel et al. (2010**).** What do people ask their social networks, and why?: a survey study of status message q&a behavior. *Proceedings of the 28th international conference on Human factors in computing systems.* CHI ’10.

Shah, C. (2010). Information derivatives – a new way to examine information propagation. *Proceedings of Workshop on Human Computer Interaction and Retrieval (HCIR) 2010*. August 22, 2010. New Brunswick, NJ.

Stutzman, F. (2006). *An evaluation of identity-sharing behavior in social network communities*. Paper presented at the iDMAa and IMS Code Conference, Oxford, Ohio.

Thelwall, M. (2009). MySpace Comments. *Online Information Review,* 33 (1), 58-76.

Thelwall, M. (2010). Emotion homophily in social network site messages. *First Monday* [Online], 15 (4). Available from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2897/2483.

Author Note

Chirag Shah

School of Communication and Information, Rutgers University

Chirag Shah is an Assistant Professor in the School of Communication and Information at Rutgers University. He received his PhD in Information and Library Science from the University of North Carolina (UNC) at Chapel Hill. He received his MS in Computer Science from UMass Amherst, where he worked with Bruce Croft and James Allan on high accuracy retrieval, and topic detection and tracking. At UNC, he worked with Gary Marchionini and Diane Kelly on various issues concerning exploratory information seeking and interactive information retrieval. He has also worked at many world-renowned research laboratories, such as FXPAL in California and National Institute of Informatics in Tokyo, Japan. His dissertation is focused on collaborative information seeking. He is also interested in social search and question-answering, digital preservation, and contextual information extraction. He has developed several tools for exploratory information seeking and extraction, including "Coagmento" for collaborative information seeking and the award-winning "ContextMiner" for capturing contextual information from multiple online sources.

Correspondence concerning this article should be addressed to Chirag Shah, Rutgers University, 4 Huntington St. New Brunswick, NJ 08901 or to chirags@rutgers.edu.

Tayebeh Yazdani nia

School of Communication and Information, Rutgers University

Tayebeh Yazdani nia is a student of Master in Library & Information Science (MLIS) program in the Dept. of Library & Information Science within the School of Communication & Information (SC&I) at Rutgers University. She is interested in studying social media usage for understanding various socio-political issues.

4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community

Michael S. Bernstein[1], **Andrés Monroy-Hernández**[1], **Drew Harry**[1], **Paul André**[2], **Katrina Panovich**[1] **and Greg Vargas**[1]

[1]Massachusetts Institute of Technology
Cambridge, MA 02139
{msbernst, amonroy, dharry, kp, gvargas}@mit.edu

[2]University of Southampton
SO17 1BJ, United Kingdom
pa2@ecs.soton.ac.uk

Abstract

We present two studies of online ephemerality and anonymity based on the popular discussion board /b/ at 4chan.org: a website with over 7 million users that plays an influential role in Internet culture. Although researchers and practitioners often assume that user identity and data permanence are central tools in the design of online communities, we explore how /b/ succeeds despite being almost entirely anonymous and extremely ephemeral. We begin by describing /b/ and performing a content analysis that suggests the community is dominated by playful exchanges of images and links. Our first study uses a large dataset of more than five million posts to quantify ephemerality in /b/. We find that most threads spend just five seconds on the first page and less than five minutes on the site before expiring. Our second study is an analysis of identity signals on 4chan, finding that over 90% of posts are made by fully anonymous users, with other identity signals adopted and discarded at will. We describe alternative mechanisms that /b/ participants use to establish status and frame their interactions.

Introduction

Identity representation and archiving strategies are central features to the design of online communities. However, our current understanding of them focuses mainly on strong identity and permanent archival. Researchers and practitioners argue that real names and pseudonyms can help "promote trust, cooperation, and accountability" (Millen and Patterson 2003), whereas anonymity may make communication impersonal and undermine credibility (Hiltz, Johnson, and Turoff 1986; Rains 2007). Influential industry players like Facebook argue that pseudonyms and multiple identities show "a lack of integrity" (Kirkpatrick 2010). Similarly, data permanence is also the norm: search engines will resurface content years after it is created (Rosen 2010), social network sites allow friends to browse updates and photos from years ago, and online communities will often expect newcomers to read their archives (Millen 2000). Some scholars have questioned these design approaches, suggesting that anonymous contributions and ephemeral participation online can be desirable (Lampe and Resnick 2004; Ren, Kraut, and Kiesler forthcoming; Grudin 2002). However, we have a limited understanding of how an anonymous and ephemeral community design might actually play out — especially at large scale.

In this paper we analyze one such large-scale, anonymous, and ephemeral community: the imageboard website 4chan. We focus on 4chan's first and most popular board, the "random" board known as /b/. Our goal is to use /b/ as a lens to understand the concepts of *anonymity* and *ephemerality* online. /b/ implements these concepts in more extreme ways than most other online communities. First, posts are fully anonymous by default and very rarely contain pseudonyms or other identity signals. This lack of identity makes traditional reputation systems unworkable. Second, instead of archiving conversations, /b/ deletes them when newer content arrives — often within minutes — which leads to a chaotic, fast-paced experience. By making complete anonymity and content deletion the norm, /b/ lets us study these concepts *in situ* at a larger scale than before.

This work quantifies the outcomes of 4chan's design decisions, and starts a discussion on how those decisions affect the community and its culture. Although /b/'s implementation of anonymity and ephemerality are extreme and unusual, the concepts themselves are not unique. Sites like Twitter, for instance, feel ephemeral because of their continuous content stream, while others like Formspring use anonymity as a core feature. By studying the impact of different points on the identity and archival continuums, we can broaden our understanding of community design strategies.

Readers may know of 4chan and /b/ (Figure 1) from their influence on Internet culture and media coverage of their off-site activities. The site boasts over seven million users (Poole 2010) and is a prolific meme factory: it originated popular memes like LOLcats (Rutkoff 2007) and rickrolling (Leckart 2009). Additionally, some 4chan and /b/ members have been known to participate in highly visible off-site activities. These activities include manipulating a Time Magazine poll to elect 4chan's creator the "World's Most Influential Person" and participating in hacktivist group 'Anonymous'. Anonymous has executed highly visible protests of the Church of Scientology (Coleman 2011) and DDoS attacks against Mastercard and Paypal in support of Wikileaks (Mackey 2010). Reactions are diverse: while memes have brought the site positive media attention (Brophy-

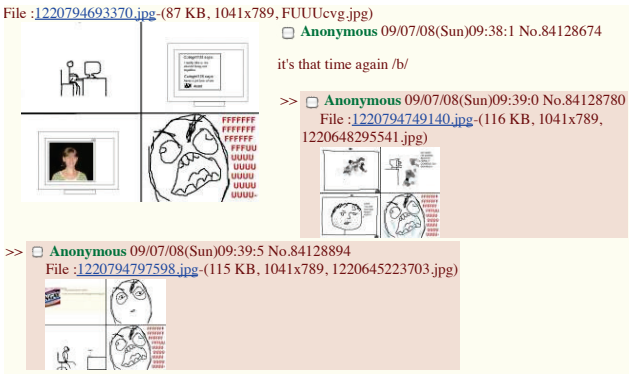


Figure 1: A "rageguy (FFFFUUUU) comic" themed thread on /b/. Source: 4chanarchive.org, where /b/ members save memorable threads.

Warren 2008), harassment and activism have often been been negatively and sensationally covered. For example, a Fox News affiliate called 4chan the "internet hate machine" (Shuman 2007). However, some media outlets have profiled the site in more nuanced ways (Dibbell 2009; 2010; Poole 2010). Much of this coverage is a response to /b/'s distinctive culture – a culture that merits critical analysis beyond the scope of this paper. Here, we focus on the board's design choices of anonymity and ephemerality, and how they may support or influence its culture.

In this paper, we perform a content analysis and two data-driven studies of /b/, focusing on anonymity and ephemerality. To begin, we survey related work. We then introduce 4chan, its design, and the /b/ board. To ground our discussion of the site, we perform a content analysis on a sample of /b/ threads. We then turn to our two studies: 1) ephemerality, tracking the site's tempo and content deletion dynamics; 2) anonymity, examining participant practices around identity. A note before proceeding: large portions of the 4chan site, and /b/ in particular, are offensive or obscene. We warn that quotes and vocabulary in this paper may offend.

Related Work

Our work builds on prior literature on anonymity and ephemerality. 4chan and /b/ can contribute insights into how this literature plays out in the wild at large scale.

Online communities choose points on the spectrum of anonymity — from completely unattributed to real names. For example, while Facebook embraces real names (Facebook 2010), Myspace does not (Dwyer, Hiltz, and Passerini 2007), and some Usenet boards allow posting from anonymous e-mail addresses (Donath 1999). Slashdot decided to enable anonymous commenting so users could feel more free to speak their minds, then controlled behavior with user moderation of comments (Lampe and Resnick 2004). However, while they may allow fully anonymous posting, anonymity is much less common in these communities than on /b/. Gómez et al (2008) found that fully anonymous posts made up only 18.6% of Slashdot comments. Instead, pseudonymity tends to become the norm as usernames allow members to build a reputation. Our work extends this research by studying the dynamics of a more thoroughly anonymous community.

Evidence is mixed on how anonymity may affect an online community. In many scenarios, researchers argue for the importance of identity-based reputation systems in promoting pro-social behavior (Millen and Patterson 2003). However, anonymity may foster stronger communal identity, as opposed to bond-based attachment with individuals (Ren, Kraut, and Kiesler forthcoming). Anonymity may impact participation: it increases equity in classrooms (Collins and Berge 1995), but results in more 'flaming' on e-mail lists (Thompsen and Ahn 1992). 4chan and /b/ play out these concepts on a larger stage than has been previously studied.

Computer-mediated communication has studied anonymity in small groups, and our work reconsiders their results in larger online communities. Removing traditional social cues can make communication impersonal (Short, Williams, and Christie 1976) and cold (Hiltz, Johnson, and Turoff 1986), and choosing to remain anonymous will undermine credibility (Rains 2007). However, we will argue that /b/'s community has developed alternative credibility mechanisms — via language and images — that still function effectively. Anonymity can also have positive outcomes: groups working anonymously and with critical confederates produce more ideas (Jessup, Connolly, and Galegher 1990); non-anonymous groups feel more personal, but have less overall cohesion (Tanis and Postmes 2007).

Ephemerality is rare in a large-scale online community, and to our knowledge, we are the first to study it directly *in situ*. Most communities that have been studied rely heavily on archives. For example, Millen (2000) reports that community members often expect each other to search group archives before asking new questions. Ephemerality may have community-wide downsides – a lack of history tends to decrease cooperation in social dilemma games (Fehr and Gächter 2000). However, instituting permanence in previously-unarchived chat rooms has elicited strong negative reactions (Hudson and Bruckman 2004).

Through our investigation of /b/, we hope to contribute to scholarly conversations about data permanence. For example, Grudin (2002) suggests that we evolved to live in an ephemeral world, yet our technology takes us from the "here and now" to the "everywhere and forever." Similarly, Mayer-Schonberger (2009) emphasizes the value of "societal forgetting," where "the limits of human memory ensure that people's sins are eventually forgotten." Blanchette and Johnson (2002) pointed to the recognition of social forgetfulness in three areas of social policy (bankruptcy law, juvenile crime records and credit reports), arguing that "data retention and disposal should be addressed as a fundamental characteristic of information."

These topics are not just of academic interest, but have clear practical implications for online social environments. Practitioners face similar challenges. For example, online game retailer Blizzard recently reversed a decision to require the use of out-of-game identities in its online forums (M.G. 2010), Formspring found both popularity and controversy by allowing teens to ask anonymous questions of their friends (boyd 2010a), and AOL angered its users when a hacker deanonymized old search logs (Barbaro and Jr. 2006).

Type	%	Description	Example
Themed	28%	Setting theme, often with an exemplar image.	*ITT ["In this thread"] we only post stuff we have laughed at so hard we had tears*
Sharing content	19%	Offering content for the community to enjoy or critique.	*This guy is a hero. :) http://www.youtube.com/[xxxxx]*
Question, advice and recommendation	10%	Asking for suggestions or, often quite intimate, life advice.	*Soup /b/ Recently I've been hanging out with a girl a lot, we're both in college. I spend the night at her place all the time and we kiss and whatnot. Problem is, she just broke up with her ex [...] and I know she's not over him. I really like her but I dunno what to do, so what do /b/?*
Sharing Personal Information	9%	Sharing or requesting content with personal information.	*U JELLY? ["You jealous?"] This is me suiting up at my formal looking fucking brilliant, then there is you fags sitting back and watching.*
Discussion	8%	Calling for discussion, debate or some back-and-forth over a topic.	*Hi Anonymous! So ive started this game called League of Legends a few days ago. [...] Is there anyone here who also plays it? Lets talk about it!*
Request for item	8%	Requesting information about previously-seen images, or other valuable information such as credentials for pay sites.	*anyone has a pic of that star wars battle tank with the german insignia shooped ["photoshopped"] on it? in return tits [a misogynous but common mechanism for "paying back" a favor with pornographic images]*
Request for action	7%	Intending to agitate for real-life action, like harassing another website	*Make a group saying [name] is awesome on face book. DO IT FAGGOTS*
Meta	5%	Discussing /b/ itself or playing with the site's mechanics (e.g., post numbers)	*Heidi Ho there /b/ I'm a Newfag and now that i've been here all Summer I was wondering if i need a letter of recommendation From a Registered OldFag?*
Other	6%	Unable to categorize.	*excuse all the blood*

Table 1: Content typology of threads on /b/: appearance frequency and an exemplar (some quotes are paraphrased). ($n = 598$)

4chan and /b/

4chan was created in 2004 by Christopher Poole as an online discussion board focused on Japanese anime (Sorgatz 2009). It has grown from its anime roots to encompass sixty boards on topics ranging from politics to fashion, science and "sexy beautiful women." Poole, better known by his pseudonym *moot*, created 4chan by copying the format of Futaba Channel, a popular Japanese discussion forum.[1] 4chan's aesthetic is simple, though it can appear confusing and cluttered: the Wall Street Journal describes it as "archaic [...] a quaint throwback to the earliest webpages" (Brophy-Warren 2008).

4chan is composed of boards, threads and posts. Each board is themed (e.g., /v/ is "Video Games"). Like most discussion boards, 4chan groups posts into threads (Figure 2). Posts starting a thread are required to include an image while images on replies are optional. Threads are organized into pages, where each page previews fifteen threads with their original post and a small sample of replies — users can click through to read the entire thread.

In this paper, we focus on the 4chan "random" board, known as /b/.[2] We focus on /b/ not only because it is 4chan's first and most active board — it claims 30% of all 4chan traffic — but also because, in the words of its creator, it is the "life force of the website", and the place where "rowdiness and lawlessness" happen (Sorgatz 2009).

Content ephemerality on 4chan is enforced by thread expiration and a large volume of incoming content. Threads begin on page one and are pushed down as new threads are added. If a user replies to a thread, it is *bumped* back to the top of the first page. If the thread reaches the bottom of the fifteenth page, the thread is removed permanently and its URL returns a 'Page Not Found' error. This entire process can take place over a matter of minutes, as Study 1 will demonstrate.

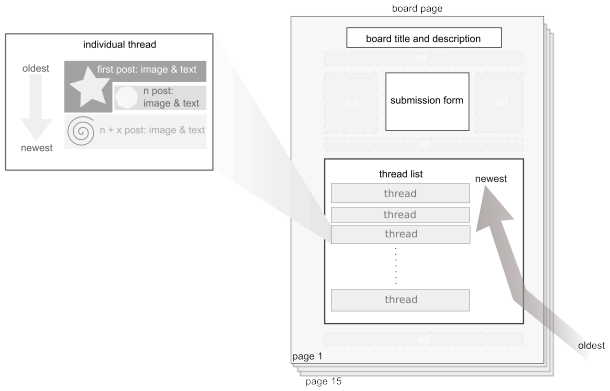


Figure 2: Structure of a 4chan board.

4chan's anonymity plays out through its posting mechanisms and defaults. Unlike other sites, where being anonymous usually means not using your "real" name or identity, most posts on /b/ are disconnected from *any* identity. There are no accounts; all information is entered on a per-post basis. If the user does not change the default empty name field, 4chan assigns the name *Anonymous*. Even if a user claims a pseudonym, any other user can claim it themselves in any subsequent post. 4chan does have a cryptographically-

[1] http://www.4chan.org/faq#what4chan
[2] http://boards.4chan.org/b/

powered feature for users to guarantee their identity, called a *tripcode*. Tripcodes use a password to generate a unique string after the username (e.g., *username!!Oo43raDvH61*), giving the password-holder a unique and inimitable identifier. However, 4chan users largely eschew tripcodes and pseudonyms, as Study 2 will show.

Content posted on /b/

/b/'s content is frequently intentionally offensive, with little held sacred. There is racist, sexist, homophobic language, groups are often referred to using a "fag" suffix (e.g., new members are "newfags", British users are "britfags"), and a common response to any self-shot picture by a woman is "tits or GTFO" (post a topless photo or get the f*** out). This language is part of the group identity: pushing the bounds of propriety in order to "hack the attention economy" and turn heads (boyd 2010b). While the content on /b/ can be offensive, it can also be funny, open, and creative, as its creation and promotion of numerous memes attests to.

In order to characterize the content and discourse on the board, we began with eight months of participant observation on the site. Using a series of informal samples of thread-starting posts on /b/, we conducted a grounded analysis (Charmaz 2006) similar to that which has been applied to other kinds of online participation (Naaman, Boase, and Lai 2010). After a few rounds of iteration, we settled on a scheme with nine high-level categories to describe the different kinds of posts that initiate threads on /b/. To measure the relative frequency of these thread categories, we collected a sample of 598 thread-initiating posts over the course of ten days (November 16–26, 2010). The threads were selected such that their temporal distribution matches the underlying distribution of 4chan posts. Our sample included the text and accompanying image from the post that started the thread.

Table 1 reports the thread composition in our sample. In keeping with /b/'s identity as an image board and the requirement that each thread-starter post an image, a common purpose of the board is to share images and web content. The two most prevalent thread types (Themed, 28%, and Sharing Content, 19%) both revolve around images and make up nearly half of all threads in our sample. There is also evidence for the off-/b/ activities that the media focuses on: threads attempting to organize such activities (Request for Action) make up 7% of the sample. Of those threads, the poster often attempted to generate comments on Facebook or Youtube pages or get people to call a particular phone number. /b/'s posters often disdain such calls because they are seen as self-serving, dismissing them by replying "/b/ is not your personal army." In the future, we plan to examine whether content category is associated with outcomes like thread length or reply frequency.

Study 1. Ephemerality

Ephemerality is one of /b/'s most striking qualities. The board moves at such a fast pace, and threads expire so quickly, that the site is largely different with each page refresh. In this section, we quantitatively describe the temporal properties and dynamics of posts on /b/. We compare /b/'s volume to other sites to provide a sense of turnover speed. We then relate how /b/ users have developed coping mechanisms like personal archives to cope with the quick expiration of material, as well as practices like "bumping" to keep threads alive. Finally, we explore how strict ephemerality policies may actually encourage increased community participation.

Method

We collected a dataset of activity on /b/ for two weeks: July 19–August 2, 2010. [3] This data includes 5,576,096 posts in 482,559 threads. Although there are likely some phenomena that influence /b/'s posts over longer time scales than two weeks (e.g. holidays), most of the major daily and weekly cycles are represented in a sample of this size. The dataset is missing a negligible number of posts during high-load periods. These missing posts are relatively randomly distributed and we do not believe that they impact our analysis. We did not capture images due to concerns over the nature of the material that may be posted to the website.

We calculated the time each thread spent on /b/ by replaying the history of all post events from our two-week dataset. We used creation timestamps as reported by the website to simulate the positions of each thread. For example, after a thread has a new reply, it moves to the top position on the first page; after a post is made in another thread, the first thread is pushed down to the second position. By replaying this history, we calculated the lifetime for each thread.

Results

Entire Lifetime The majority of threads have a short lifespan and a small number of replies; the median life of a thread is just 3.9 minutes. Thread lifetimes are right-skewed similar to a power law, making the mean less meaningful: 9.1 minutes (σ = 16.0 min). The fastest thread to expire was gone in 28 seconds (i.e., a thread with no responses during a very high activity period); the longest-lived lasted 6.2 hours (i.e., a thread with frequent new posts to bump it). Six hours is a very long time in /b/, but it is near-instantaneous when compared to the forever-archived nature of most other websites.

The longest-lived thread in our dataset was a discussion of paganism. The original poster was a pagan who advertised the opportunity to Ask A Pagan Anything, remarking "go on do your worst (or alternatively actually get my respect and actually ask something useful)." One question was "how do you worship your so called gods?", with the answer "From day to day just by reveling in the beauty and wonder of life, in all it's forms [sic] from studying martial arts to satifie [sic] my masculinity to taking care of my garden." Other questions included "How does it feel knowing Christianity raped your religion?", "What exactly do you worship?", and "Do you believe in magic?" Other long-lived threads fell in the "themed" category, like a "creepypasta thread", "info threads" (posting useful knowledge about some topic, like how to tie a tie or keyboard shortcuts), and self-shot nudity.

[3] More information about our data collection tool available at `http://projects.csail.mit.edu/chanthropology`

Short-lived threads on /b/ varied, but many were failed attempts to get the community's attention (e.g., "Well guyz, I hope you're glad the captcha is gone, woo yeah, let's all be random lolz!"). The short-lived posts often came in spurts during high-activity periods when it was easy to miss them.

As might be implied by short lifetimes, a large number of threads (43%) get no replies at all; the median is 2 posts per thread, the original post and one reply. This 43% figure is roughly consistent with Usenet, where 40% of posts get no reply (Joyce and Kraut 2006). Again, some threads become quite large, resulting in a mean of 13.27 posts/thread ($\sigma = 37.28$ posts, min = 1, max = 519).

First Page Only Another way to look at how ephemerality plays out in the board is by looking at each thread's exposure to the first page. The first page of results is where many items, like search results, get much of their overall visibility and click-throughs (Joachims et al. 2005).

The median thread spends just *5 seconds* on the first page over its entire lifetime. The mean time on the first page is 36 seconds ($\sigma = 109$ sec). The fastest thread was pushed off the first page in less than one second (actually, 58 of them shared this dubious honor), and the most prominent thread spent 37 minutes on the first page cumulatively over its lifetime.

The thread that spent the most time on the first page was a "roll" thread (a meta thread in our content analysis, playing with the mechanics of the board). In a roll thread, /b/ posters reply to get a 4chan-assigned post number (e.g., 1234567), and the last digit of their post number instructs them on an action to take. In this case, participants had to share personal secrets, a game that combines Spin the Bottle with Truth or Dare. Example responses included: "honestly, i dont know. Nothing has ever made me feel too terrified for my life. Maybe when I had taken too much cocaine and thought my heart was gonna overwhelm itself", "I gave into temptation by having two boyfriends at once instead of 1", "chose 9 i get angry at the idea of people having control over me and abusing it or manipulating me, a weird contrast to my mind control fetish", and "blue 5: my biggest regret is not asking her out".

User Control over Ephemerality: Bumping and Sage 4chan has developed two main ways for users to control thread ephemerality: *bumping* and *sage*. *Bumping* means replying to a thread to keep it alive, sometimes explicitly with a phrase like "bump", "bumping" or "bamp". In our sample, we observed that 2.16% of all replies contained these words or similar inflections. This is a lower-bound estimate, since any post will effectively "bump" a thread. The second method of control is *sage*, which allows a user to comment on a thread but not bump it to the first page (i.e., bury it). This lets users comment on a disliked thread without attracting attention to it, and to count the reply toward a system-enforced bump limit, thus ensuring the thread will expire more quickly. We found that 0.77% of all replies used the sage feature. So, /b/ posters will explicitly manipulate the ephemerality of some threads, though they seem more likely to promote threads than bury them.

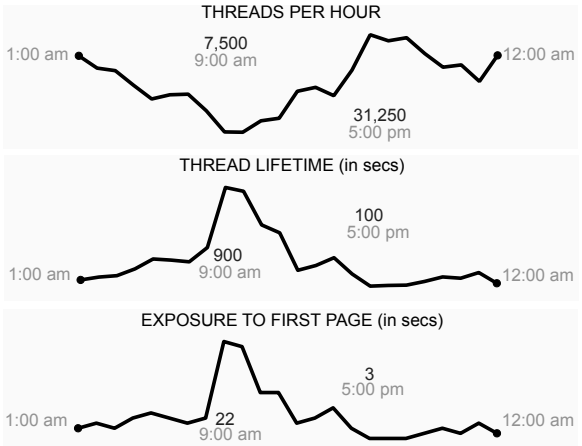


Figure 3: Daily board activity in our two-week dataset. Thread Lifetime and Exposure to First Page are medians; Threads per Hour averages the volume during that hour over fourteen days. All times in EST.

Time of Day /b/'s daily trends can help us understand how ephemerality is affected by time of day. The slower the traffic on /b/, the longer a thread will last.

Threads last the longest between 9am and 10am EST and expire fastest between 5pm and 7pm EST. High activity is sustained until 3am or 4am EST. This result suggests that, despite the not infrequent references to European and British users (e.g. "eurofags" and "britfags"), the demographics of /b/ are primarily North Americans that use the website after business or school hours. Figure 3 shows the synchronized spikes in lifespan and drop in number of posts per thread.

Discussion

/b/ creates a sense of ephemerality through a fast tempo and content deletion. In regards to tempo, we found that the board had roughly 35,000 threads and 400,000 posts per day. For comparison, Usenet volume (which still continues to grow) across all Big-8 newsgroups is 25,000 posts per day[4], or $1/16$th of /b/'s volume. Szabo and Huberman (2010) found that Digg has about 7,100 "threads" each day (1.3 million over six months), and 65,000 new YouTube videos each day. So, /b/ has roughly the same amount of posting activity as arenas like Usenet and Youtube, but *all of this activity happens in one forum board*.

Though the site may be ephemeral, /b/ users have developed other mechanisms to keep valuable content. For example, users often refer to having a "/b/ folder" on their computers where they preserve images for future enjoyment or remixing. /b/ posters ask others to dig into their archives; for example, this user wants to shock a friend, and donates an image of a cat in return: "Have a friend here, need the most fucked up shit you have in your /b folder, can't provide much, considering not on my comp, but here is a cat." /b/ users have also developed sites like 4chanarchive.org to save particularly important or "epic" threads.

Content deletion may play a role in pushing the /b/ community to quickly iterate and generate popular memes like LOLcats, Advice Dog and Archaic Rap. Having no history

[4] http://www.newsadmin.com/top100tmsgs.asp

moderates some the "rich get richer" phenomena (Barabási and Albert 1999). 4chan's founder has argued that /b/'s "lack of retention [...] lends itself to having fresh content," so only the fittest memes survive (Sorgatz 2009). To keep content around, users must make an explicit decision to save it to their hard drive and repost it later.

Finally, and perhaps unintuitively, ephemerality may raise community participation. One may think users would see no point to contributing if their actions will be removed within minutes. However, if /b/ users want to keep a thread from expiring within minutes, they need to keep conversation active. This "bump" practice, combined with a norm of quick replies, may encourage community members to contribute content. This hypothesis was derived from our observations, and will need to be tested more rigorously.

Study 2. Identity and Anonymity

/b/ makes it easy to participate without requiring a real or even pseudonymous identity. In this section, we investigate the frequency of the most common names used by /b/ posters, we discuss the impact of anonymity on /b/'s culture, and report on /b/'s alternative status and authenticity signals. We also discuss the effects of anonymity-fuelled disinhibition, like those seen in the "relationship advice" threads and "Anonymous" raids.

Method

Using our two-week data sample, we analyzed the identity metadata of each post. /b/ allows posters to enter no name ("Anonymous"), choose any name, or use a cryptographic identity mechanism known as a tripcode. We investigated the prevalence of each of these identity markers in our dataset.

Results

It is extremely uncommon to post using a name or pseudonym on /b/. In our sample, 90.07% (5,022,149) of posts were credited to the default name "Anonymous" (Table 2). The closest comparison available in the literature is that anonymous commenting makes up 18.6% of Slashdot comments (Gómez, Kaltenbrunner, and López 2008). The remaining 10% use a wide diversity of names. Some relate to an inside joke where many users claimed one name, David,[5] and others show mistaken uses of 4chan-specific keywords like *sage* or *noko*. Some users claimed to be "OP" (the original poster of the thread), demonstrating a way in which /b/ posters fluidly claim identity when needed.

E-mails are even less common (Table 3). Fully 98.3% (5,478,573) of posts in our sample did not contain an e-mail. Of those that did complete the e-mail field, 40.73% (39,725) are not actual emails but rather posts trying not to bump the thread (using the *sage* feature). The rest are misspellings of 4chan's special commands, some of them temporary. For example, ":stopsound:" got rid of the vuvuzela buzzing that 4chan administrators added to the page during the 2010 World Cup.

[5] http://encyclopediadramatica.com/Operation_/b/ipolar

Name	Num. of Posts	% of Posts
Anonymous	5,022,149	90.07%
David	59,320	1.06%
–	14,070	0.25%
OP	13,576	0.24%
(blank)	13,077	0.23%
sage	13,003	0.23%
anonymous	6,150	0.11%
noko	5,727	0.10%

Table 2: The most popular post names on /b/ during our two-week sample. Anonymous has over ninety percent of the posts ($n = 5,576,095$).

E-mail	Num. of Posts	% of Posts
(blank)	5,478,573	98.25%
sage	39,725	0.71%
Noko	5,037	0.09%
:stopsound:	3,377	0.06%
:soundoff:	1,627	0.02%

Table 3: The most common emails on /b/ during our two-week sample. Most posts leave the e-mail field blank ($n = 5,576,095$).

Tripcodes are the only way a 4chan user can guarantee that they are the same author of a previous post; however, they are very rarely used. Only 0.05% (281,367) of posts – one twentieth of one percent of our sample – contained a tripcode. Even this number may be inflated, because participating in the "David" in-joke mentioned above required using a shared tripcode among many users, which is uncommon. Ignoring the "David!4changtcqk" tripcode occurrences lowers the total to 0.04% (211,068).

Discussion

The usual narrative around anonymity suggests that communities benefit by revealing participants' names and reputations (Millen and Patterson 2003), and that anonymity will be a negative influence due to the "online disinhibition effect" (Suler 2005). Certainly, /b/ is a crude place and is given to antisocial behavior. Not only does anonymity invoke disinhibition on /b/, but styling the collective as "Anonymous" also suggests de-individuation and mob behavior. It may be safe for /b/ posters to act in a way they never would do offline because they can be relatively certain that their actions will not come back to haunt them.

However, the dynamics on 4chan and /b/ also suggest ways that anonymity can be a positive feature for communities. Disinhibition can be beneficial: in advice and discussion threads, anonymity may provide a cover for more intimate and open conversations. For example, in Table 1, the poster asks anonymously for advice about a potential girlfriend. Such threads are quite common. In addition, anonymity may encourage experimentation with new ideas or memes. As seen in Study 1, failure is quite common on 4chan: almost half of all threads receive no replies. Anonymity masks that failure, softening the blow of being ignored or explicitly chastised for trying to start uninteresting threads (Dibbell 2010). In communities with stronger identity mechanics, a history of poor posts would never go away. On 4chan, it is irrelevant.

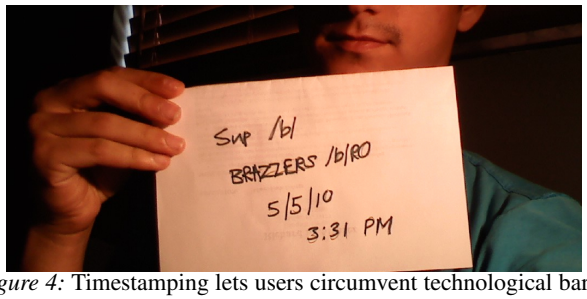


Figure 4: Timestamping lets users circumvent technological barriers to guarantee identity. Source: 4chanarchive.org.

/b/ has given rise to more fluid practices to signal identity and status in spite of, or perhaps because of, the lack of technological support. Because anyone can post a picture and claim to be that person, /b/'s posters have developed a practice of "timestamping" to guarantee authenticity. To claim identity, users often take a picture of themselves with a note containing the current day and time (Figure 4).

To communicate high status in the community, most users tend to turn to textual, linguistic, and visual cues. In many communities, including /b/, slang plays a role in delineating group membership (Eble 1996). Simply writing in 4chan dialect is non-obvious to outsiders and in-dialect writing serves as an entry-level signal of membership and status. Second, images on 4chan function a bit like fashion. Specific classes of images have periods of limited experimentation, turning into wider adoption, followed by subsequent abandonment. Fluency in the styles that are in vogue is an important way to signal status, as in fashion (Simmel 1957). Lack of fluency is dismissed with the phrase "LURK MOAR", asking the poster to spend more time learning about the culture of the board.

One example status signal in /b/ is the classic barrier for newcomers called "triforcing." Triforcing means leaving a post using Unicode to mimic the three-triangle icon of popular video game The Legend of Zelda:

Anonymous 08/30/09(Sun)17:59:38 No.156666179

▲
▲ ▲

Newcomers will be taunted by a challenge that "newfags can't triforce." Uninitiated users will then copy and paste an existing triforce into their reply. It will look like a correct triforce in the reply field; however, after posting, the alignment is wrong:

Anonymous 07/03/10(Sat)10:32:29 No.247423XXX

sweet im not a newfag anymore

▲
▲ ▲

The only way to display high status and produce a correct triforce on 4chan is to use a complicated series of Unicode character codes. In signaling theory terms, we can think of triforcing as an index: a signal whose presentation is only possible by someone with particular skill or knowledge (Smith and Harper 1995).

That communities enforce boundaries and communicate status using language and differentiated social practices is perhaps not surprising in the social psychology literature. What we see as particularly noteworthy in this case is how these boundary-sustaining practices are informed by the technical context in which they take place. Furthermore, the extreme nature of community practices on /b/ can obscure the underlying role these behaviors play to the casual observer. We see our role as partly one of translating these practices into terms familiar to scholars.

Conclusion

In this article we investigate the 4chan /b/ board as a vehicle for understanding the effects of ephemerality and anonymity in online communities.

Analyzing ephemerality via two weeks of of site activity, we found that the median thread spends just five seconds on /b/'s first page before being pushed off by newer posts, and that the median thread expires completely within five minutes. Even in a world informed by Twitter and newsfeeds, where content is out of users' attentional sphere quickly, we argue that such rapid content *deletion* drives many of /b/'s community dynamics. On /b/, ephemerality and deletion create a powerful selection mechanic by requiring content the community wants to see be repeatedly reposted, and potentially remixed. We believe this is critical to the site's influence on internet culture and memes.

We then examined anonymity on /b/. We found that over 90% of posts are made completely anonymously, and just one twentieth of one percent of posts use system mechanisms like tripcodes to guarantee identity. Instead, the /b/ community uses non-technical mechanisms like slang and timestamping to signal status and identity. Consistent with common identity theory (Ren, Kraut, and Kiesler forthcoming), /b/'s anonymity is likely shaping a strong communal identity among a very large set of individuals.

We hope to open the door to future 4chan work. 4chan is widely credited with being the source of many online memes, making it an excellent venue for studying innovation diffusion. By tracking images (and genres of images) we could see how trends spread through 4chan and into the wild. It is also clear that 4chan represents only one part of a larger online ecosystem. Future work might focus on how 4chan's users move between other tools and interaction venues to organize both online and offline action. Finally, a closer study of the content on 4chan and its users would enable us to make more substantial claims about the relationship between 4chan's design and its users' practices and culture.

As large Internet players like Facebook or Google evolve their models for identity and archiving, it becomes increasingly important to understand what happens in large communities that occupy the opposite positions on the user identity and data permanence design continuums. Communities like 4chan have immense impact on Internet culture, and /b/'s anonymous, ephemeral community design is playing a strong role in that cultural influence.

Acknowledgements

We would like to thank for their feedback: danah boyd, Amy Bruckman, Gabriella Coleman, Danyel Fisher, Benjamin Mako Hill, Alex Leavitt, Rob Miller, Lisa Nakamura, Chris Schmandt, Christina Xu and Sarita Yardi.

References

Barabási, A., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286(5439):509.

Barbaro, M., and Jr., T. Z. 2006. A Face Is Exposed for AOL Searcher No. 4417749. *New York Times*.

Blanchette, J., and Johnson, D. G. 2002. Data retention and the panoptic society: The social benefits of forgetfulness. *The Information Society: An International Journal* 18(1):33.

boyd, d. 2010a. Harassment by q & a: Initial thoughts on formspring.me. http://dmlcentral.net/blog/danah-boyd/harassment-qa-initial-thoughts-formspringme.

boyd, d. 2010b. "for the lolz": 4chan is hacking the attention economy. http://www.zephoria.org/thoughts/archives/2010/06/12/for-the-lolz-4chan-is-hacking-the-attention-economy.html.

Brophy-Warren, J. 2008. Modest web site is behind a bevy of memes. *Wall Street Journal*.

Charmaz, K. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London: Sage Publications.

Coleman, G. 2011. "'I did it for the lulz! but i stayed for the outrage:' anonymous, the politics of spectacle, and geek protests against the church of scientology.". http://techtv.mit.edu/videos/10237.

Collins, M., and Berge, Z. 1995. Introduction: Computer-mediated communications and the online classroom in higher education. In Berge, Z., and Collins, M., eds., *Computer mediated communication and the online classroom: Vol. 2. Higher education*. Hampton Press. 1–10.

Dibbell, J. 2009. The Assclown Offensive: How to Enrage the Church of Scientology. *Wired*.

Dibbell, J. 2010. Radical Opacity. *Technology Review*.

Donath, J. 1999. Identity and Deception in the Virtual Community. In Kollock, P., and Smith, M., eds., *Communities in Cyberspace*. London: Routledge.

Dwyer, C.; Hiltz, S.; and Passerini, K. 2007. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In *Proc. of AMCIS*.

Eble, C. 1996. *Slang & sociability: In-group language among college students*. The University of North Carolina Press.

Facebook. 2010. Statement of rights and responsibilities. http://www.facebook.com/terms.php.

Fehr, E., and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *The Am. Econ. Rev.* 90(4):980–994.

Gómez, V.; Kaltenbrunner, A.; and López, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proc. of WWW*, 645–654. New York, NY, USA: ACM.

Grudin, J. 2002. Group dynamics and ubiquitous computing. *Commun. ACM* 45(12):74–78.

Hiltz, S.; Johnson, K.; and Turoff, M. 1986. Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research* 13(2):225–252.

Hudson, J. M., and Bruckman, A. 2004. "Go away": Participant objections to being studied and the ethics of chatroom research. *The Information Society: An International Journal* 20(2):127.

Jessup, L.; Connolly, T.; and Galegher, J. 1990. The effects of anonymity on GDSS group process with an idea-generating task. *MIS Quarterly* 14(3):313–321.

Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of SIGIR*, 154–161.

Joyce, E., and Kraut, R. 2006. Predicting continued participation in newsgroups. *J. of Computer-Mediated Comm.* 11(3):723–747.

Kirkpatrick, D. 2010. *The Facebook Effect*. Simon & Schuster.

Lampe, C., and Resnick, P. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proc. of CHI*, 543–550.

Leckart, S. 2009. The Official Prankonomy: From rickrolls to malware, a spectrum of stunts. *Wired* 17(9):91–93.

Mackey, R. 2010. 'Operation Payback' Attacks Target MasterCard and PayPal Sites to Avenge WikiLeaks. *New York Times*.

Mayer-Schonberger, V. 2009. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton University Press.

M.G. 2010. A Blizzard of protest over privacy. *Economist*.

Millen, D. R., and Patterson, J. F. 2003. Identity disclosure and the creation of social capital. In *Proc. of CHI*, 720–721.

Millen, D. R. 2000. Community portals and collective goods: Conversation archives as an information resource. *HICSS*.

Naaman, M.; Boase, J.; and Lai, C. 2010. Is it really about me?: message content in social awareness streams. In *Proc. of CSCW*, 189–192.

Poole, C. 2010. The case for anonymity online. *TED2010*.

Rains, S. 2007. The impact of anonymity on perceptions of source credibility and influence in computer-mediated group communication: A test of two competing hypotheses. *Communication Research* 34(1):100.

Ren, Y.; Kraut, R. E.; and Kiesler, S. forthcoming. Encouraging commitment in online communities. In Kraut, R., and Resnick, P., eds., *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press.

Rosen, J. 2010. The web means the end of forgetting. *The New York Times*.

Rutkoff, A. 2007. With 'LOLcats' Internet Fad, Anyone Can Get In on the Joke. *Wall Street Journal* 25.

Short, J.; Williams, E.; and Christie, B. 1976. *The social psychology of telecommunications*. Wiley London.

Shuman, P. 2007. Fox 11 investigates: 'anonymous'. http://www.youtube.com/watch?v=DNO6G4ApJQY.

Simmel, G. 1957. Fashion. *Am. J. of Soc.* 62(6):541–558.

Smith, M., and Harper, D. 1995. Animal signals: models and terminology. *Journal of Theoretical Biology* 177(3):305–311.

Sorgatz, R. 2009. Macroanonymous is the new microfamous. http://fimoculous.com/archive/post-5738.cfm.

Suler, J. 2005. The online disinhibition effect. *International Journal of Applied Psychoanalytic Studies* 2(2):184–188.

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Commun. ACM* 53:80–88.

Tanis, M., and Postmes, T. 2007. Two faces of anonymity: Paradoxical effects of cues to identity in CMC. *Computers in Human Behavior* 23(2):955–970.

Thompsen, P. A., and Ahn, D. 1992. To be or not to be: An exploration of e-prime, copula deletion and flaming in electronic mail. *Et Cetera: A Review of General Semantics* 146–164.

Exploring Millions of Footprints in Location Sharing Services

Zhiyuan Cheng, James Caverlee, Kyumin Lee
Texas A&M University
College Station, TX 77843
{zcheng, caverlee, kyumin}@cse.tamu.edu

Daniel Z. Sui
Ohio State University
Columbus, OH 43210
sui.10@osu.edu

Abstract

Location sharing services (LSS) like Foursquare, Gowalla, and Facebook Places support hundreds of millions of user-driven footprints (i.e., "checkins"). Those global-scale footprints provide a unique opportunity to study the social and temporal characteristics of how people use these services and to model patterns of human mobility, which are significant factors for the design of future mobile+location-based services, traffic forecasting, urban planning, as well as epidemiological models of disease spread. In this paper, we investigate 22 million checkins across 220,000 users and report a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with these footprints. We find that: (i) LSS users follow the "Lèvy Flight" mobility pattern and adopt periodic behaviors; (ii) While geographic and economic constraints affect mobility patterns, so does individual social status; and (iii) Content and sentiment-based analysis of posts associated with checkins can provide a rich source of context for better understanding how users engage with these services.

1 Introduction

In many ways analogous to the sensor systems embedded in the physical environment of planet earth, emerging real-time social systems are rapidly creating a web of social sensors that can potentially be used as sociometers to gauge diverse social indicators ranging from political views to consumer tastes to public opinions about key social issues to the mood of people at particular places and times. In practice, highly-dynamic real-time social systems like Twitter, Facebook, and Google Buzz have already published exabytes of real-time human sensor data in the form of status updates. Coupled with growing location sharing services like Foursquare, Gowalla, Facebook Places, and Google Latitude, we can see unprecedented access to the activities, actions, and trails of millions of people, with the promise of deeper and more insightful geospatial understanding of the emergent collective knowledge embedded in these activities and actions.

In terms of scale, the Foursquare service alone claims over 6 million registered users (Foursquare 2011) and nearly 1 million check-ins per day (Grove 2010). Like similar services, Foursquare allows users to "check in" at different venues (e.g., grocery stores, restaurants), write tips, and upload pictures and videos.[1] As in other online social networks, Foursquare users can make friends with each other, and monitor their friends' status and location. While users of Foursquare and related location sharing services may not be a representative cross-section of the whole human society, the data revealed through these services provides a fascinating and unique opportunity to study large-scale voluntarily contributed human mobility data, which could impact the design of future mobile+location-based services, traffic forecasting, urban planning, and models of disease spread.

Toward understanding the spatial, temporal, and social characteristics of how people use these services, we present in this paper a large-scale study of location sharing services. Concretely, we study the wheres and whens of over 22 million checkins across the globe. We study human mobility patterns revealed by these checkins and explore factors that influence this mobility, including social status, sentiment, and geographic constraints. To the best of our knowledge, this is the first effort to utilize location sharing services to study human mobility patterns and the corresponding factors which can affect mobility patterns.

2 Related Work

The role of geography and location in online social networks has recently attracted increasing attention. Facebook researchers analyzed the distance between Facebook users' social relations, and utilized locations of a user's friends' to predict the user's geographical location (Backstrom, Sun, and Marlow 2010). (Cheng, Caverlee, and Lee 2010) modeled the spatial distribution of words in Twitter's user-generated content to predict the user's location. Characterizing network properties in relation to local geography is studied in (Yardi and Boyd 2010). User behavior with regard to the location field in Twitter user profiles has been studied in (Hecht et al. 2011). (Lindqvist and others 2011) analyzed how and why people use location sharing services, and discussed the privacy issues related to location sharing services. Besides locations, researchers have also explored temporal dynamics associated with on-line social activities (Golder, Wilkinson, and Huberman 2007).

[1] According to the Wall Street Journal, "check in" was the 12th most popular word of 2010 (Cholera 2011)

Table 1: Distribution of Sources of Checkins

Name	Percentage
Foursquare	53.5%
UberTwitter	16.4%
Twitter for iPhone	10.2%
Twitter for Android	3.4%
TweetDeck	3.1%
Gowalla	2.9%
Echofon	2.0%
Gravity	1.3%
TwitBird	1.1%
Others	6.0%

Figure 1: Global Distribution of Checkins

Analyzing and modeling mobility patterns has long attracted attention by researchers in fields like statistical physics, ubiquitous computing, and spatial data mining. For example, an analysis of 100,000 cellphone users' trajectories (Gonzalez, Hidalgo, and Barabasi 2008) showed that human mobility displayed simple reproducible patterns. The authors of (Brockmann, Hufnagel, and Geisel 2006) analyzed the circulation of bank notes in the US and concluded that human traveling behavior can be described mathematically on many spatio-temporal scales by a two parameter continuous time random walk model. A 93% potential predictability in user mobility was found across 50,000 cellphone users in (Song et al. 2010). (Zheng and others 2009) proposed a system to mine interesting locations and travel sequences from users' GPS trajectories. Researchers of (Humphries and others 2010) observed Lèvy Flight search patterns across 14 species of marine predators, with a few individuals switching between Lèvy Flight and Brownian motion as they traversed different habitat types.

Different from cellphone data and trajectories derived from GPS trackers, checkins have several unique features: (i) they are inherently social, since users reveal their location to their friends, meaning that social structure and its impact on human mobility can be directly observed; (ii) checkins are associated with particular venues (e.g., a restaurant), allowing for greater analysis of venue type; (iii) checkins can be augmented with short messages, providing partial insight into the thoughts and motivations of users of these services.

3 Gathering Checkins

To begin our study, we first require a collection of checkins. Since personal checkin information on location sharing services like Foursquare, Gowalla, and Facebook Places is typically restricted to a user's immediate social circle (and hence unavailable for sampling) we take an approach in which we sample location sharing status updates from the public Twitter feed. Twitter status messages support the inclusion of geo-tags (latitude/longitude) as well as support third-party location sharing services like Foursquare and Gowalla (where users of these services opt-in to share their checkins on Twitter). We monitor Twitter's gardenhose streaming API ($\sim$1% of the entire Twitter public timeline), and retrieve users who post geo-tagged status updates. For each sampled user, we crawl up to a maximum of the most recent 2,000 geo-labeled tweets.

The location crawler ran from late September 2010 to late January 2011, resulting in a total collection of 225,098 users and 22,506,721 unique checkins. The 22 million checkins were posted from more than 1,200 applications, and the distribution of sources is displayed in Table 1. More than 53% of the checkins are from Foursquare, and most of the other checkins are from Twitter's applications on mobile platforms like Blackberry, Android, and iPhone. A few hundred thousands checkins are from other location sharing services like Gowalla, Echofon, and Gravity.

Format of the Data: Each checkin is stored as the tuple *checkin(userID, tweetID)* = {*userID, tweetID, text, location, time, venueID*}. An example checkin tuple is: checkin(14091113, 9710376274) = {14091113, 9710376274, "I'm at MTA - Atlantic Ave-Pacific St Subway Station. http://4sq.com/2nWVD0", 40.685307, -73.980719, "2010-02-26 21:42:04", "cd979d2e352c4f54"}. We additionally store a user as the tuple: *user(userID)* = {*userID, status_count, followers_count, followings_count*}; for the example checkin, the user has 2,771 total status updates, 255 followers and is following 926 users.

Filtering Noise: Many location sharing services provide some mechanism to verify that a user is actually at or near the venue where they are checking in (e.g., by cross-checking with a user's cellphone GPS) (Foursquare 2010), however, there can still be incidents of false checkins. Hence, we additionally filter out all checkins from users whose consecutive checkins imply a rate of speed faster than 1000 miles-per-hour (or faster than an airplane). In total, we filtered 294 users (0.1%) with sudden moves, yielding a final collection of 224,804 users and 22,388,315 checkins. More than 72% users have fewer than 100 checkins; 7.8% users have more than 300 checkins; and 3.6% users have more than 500.[2]

Locating Each User's "Home": Some of the analysis in the following sections requires that we first associate each user with a natural "home", so, for example, we can compare the properties of all users "from" New York City versus users "from" Los Angeles. Since users of location sharing services are not required to register a home location, we must algorithmically determine the home location. Note that choosing a user's home based on the center of mass of all

[2]Data are available at http://infolab.tamu.edu/data/

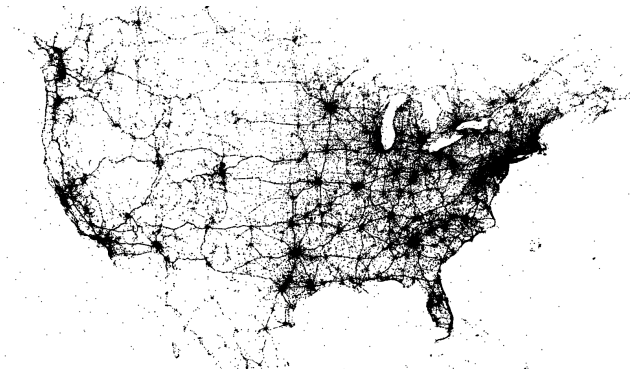


Figure 2: Detail: Checkins in the United States

checkins suffers from splitting-the-difference, by placing a user from Houston who occasionally travels to Dallas somewhere in between the two cities; alternatively, directly considering the user's most frequently checked-in venue may overlook a cluster of closely-located but less individually checked-in venues. To avoid these drawbacks, we propose a simple method to geo-locate a user's home based on a recursive grid search. First, we group checkins into squares of one degree latitude by one degree longitude (covering about 4,000 square miles). Next, we select the square containing the most checkins as the center, and select the eight neighboring squares to form a lattice. We divide the lattice into squares measuring 0.1 by 0.1 square degrees, and repeat the center and neighbor selection procedures. This process repeats until we arrive at squares of size 0.001 by 0.001 square degrees (covering about 0.004 square miles). Finally, we select the center of the square with the most checkins as the "home" of the user.

4 Spatio-Temporal Analysis of Checkins

In this section, we begin our study of large-scale location sharing services with an investigation of the temporal and geographic characteristics of how people use these services.

4.1 Wheres of the Checkins

First, we plot the locations of the 22 million checkins in Figure 1, where we see that while checkins are globally distributed, the density of checkins is highest in North America, Western Europe, South Asia, and Pacific Asia. Zooming in on the US, Figure 2 shows the reach of location sharing services, revealing the boundaries of cities and the lines of highways. Further zooming in, we can see in Figure 3 how New York City is densely covered by more than $\frac{1}{2}$ million checkins. While these figures convey the scale and density of location sharing services, we can further explore the nature of these checkins by aggregating keywords across all 22 million checkin tuples. The aggregated view in Figure 4 shows that the most popular checkin venues are restaurants, coffee shops, stores, airports, and other venues reflecting daily activity (e.g., fitness, pubs, church).

4.2 Whens of the checkins

Considering the temporal distribution of checkins, we can uncover both the aggregate daily patterns of users of location sharing services and their weekly patterns. By normalizing the timestamps of every checkin so that all local times are treated as the same time (i.e., aggregating all checkins at 1pm, whether they be in Chicago or Tokyo), we show in Figure 5 the mean checkin pattern per day. This pattern provides a glimpse into the global daily "heartbeat", with three major peaks: one around 9am, one around 12pm, and one around 6pm. The diurnal pattern is clearly displayed as more people are active during the daytime than at night.

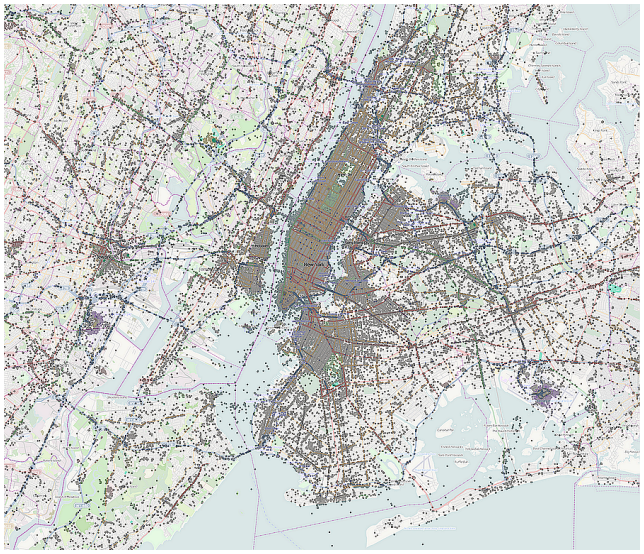


Figure 3: Detail: Checkins in New York City

Figure 4: Venue Cloud for Checkins

To illustrate the potential of location sharing services as sociometers of city health and activity, we show in Figure 6, the disaggregated daily checkin patterns of users in New York City, Los Angeles, and Amsterdam. The checkin patterns show that Amsterdam's daily "heartbeat" reflects an early-rising city, with more activity than either LA or New York in the morning hours. LA peaks around noon, whereas New York has the highest checkin rate during the night ("The City That Never Sleeps"). We are interested to further explore the reasons for these differences. Are the daily differences artifacts of local culture? Or the proclivity of users in certain locations to more willingly reveal certain aspects of their daily lives than others (e.g., checkin in while at work, but not at play?) Or do the differences reflect biases in the data, so that certain demographics are overrepresented in one city versus another?

Moving from the daily pattern to the weekly pattern, we

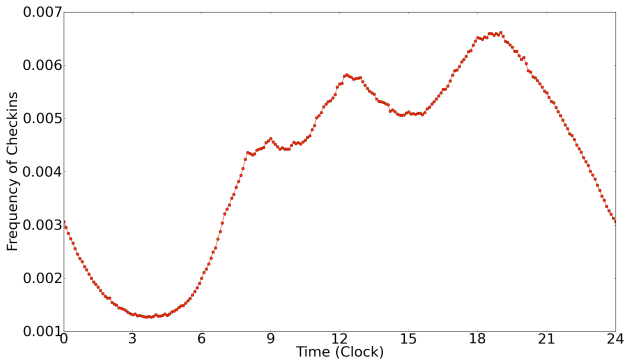


Figure 5: Mean Daily Checkin Pattern

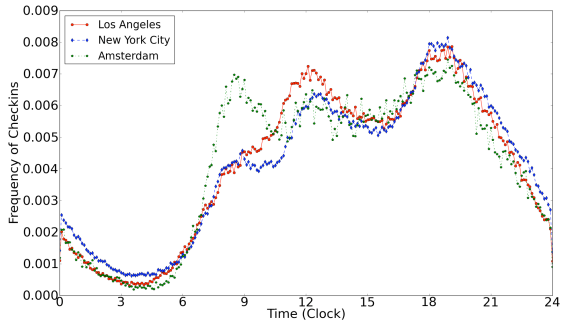


Figure 6: Daily Checkin Patterns: NYC, LA, Amsterdam

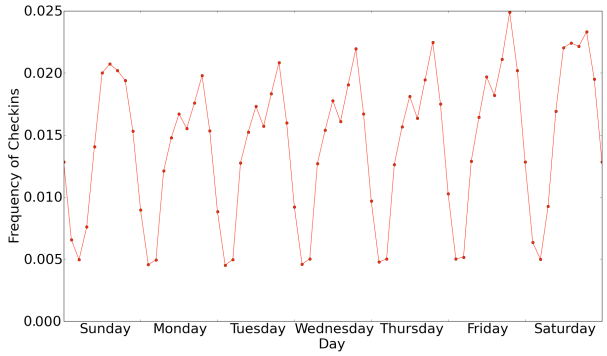


Figure 7: Mean Weekly Checkin Pattern

see in Figure 7 the aggregate global patterns over the days of the week. Weekdays clearly indicate two peaks during lunch time and dinner time, while over the weekend these two peaks blend, reflecting a fundamentally different weekend schedule for most users of location sharing services. We can also observe that the relative daily activity increases from Monday to Friday, peaking on Friday evening.

5 Studying Human Mobility Patterns

Given the global coverage of location sharing services and the potential of user checkins to reveal temporal patterns of human behavior, we next turn to an examination of mobility patterns reflected in the checkin data. We consider three statistical properties often used in the study and modeling of human mobility patterns – displacement, radius of gyration,

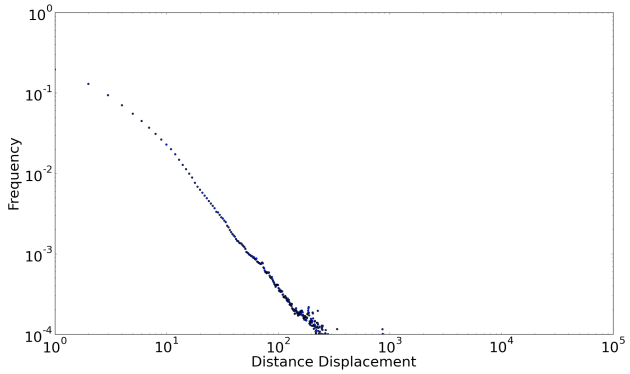


Figure 8: Distribution of Displacements

and returning probability. Taken together, these properties can inform whether humans follow simple reproducible patterns, and can have a strong impact on all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and agent-based modeling.

5.1 User Displacement

We begin with an investigation of the distance-based *displacement* of consecutive checkins made by users. Considering all pairs of consecutive checkins yields 22,163,511 separate displacements, reflecting the distance between these consecutive checkins (and hence, how far a user has traveled). We plot the distribution of displacement for the dataset on a log-log scale in Figure 8. The x-axis is the displacement in miles, and the y-axis is the frequency of displacements in the same bucket. The trend is approximated by a power-law:

$$P(\delta_r) \propto \delta_r^{-\beta}$$

where δ_r represents the displacement and $\beta = 1.8845$. The formula indicates that human motion modeled with checkin data follows a Lévy Flight (Rhee et al. 2008), in which a random walk proceeds according to steps drawn from a heavy-tailed distribution. A Lévy Flight is characterized by a mixture of short, random movements with occasional long jumps. Flight models with a similar scaling exponent have been observed separately in a study of displacements based on cellphone call data with $\beta = 1.75$ (Gonzalez, Hidalgo, and Barabasi 2008) and in a study of displacements based on bank note dispersal with $\beta = 1.59$ (Brockmann, Hufnagel, and Geisel 2006).

5.2 Radius of Gyration

Second, we consider the *radius of gyration* of each user, which measures the standard deviation of distances between the user's checkins and the user's center of mass. The radius of gyration measures both how frequently and how far a user moves. A low radius of gyration typically indicates a user who travels mainly locally (with few long-distance checkins), while a high radius of gyration indicates a user with many long-distance checkins. The radius of gyration for a user can be formalized as:

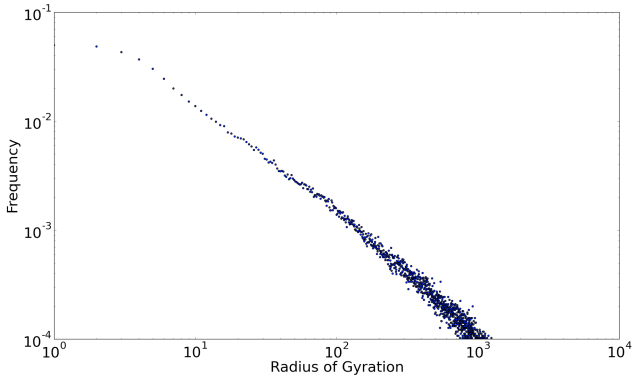


Figure 9: Distribution of Radius of Gyration

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2}$$

where n is the number of checkins of the user, and $(r_i - r_{cm})$ is the distance between a particular checkin r_i and the user's center of mass r_{cm} (which is a simple average location over all checkins). We calculate the radius of gyration for each user in our collection and the distribution of radius of gyration is displayed on A log-log scale in Figure 9. The x-axis identifies the radius of gyration in miles and the y-axis shows the number of users with that radius of gyration. The trend in Figure 9, like the distribution of displacements, also follows a power-law:

$$P(r_g) \propto r_g^{-\beta}$$

where r_g represents the radius of gyration, and $\beta = 0.9864$. 34.5% of all users display a radius of gyration of less than 10 miles, while only 14.6% have a radius of gyration larger than 500 miles.

To illustrate how radius of gyration can give further insight into the dynamics of cities, Figure 10 plots the average radius of gyration of users in major cities (with 100,000+ population and at least 20 users in the checkin dataset) in the continental US. The red bubbles are cities with a radius of gyration larger than 500 miles; blue ones are cities with a radius larger than 250 miles; cyan ones have a radius larger than 125 miles, and yellow ones are the rest of major cities. Users in coastal cities tend to have a higher radius of gyration than users in inland cities, and people in central states tend to have a high radius of gyration due to long distance travels to the coasts. Even so, there are some interesting regional variations worth further study, for example, the low radius of gyration for El Paso compared to the higher radius for nearby Albuquerque.

5.3 Returning Probability

The third property we study – returning probability – is a measure of periodic behavior in human mobility patterns. Periodic behavior is common in people's daily life (e.g., visits to work or school every weekday; visits to the grocery store on weekends) and echoes periodic behavior observed

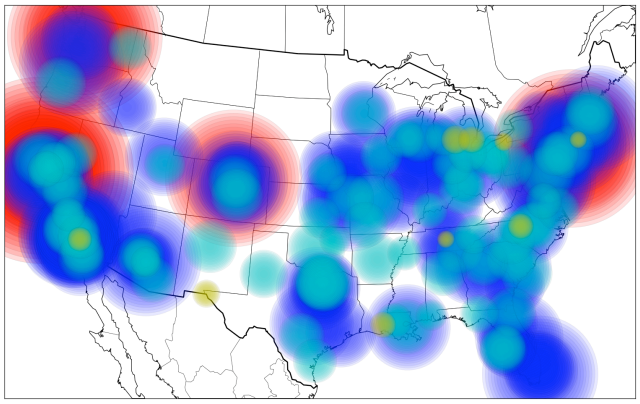


Figure 10: Mean Radius of Gyration for Users in US Cities

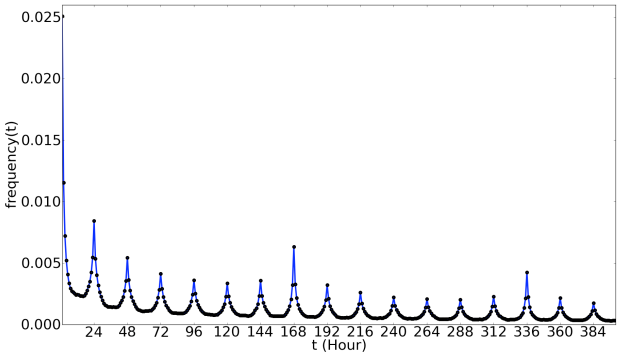


Figure 11: Distribution of Returning Probability

in animal migrations when animals visit the same places at the same time each year. Do users of location sharing services display a similar periodicity?

We measure periodic behavior by the *returning probability* (or, first passage time probability), which is the probability that a user returns to a location that she first visited t hours before. Grouping all returning times of all checkins into buckets of one-hour, we plot the distribution of returning times in Figure 11, in which the x-axis represents the bucket of returning time, and the y-axis is the corresponding frequency for a bucket. For example, at 168 hours, the returning probability peaks, indicating a strong weekly return probability. Similarly, we see daily return probabilities. As time moves forward, the returning probability shows a slight negative slope, indicating the aggregate forgetfulness of visiting previously visited places (that is, the return probability is strongest for places we have visited most recently).

6 Exploring Factors that Influence Mobility

In this final section, we turn our attention to exploring the factors that may impact human mobility. While factors like geography and economic status are natural to investigate, the unique properties of location sharing services provide an unprecedented opportunity to consider heretofore difficult to measure aspects of human behavior. For example, does social status as measured through popularity in these services impact a user's radius of gyration? Does user-generated con-

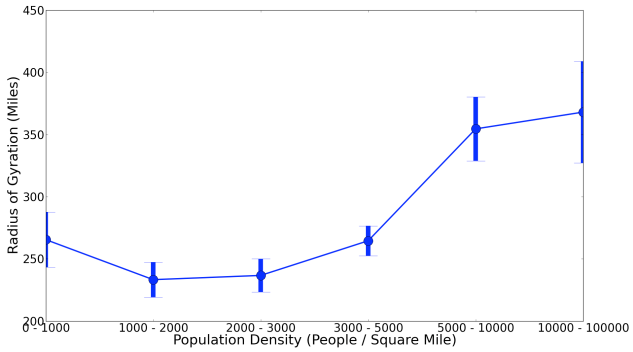


Figure 12: Average R_g versus City Population Density

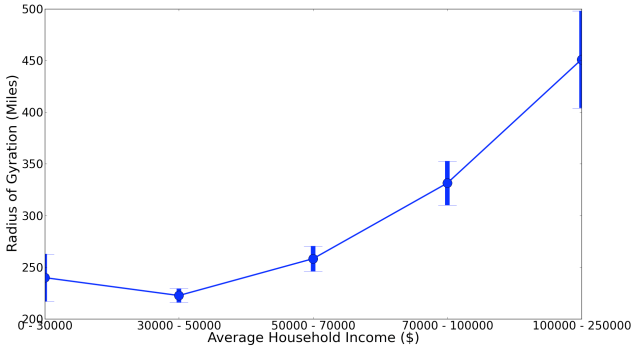


Figure 13: Average R_g versus City Avg Household Income

tent implicitly reveal characteristics of the mobility of users?

6.1 Geographic and Economic Constraints

We begin by illustrating how geographic and economic constraints can influence human mobility patterns as revealed by location sharing services. We focus on users who are located in US cities with a population of more than 4,000. As one type of geographic constraint we consider population density and compare the radius of gyration for users from cities of differing density.[3]

As shown in Figure 12, we can clearly see that people in the densest areas travel much more than people in sparse areas, but that people in the sparsest areas travel farther than people in slightly denser areas. One possible explanation for both of these observations can be that: people living in metropolitan areas have more opportunities to travel for business to distant cities or countries; and people living in sparse areas (small towns) require longer travel to nearby mid-size cities.

Similarly, we can examine the economic properties of a city to understand whether economic capacity inhibits or encourages more travel by its residents. Specifically, we measure the influence of a city's average household income on its residents' radius of gyration, which is plotted in Figure 13. The figure shows that people in wealthy cities travel more frequently to distant places than people in less rich cities. In the meantime, people in cities with the least incomes travel slightly more than people in richer cities.

[3]Data for each US city is parsed from www.city-data.com.

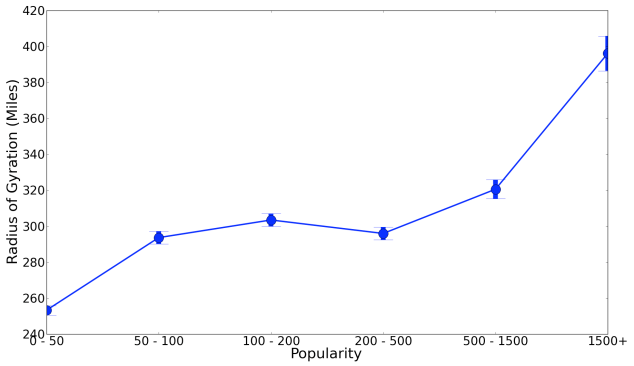


Figure 14: Average R_g versus Popularity

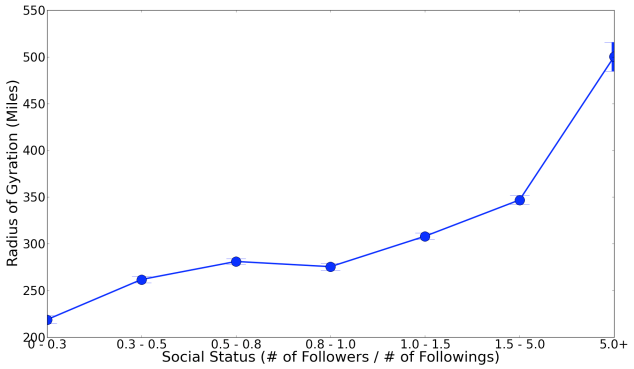


Figure 15: Average R_g versus Social Status

What is encouraging about both these example observations is that location sharing services provide a new window for measuring and studying fundamental properties of cities and their residents.

6.2 Social Status

We next turn to one of the more exciting possibilities raised by the social structure inherent in location sharing services. Does social status impact human mobility? We consider two simple measures of status. The first is a simple measure of popularity, where we count the user's number of followers from their Twitter profile (recall the data collection method described earlier in the paper; followers are one-sided friendships). The second is a measure of status that considers the ratio of a user's number of followers to the number of users that the user follows (followings):

$$status(u) = \frac{n_{followers}(u)}{n_{followings}(u)}$$

High-status users have many followers but follow very few other users themselves. Figure 14 and Figure 15 show the relationship between both of these social status factors and the radius of gyration. We see that in both cases highly social users have higher radii of gyration than less social users. Our initial hypothesis is that users who travel have more chances to meet friends, and thus get involved in more social activities. But perhaps users with lower measured "status" engage with these social media technologies differ-

ently? For example, some Twitter users may primarily only follow other users as a form of news gathering, rather than treating Twitter as a social network of friends, resulting in lower measured status. We are interested to explore these and related questions in our ongoing research.

6.3 Content and Sentiment Factors

Finally, we turn to an analysis of user-generated content in location sharing services and its impact on mobility. Users of location sharing services, in addition to recording their location, can also post short messages, tips, and other annotations on the locations they visit. Unlike purely GPS-driven or cellphone trace data, these short messages provide a potentially rich source of context for better understanding how users engage with location sharing services.

Significant Terms vs. Radius of Gyration: Our first goal is to identify *significant terms* for users associated with varying degrees of radius of gyration, much like in our previous studies of economic, geographic, and social factors. Do high mobility users describe the world differently than low mobility users? We focus our study here on English-language messages only by using the language identification component in the NLTK toolkit (Loper and Bird 2002). We find that 49% of all users (110,559) in our collection are primarily English-language users.

To identify significant terms for these users, we identify terms with high mutual information for each category of radius of gyration. Mutual information is a standard information theoretic measure of "informativeness" and, in our case, can be used to measure the contribution of a particular term to a category of radius of gyration. Concretely, we build a unigram language model for each category of radius of gyration by aggregating all posts by all users belonging to a particular category of radius of gyration (e.g, all users with a radius of gyration between 0 and 10). Hence, mutual information is measured as: $MI(t,c) = p(t|c)p(c)log\frac{p(t|c)}{p(t)}$ where $p(t|c)$ is the probability that a user which belongs to category c has posted a message containing term t, $p(c)$ is the probability that a user belongs to category c, and $p(t)$ is the probability of term t over all categories. That is, $p(t) = count(t)/n$. Similarly, $p(t|c)$ and $p(c)$ can be simplified as $p(t|c) = count(c,t)/count(c)$ and $p(c) = count(c)/n$ respectively, where $count(c,t)$ denotes the number of users in category c which also contain term t, and $count(c)$ denotes the number of users in category c.

In Table 2, we report the top-10 most significant terms from users with different radii of gyration. In the table, we can clearly see the differences between frequent travelers with a large radius of gyration and the more local people with a small radius of gyration. Travelers talk a lot about long journey related terms: "international airport" (and abbreviations of international portals: "SFO", "JFK"), major metropolitan areas (e.g., New York, San Francisco, London, Paris, Los Angeles), "flight", and "hotel". At lower levels of mobility, we see significant words like "railway station" and "bus", as well as discussion of "home", "work", "church", grocery stores (e.g., HEB, Walmart, "mall"), "college", and "university". People with different mobility patterns significantly differ in the topics they talk about and terms they use, indicating a fruitful area of further study.

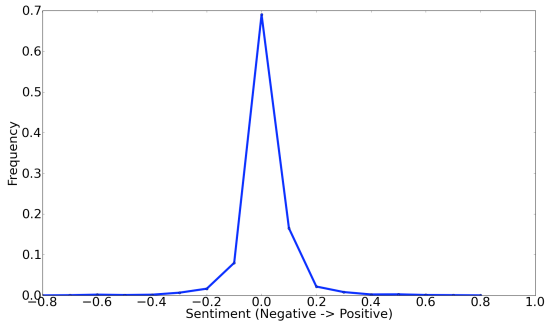


Figure 16: Frequency of Users in Categories of Sentiment

Table 3: Top-10 Significant Terms for Sentiment Category

Sentiment	Top 10 Terms				
$(0.1, 1.0]$	good	like	love	lol	well
	thanks	great	haha	awesome	nice
$(-0.1, 0.1]$	ave	mayor	street	New York	park
	road	blvd	airport	center	home
$[-1.0, -0.1]$	not	hate	bad	f**k	s**t
	damn	wrong	hell	stupid	hiv

Capturing User's Sentiment: We can additionally measure the relative viewpoint of users and their locations by considering the sentiment of each user's posted messages. To capture the sentiment associated with the checkins, we use the public SentiWordNet (Esuli and Sebastiani 2006) thesaurus to quantify sentiment for each English speaking user. For each message, we extract the words that have a quantified sentiment value in SentiWordNet and consider the sentiment of the post as the mean value for the sentiments for words in the post. For each user, the user's sentiment is calculated as the mean value of the sentiments of all the user's posts. In this way, we capture the sentiment for each of the 110,559 English speaking users in the dataset. The distribution of sentiment of the users is plotted in Figure 16, and we can clearly see that most users have a neutral sentiment, and only a small portion of users express strong sentiment when using location sharing services.

When we drill down to see which words are associated with a positive, neutral, and negative sentiment (again, using mutual information) we see in Table 3 that most of the top neutral terms are likely to be extracted from the auto-generated checkins. In the two categories with non-neutral sentiment, we can clearly see typical words which indicate strong positive and negative sentiment.

However, when we filter the top-100 most positive and most negative terms to only consider location-related terms, we find that there are no location-specific positive terms, but there are many location-specific negative terms. Examples of the words are listed in Table 4. On further inspection of the messages containing these words, we can clearly see the strong negative sentiment associated to the content. For example, when people talk about "MTA", they complain a lot about price increases of MTA's tickets, and its poor service

Table 2: Top 10 Significant Terms for Each Radius of Gyration R_g Category

R_g (miles)	Top 10 Terms				
(1000,+∞)	international airport	New York	San Francisco	London	terminal
	SFO	flight	JFK	Jakarta	Paris
(500,1000]	international airport	San Francisco	New York	Las Vegas	Los Angeles
	Chicago	hotel	Seattle	terminal	Washington
(300,500]	international airport	Chicago	Dallas	New York	hotel
	Lake	Austin	Beach	Orlando	Seattle
(100,300]	airport	Chicago	Atlanta	Jakarta	hotel
	Berlin	church	center	bar	beach
(50,100]	mayor	railway station	Pittsburgh	university	Stockholm
	church	Madrid	Greenville	center	college
(10,50]	mayor	station	home	work	Bangkok
	house	HEB	school	Walmart	road
(0,10]	Singapore	home	Jakarta	Indonesia	university
	center	mall	bus	woodlands	road

Table 4: Top-20 Location Terms with Negative Sentiment

MTA	Jersey	Redmond	Memphis
Winooski	Ridgewood	Toronto	Greece
Chicago	Cleveland	Calgary	Scottsdale
Beaumont	Petersburg	Ashburn	Buffalo
Richmond	Montreal	Durham	Eugene

(e.g., "Ticket to the country home has increased by $3. NJ-Transit is worse than the MTA! (@ New York Penn Station w/ 23 others)", and "I know the MTA is a disaster but 2 of 4 machines being unable to read credit cards at AirTrain station is a new low."). This preliminary analysis indicates that users are more likely to express negative sentiment about location, and that locations and location-related concepts associated with negative sentiment can be automatically identified based on location sharing services.

7 Conclusion

In this paper, we provide the first large-scale quantitative analysis and modeling of over 22 million checkins of location sharing service users. Concretely, three of our main observations are: (i) LSS users follow simple reproducible patterns; (ii) Social status, in addition to geographic and economic factors, is coupled with mobility; and (iii) Content and sentiment-based analysis of posts can reveal heretofore unobserved context between people and locations. As future work, we are interested to further explore the social structure inherent in location sharing services to study group-based human mobility patterns (e.g., flock behavior). We are also interested in personalized location recommendation based on checkin history and friend-based social mining.

References

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW '10*.

Brockmann, D.; Hufnagel, L.; and Geisel, T. 2006. The scaling laws of human travel. *Nature* 439(7075):462–465.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *CIKM '10*.

Cholera, R.-S. 2011. Words of the year 2010 (the wall street journal). http://on.wsj.com/e7AyTt.

Esuli, A., and Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*.

Foursquare. 2010. Cheating, and claiming mayorships from your couch. http://blog.foursquare.com/2010/04/07/503822143/.

Foursquare. 2011. So we grew 3400% last year. http://blog.foursquare.com/2011/01/24/2010infographic/.

Golder, S. A.; Wilkinson, D. M.; and Huberman, B. A. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Proceedings of the Third Communities and Technologies Conference*.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Grove, J. 2010. Foursquare nearing 1 million checkins per day (mashable). http://mashable.com/2010/05/28/foursquare-checkins/.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from justin biebers heart: the dynamics of the location field in user profiles. In *SIGCHI '11*.

Humphries, N. E., et al. 2010. Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature* 465(7301):1066–1069.

Lindqvist, J., et al. 2011. I'm the mayor of my house: Examining why people use foursquare - a social-driven location sharing application. In *SIGCHI '11*.

Loper, E., and Bird, S. 2002. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*.

Rhee, I.; Shin, M.; Hong, S.; Lee, K.; and Chong, S. 2008. On the Levy-Walk nature of human mobility. In *INFOCOM '08*. IEEE.

Song, C.; Qu, Z.; Blumm, N.; and Barabasi, A.-L. 2010. Limits of Predictability in Human Mobility. *Science* 327(5968):1018–1021.

Yardi, S., and Boyd, D. 2010. Tweeting from the town square: Measuring geographic local networks. In *ICWSM '10*.

Zheng, Y., et al. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *WWW '10*.

IV

Big Data Spaces: Wikileaks

Lynch, Lisa (2010). "A Toxic Archive of Digital Sunshine: Wikileaks and the Archiving of Secrets." Paper presented at the MIT6 conference, Cambridge, MA. http://web.mit.edu/comm-forum/mit6/papers/Lynch.pdf

Lovink, Geert and Patrice Riemens (2010). "Twelve Theses on WikiLeaks," *Eurozine.* http://www.eurozine.com/articles/2010-12-07-lovinkriemens-en.html

Stalder, Felix (2010). "Contain this! Leaks, whistle–blowers and the networked news ecology." *Eurozine.* http://www.eurozine.com/articles/2010-11-29-stalder-en.html

Sterling, Bruce (2010), "The Blast Shack," Webstock. http://www.webstock.org.nz/blog/2010/the-blast-shack/

Žižek, Slavoj (2011). "Good Manners in the Age of WikiLeaks," The London Review of Books, 33:2, 9-10. http://www.lrb.co.uk/v33/n02/slavoj-zizek/good-manners-in-the-age-of-wikileaks

Video

Manuel Castells (2011), "From WikiLeaks to Wiki-revolutions," SONIC Media, Technology and Society Speaker Series, Lecture on 8 March 2011 at Northwestern University, Video Registration available at: http://lecture.soc.northwestern.edu/mediasite/Viewer/?peid=4e192796ace943fabf8172b463ce74381d

Lisa Lynch

A Toxic Archive of Digital Sunshine: Wikileaks and the Archiving of Secrets

DISCLAIMER: MANY THINGS HAVE BEEN HAPPENING IN RELATION TO WIKILEAKS OVER THE PAST FEW WEEKS, SO THIS VERSION OF THE PAPER MAY DIFFER FROM THE VERSION I ACTUALLY DELIVER AT MIT6. I ALSO ASK THAT THIS PAPER NOT BE CIRCULATED IN THIS VERSION.

On March 19th, Australian citizens learned that their government was considering instituting a mandatory national filtering system that would prevent them from accessing a list of websites identified as having connections to child pornography.[1] The origin of this revelation -- which engendered a substantial political fallout and the likely consequence that the list will not be approved in the Australian senate -- was not, as might be expected, a journalistic investigation, or a TV press conference featuring an indignant whistle-blower. Instead, the plan was made public through a leaked copy of the proposed blacklist posted on Wikileaks, a Swedish-hosted website run by an international collective and dedicated to "untraceable mass document leaking and analysis." [2]

For followers of the Wikileaks site, the fact that the blacklist appeared on Wikileaks first was hardly surprising. Since its launch in early 2007, Wikileaks has published scores of documents never intended for public view, and its professed ability to safeguard the security of those who wish to upload and circulate such documents has meant that the site has become a primary destination for leakers, for the media, and for members of the interested public. But what *was* surprising about the list's publication was the disabling

effect that it had on Wikileaks. On March 22nd -- two days after Australia's Minister for Broadband, Communications and the Digital Economy, Senator Steven Conroy, threatened legal action against the site -- Wikileaks became unavailable.[3]

Soon after Wikileaks went offline, bloggers and some media outlets began speculating whether the site had been the victim of a court-ordered shutdown.[4] After all, Conroy was not Wikileak's only enemy; over the past two years, governments, corporations, and the Scientologists had repeatedly tried to quash Wikileaks, and the Swiss bank Julius Baer had been successful in getting Wikileaks' American domain temporarily disabled in February of 2008.[5] Soon, however, the members of the Wikileaks editorial board announced via Twitter that the site had not been forced offline by any political, religious or corporate entity; rather, global interest in the blacklist had overwhelmed the Wikileaks servers.[6] Later that day, visitors who came to the site were met with a static page which contained an apology, as well as a request for donations to enable Wikileaks to upgrade their equipment in the face of increased demand. Over the following week, service remained spotty, and the online community of Wikileaks followers began to express their concern about whether the Wikileaks collective would eventually become the victim of its own success.

I'm interested in this story about Wikileaks and the Australian blacklist on several levels; among other things, it's a good example of the ever-increasing boundary skirmishes between traditional, institutional sites of facticity and newer sites, a topic I'm exploring in a longer project. Here, however, in accordance with the theme of the conference, I am

going to focus on the idea of Wikileaks as an archive -- a digital archive of censored documents that are either revealed or yet-to-be-revealed, thus an archive of secrets both expired and untold. What I'll be arguing here is that the Wikileaks collective, by creating what is arguably the safest and easiest way to anonymously upload classified documents for publication, has paradoxically engineered a suicidal archive in which each subsequent release of a document poses a threat to the entire archive's existence. As my opening anecdote suggests, this threat is both legal and operational, since publishing a document can spur government action to have the site taken offline, but it can also create such a level of interest that Wikileaks is ultimately unable to keep up with the demand.

Thus far, the disruptions Wikileaks has experienced as a result of its actions have been temporary, and the site's founders continue to argue that the system Wikileaks has created is robust enough to withstand future legal and operational assault.[7] But I presume the opposite here – namely, that there is something inherently fatal about the enterprise of Wikileaks, a death-drive built into the very structure of its archive.[8] As I'll argue, this potential for self-destruction coexists uneasily with Wikileaks' aggressive positioning itself as the go-to repository for classified documents. As it faces a new wave of challenges to its continued existence, Wikileaks serves to remind us of the fragility of the digital archives that are increasingly mediating our experiences of both historical and present-day records.

So what exactly is Wikileaks? It is not an affiliate of Wikipedia, or the Wikimedia Foundation; rather, it is one of a number of websites -- including *Cryptome, The Memory*

Hole, *National Security Archive*, *GlobalSecurity.org,* and the *Nautilus Institute* -- dedicated to providing an outlet for information that might otherwise remain secret. What is distinctive about Wikileaks is the extent to which it has been aggressively proactive in soliciting and publicizing material on a broad range of topics. The site's founders, a mainly anonymous collective which, according to the site's "About" section, includes Chinese dissidents, journalists, mathematicians, and 'startup company technologists' from the US, Taiwan, Europe and South Africa, claim they are now processing over a million documents uploaded from locations around the world, selecting and vetting those which have political, diplomatic, ethical or historical significance. Those they select as meriting attention are posted, translated when possible (currently, about 30 languages are represented), and announced via RSS, Twitter, and media outreach. According to the site, Wikileaks' goal is to create "a social movement emblazoning the virtues of ethical leaking,"[9] that will shine light on corrupt practices everywhere —particularly, they claim, in Asia, the former Soviet bloc, Latin America, Sub-Saharan Africa and the Middle East.

This is an extraordinarily ambitious agenda, and some have accused Wikileaks of drifting off mission -- as several observers have pointed out, the site's disclosures seemed to have shone far more light on European and North American corruption than on corruption elsewhere. However, in a short amount of time Wikileaks has facilitated some truly revelatory and consequential leaking, emerging as both ally and competitor to media outlets around the globe. The documents they have published include the 51,000-name supporter database of ex-US Senator Norman Coleman (with names and addresses); the

partial contents of Sarah Palin's Yahoo inbox; a list of military equipment in Iraq; the complete text of the officer's handbook used at the detention center in Guantanamo Bay Cuba; information on the DOD's Warlock Green and Warlock Red IED jamming technology; the final draft of a US Army Intel brief on Afghani insurgent groups; a selection of Scientology's "Operating Thetan" missives; the membership list of the far-right British National Party (also with addresses); and a series of documents suggesting that Barclay's Bank was engaged in sustained practices of tax avoidance.[10] Though the majority of these documents are anonymously sourced and illegally released, this is not always the case: on April 3, for example, a Canadian academic named Michael Geist uploaded a copy of the 2008 Canadian ACTA Consultation report that he had procured through the Canadian Access to Information Act.[11]

As this very partial list makes clear, these documents have little in common save for three things: first, someone has attempted to hide them from public view, second, someone else acquired them and sent them to Wikileaks, and third, the Wikileaks editorial board decided they were worthy of publication. This last point is important: unlike most wikis, Wikileaks does not allow documents to be published directly to the web or collectively edited. Rather, the Wikileaks site uses a modified version of the Wikimedia platform, which allows users to post documents anonymously to the server for publication following review. The platform also allows anyone interested to comment on the reliability or implication of published documents in a linked comments area.

Another aspect of Wikileaks that distinguishes the site from a conventional wiki is the

manner in which documents are submitted. In order to protect the identity of leakers, Wikileaks uses customized versions of readily available cryptographic and rerouting techniques including Free Net, PGP and Tor. And if leakers remain concerned their computer still might be traced, they can encrypt the documents using online tools provided by Wikileaks and then mail them to designated postal boxes, where they are collected by volunteers and sent on, still encrypted, to a member of the Wikileaks editorial board.[12]

However it reaches Wikileaks, once a document has been received it goes through a vetting process by the investigative journalists on the Wikileaks board.[13] After its authenticity has been established, there are digital encryption procedures that sever the verified document from its forensic trace before publication. When finally published, documents are hosted on a server physically located in Sweden, a country with extremely strong press-freedom protections. They can then be accessed by users either from the central site (Wikileaks,org), or through one of about 50 alternate Wikileaks domains. These domains include both Wikileaks sites (such as http://wikileaks.la/), and "cover" domains established to combat Chinese filtering of all "Wikileaks" sites, (such as http://ljsf.org/). These alternate sites redirect users either to Wikileaks,org, or to a Swedish web proxy that in turn points to Wikileaks' real server – in other words, the system is design to withstand a DNS issue that affects Wikileaks.org.

All of this goes to suggest that Wikileaks has taken great care to engineer a system that protects their sources. However, this concern for safety of sources does not stem from

any general policy on the part of the Wikileaks editorial collective to protect those endangered by the publication process. Wikileaks has often been criticized for publishing information that arguably endangers not only the malefactors it exposes, but also innocent parties -- for example, American military personnel who might be endangered by information about troop equipment. In an interview with NPR's *On The Media* that focused on the possible harm caused by Wikileaks releases, Wikileaks spokesperson Julian Assange told interviewer Bob Garfield that though Wikileaks would consider notifying those they might endanger through publication, they would publish a document "even if there was a possibility of loss of life" as a result of publication. The exception, Assange conceded, would be if publication might result in the loss of life of a *source*, in which case they would "find a way to sit on the information." [14] Otherwise, the collective's commitment to free information would allow for no redaction in the interest of safety or propriety.[15]

As Assange explained during the NPR interview, the Wikileaks collective believes that "their primary loyalty is to their sources, not to their readers;" creating a climate of trust is the most important step towards encouraging sources to reveal information, which in turn is the best way to get more information from sources. Thus, if we take Wikileak's words at face value, we can understand this difference between Wikileak's concern about the safety of their sources, and their resigned acceptance of the harm that others might suffer due to documents submitted by these sources, as a consequence of the editorial collective's belief in the ultimate primacy of the freedom of information.

But there are at least two other ways to think about Wikileaks approach, both tied to the

archival impulse at the heart of Wikileaks' enterprise.

First, we can connect Wikileaks' cultivation of their sources to the compulsion evidenced on the site for the collection of the greatest number and greatest variety of secret documents. For all of its focused sense of mission, the Wikileaks project, as I've suggested, is in reality a vast cabinet of miscellany. Ranked chronologically instead of according to importance, the overwhelming variety of documents on the site's homepage are alternately exhilarating and exhausting to peruse. Following Wikileaks' random-seeming, rapid-fire release of document after document, one senses the mixed fatigue and urgency that marks each subsequent release, and senses also that nothing else is as important to those at the center of this enterprise than this endless cycle of collection, release, and collection.

Second, if Wikileaks' approach to their sources can therefore be read as a sort of symptom of archival compulsion, it can also be read as an indicator of the radical rupture in the constitution of the Wikileaks archive -- the schism between those documents which are published and those which are still queued for publication. The difference between these two kinds of documents cannot be overstated, nor can we underestimate the tension between them. To clarify, consider exactly what Wikileaks does in publishing a document. Whatever function these documents serve in their original context – be they contracts, handbooks, correspondence, blacklists, etc -- publication transforms them into performative acts, interventions into ongoing political, financial, military or legal crises. In this sense, one could draw a connection between the Wikileaks archive and other

collections of official documents -- the Guatemalan Police Archives, released to the public at the beginning of April, come to mind -- that consist of a series of records whose function changes radically when circulated among different audiences. But Wikileaks is also different from archives such as these in that *from the moment a document enters the Wikileaks system,* it begins a process of transformation which liberates the record's potential as radical act. In this process of digital reincarnation –or reinscription might be the better word, since it is essentially stripped of its prior material identity -– the document loses its ability to cause harm to the person who has submitted it to Wikileaks, while at the same time gaining the ability to harm those it directly or indirectly indicts.

But the danger posed by each published Wikileaks document extends beyond its ability to harm those it names – or even to inadvertently harm innocents.[16] As we have seen, each publication can also pose a threat to the architecture of the archive itself, and thus, to the documents that have not yet been released for publication. As I suggested earlier, some of the danger posed by these documents is purely operational: as Wikileaks documents gain more public attention, the increased demand for the archive threatens to shut it down completely. On March 24, while Wikileaks was offline, this issue of site traffic became the topic of a forum discussion on the social news site Reddit that encapsulated the site's current dilemma. One poster argued that the problems Wikileaks was having with site traffic were inevitable, given that Wikileaks relied on donor financing to support a centralized infrastructure: "Every year their site resources requirements balloon and they beg for more and more money...no amount of money can buy the redundancy that the public needs from these sites." His solution – proposed by several others on the forum

as well – was that Wikileaks switch to a peer-to-peer distribution system that would avoid the expenses and managerial hassles of centralized hosting. However, another poster pointed out that the physical location of the Wikileaks servers in a country with strong press freedom laws was vital to the site's survival, adding, "I do not want my non-Swedish IP address on a public bittorent tracker, seeding the sort of documents that Wikileaks publishes, without that sort of legal backing."[17] And another poster objected to the idea of P2P technology because the more complicated interface would interfere with the site's public mandate. "Wikileaks is for the masses…the reality is that if one wants to get content distributed far and wide over the Internet, it need to be delivered via http."

Over the past several weeks, the potential for documents to have increasing consequences for Wikileaks has become quite clear, due to a series of events that in Germany connected with the German Wikileaks domain, Wikileaks, de.[18] On March 24, police officers in Dresden and Jena searched the two homes of Theodor Reppe, the designated owner of the domain name, ordering Reppe take Wikileaks.de offline. As German officials acknowledged later, this search was conducted as part of a child pornography investigation. Germany has just finalized its own proposal to introduce a nation-wide child pornography filter, and the Wikileaks site, which published a list of links of to child pornography, was assumed to be a pornography portal.[19]

Because Reppe didn't have access to the domain name passwords, he was unable to comply, and the police left without pressing charges against Reppe or seizing equipment.

In the days following, the fact that the Wikileaks.de domain had been protected seemed to suggest that Wikileaks' system of safeguarding their domains and servers was working. However, on April 9, the Wikileaks.de was taken offline, Wikileaks notified their readers that the DENIC (Deutsches Network Information Center), the manager of the .de domain, had seized the domain without warning. An editorial on the Wikileaks website announced "Germany Muzzles Wikileaks!" The collective began a solidarity campaign, urging German activists to begin pointing their own domains to Wikileaks, or launch new domains, including http://repressionsstaat.de/. Bloggers wondered whether German censors would ferret out German—based cover domains for Wikileaks in the manner of the Chinese government, or even bloc access to Wikileaks.org.

Unfortunately for Wikileaks, by Monday morning it appeared that they had their story wrong. The problem was not with DENIC, but rather with the lower-level domain registrar, Beasts Associated, who claimed that they had given Reppe 90-day notice that his domain had been terminated. According to DENIC and Beasts Associated, any connection between the site's disappearance and the police raid was purely coincidental. Responding to the DENIC statement, Wikileaks announced that indeed, the decision to terminate the domain was linked to documents published in December detailing activities of the BND, the German CIA. The timing of the termination, they still insisted, was related to the raid on Reppe's home. On forums and on Twitter, activists who had responded to Wikileaks' outraged declaration of censorship began to deride Wikileaks for claiming that they had been censored by the German government for the porn blacklists, and some dismissed Wikileaks's response as conspiracy theory.[20] In the wake of the

incident. Wikileaks found itself in the position of facing a crisis of credibility for having cried 'censorship' too soon, while at the same time coming to terms with the fact that, even if they were wrong in this instance, it remained a possibility that the continuing presence of the Australian blacklist might give other countries license to censor Wikileaks in the name of censoring child porn, thus setting Wikileaks up for a far more charged legal battle than it had faced previously.

So – Wikileaks is heading into uncharted legal and operational territory, facing two crises that have become utterly entangled. The more attention that the site gets as a result of its DNS issues, the more difficulty the Wikileaks technical staff will have managing their bandwith crisis. And increasing concerns about the legal consequences of the pornography blacklists will mean that money that might otherwise be directed towards resolving traffic issues will be directed towards a legal fund. But the editorial collective is obviously determined to prove themselves undaunted by recent events – on April 11, even as they rallied supporters to donate money for what they believed would be a censorship battle in Germany, they published a purloined copy of a portion of the Anti-Counterfeiting Trade Agreement (ACTA), a document that the Obama Administration had classified under the State Secrets Act. Whether or not this document will attract the media attention that it deserves, it will certainly draw government interest in a moment that Wikileaks is perhaps uniquely vulnerable.

I am not suggesting here that Wikileaks should not have published their copy of ACTA – nor, for that matter, do I intend in this paper to quibble with any of the individual

procedures or actions of Wikileaks that I have described thus far. As I suggested at the beginning of this paper, Wikileaks is a paradox – brilliant, invaluable, but also damned by its ideological consistency and the seeming necessity of its current architecture (an architecture which may, in a few year's time, seem to be an artifact of technological history). I do, however, want to take issue with the manner in which Wikileaks actively discourages potential whistleblowers from going elsewhere, since I think it may have grave consequences for the archive of unpublished documents that Wikileaks has assembled.

The tenor of Wikileaks' wooing of their sources is perhaps best demonstrated by an editorial published on the site last October, in which Wikileaks enumerates the perils of leaking to anywhere else but Wikileaks. The author asserts that while no one who has uploaded to Wikileaks has been caught,[21] in three separate instances -- the hacked Sarah Palin emails, a set of documents regarding kickbacks given by Sallie Mae, and a handbook discussing rituals of the Kappa Sigma fraternity -- material uploaded to Wikileaks by an intermediary *after* it had been leaked by another party wound up incriminating the initial leaker. Detailing the traces that let to each original leaker's identification, the editorial urges potential leakers to "Communicate with Wikileaks and only Wikileaks. After the dust has settled you can consider how you may want to tell others."[22]

Whatever the concern of Wikileaks' editors for their source, the advice it prescribes is troubling. I have suggested, a world in which whistleblowers did in fact "communicate

with Wikileaks and only with Wikileaks," would not necessarily be a good thing; if Wikileaks becomes the sole archive of the world's dirty laundry, and then implodes, what becomes of the secrets left untold? Surely some of them would still emerge; others, however, would not. To gloss on digital historian Roy Rozensweig, Wikileaks has done to the world of secrecy what digital archives have done for the telling of history – it has given us the gift of abundance while threatening us with a future of scarcity. [23]

In *Archive Fever*, Derrida notes that 'effective democratization can always be measured by this essential criterion; the participation in and access to the archive, its constitution, and its interpretation.' As Wikileaks becomes increasingly important as a means to distribute classified information, it seems to precisely emulate the kind of archive that Derrida describes: an collection of documents of public importance that is designed to accrue the most possible material, to be available to the widest range of citizen, and to facilitate mass interpretation. And yet Wikileaks is also something quite different – the unstable digital doppelganger of what Peter Galison has described as the unstable rising mountain of material in the classified world. However useful it might be as a tool as we attempt to chart the tectonics of these of worlds of truth and of secrecy, Wikileaks should also be a reminder of the dangers of placing too much faith in our ability to engineer ourselves the society we desire through technological means alone.

---

[1] As noted in March 19 article on Forbes.com, the list was not limited to pornographic sites per se, but rather included certain Wikipedia entries, some Christian sites, the Web site of a tour operator and even a Queensland dentist's practice. (See http://www.forbes.com/2009/03/19/ Australia-internet-censorship-markets-economy-wikileaks.html)

[2] See http://wikileaks.org/wiki/Wikileaks:About

[3] According to an article on the website of Australia's ABC News, Conroy stated that "ACMA is investigating this matter and is considering a range of possible actions it may take including referral to the Australian Federal Police. Any Australian involved in making this content publicly available would be at serious risk of criminal prosecution." See http://www.abc.net.au/news/stories/2009/03/19/2520929.htm

[4] See http://www.mediawatchwatch.org.uk/2009/03/19/wikileaks-is-offline-was-it-the-australian-government/; http://www.scmagazineuk.com/Wikileaks-taken-offline-after-it-publishes-banned-Australian-websites/article/129213/; http://www.networkworld.com/community/node/39977; http://www.freerepublic.com/focus/chat/2210018/posts; http://news.digitaltrends.com/news-article/19544/wikileaks-publishes-list-of-banned-aussie-sites-goes-offline for examples

[5] After material appeared on Wikileaks that suggested odd doings were afoot at Julius Baer's Cayman Island outpost, the bank filed a Federal lawsuit against the site and obtained an injunction against Dyanadot, the site's registrar. However, after hearing arguments on behalf of Wikileaks, the judge reversed the injunction and allowed Wikileaks to remain online. Baer eventually dropped the suit. See http://arstechnica.com/tech-policy/news/2008/02/swiss-bank-wins-injunction-against-wikileaks.ars and http://www.guardian.co.uk/media/2008 /mar/06/digitalmedia.medialaw?gusrc=rss.

[6] This was not the first time Wikileaks was taken offline by increased site traffic; the site gone down briefly several times before, including a year earlier after Wikileaks editors chose to mirror a host a censored trailer for the controversial Dutch film *Fitna*. But it the site was offline for longer this time than before.

[7] In fact, WIkileaks spokesman Julian Assange recently told a reporter, "When we get a legal threat everyone jumps up and down with glee, (since) any attack will just draw attention to us and the document they are trying to suppress." See http://www.khaleejtimes.com/DisplayArticle08.asp?xfile=/data/theuae /2009/April/theuae_April113.xml&section=theuae

[8] In *Archive Fever* Derrida suggests that an impulse toward destruction was inherent in the conceptualization of the archive. Indeed, perhaps one way to think about Wikileaks is as the ultimate Derridean archive, in which the struggle between Eros and Thanatos plays out on a daily basis -- not least because the pleasure of Wikileaks, which is the delight in discovering someone's secrets, is intertwined with the continual and often deadly possibility that someone might find out about your secrets.

[9] See http://wikileaks.org/wiki/Wikileaks:About

[10] The Barclay's bank documents were originally leaked to the British newspaper *The Guardian*; they were uploaded to the Wikileaks site after a court injunction forced *The Guardian* to remove them from the *Guardian* website. According to *The New York Times*, the judge additionally forbid from telling their readers where they could find the documents after they had vanished from the site. See http://www.nytimes.com/2009/03/30/technology/internet/30link.html)

[11] Wikileaks published this with the disclaimer, "The document is not a leak, having been obtained under the AtIA, but is related to a number of previous leaks on ACTA released by Wikileaks."

[12] Given the mystery surrounding the actual selection and publication of documents, it is difficult to know how many individuals are involved in the vast task of collection: the site's 'About' section claims that Wikileaks has '1,200 registered volunteers,' but it is hard to know what that means in terms of actual labor.
[13] See http://wikileaks.org/wiki/Wikileaks:About. The editors go on to claim that "We have become world leaders in this, and have never, as far as anyone is aware, made a mistake." While it is true that no Wikileaks document has been revealed as inauthentic, after the recent posting of a "censored" segment of a CBC broadcast, the segment's producer wrote in to say that no one had bothered to check whether the item had been censored: it was not.
[14] See http://www.onthemedia.org/transcripts/2009/03/13/04.
[15] In fact – as Garfield pointed out during the interview -- Wikileaks even recently published a list of their own donors, including addresses, which had been presumably leaked to the site by one of their own members. The publication of this list has understandably cast a pall over the site's ongoing efforts to raise funds.
[16] I should note that publication of documents also endangers the Wikileaks collective; for example, a listener posted on the NPR forum that he hoped Assange met with 'vigilante justice' for publishing information relating to the U.S. military.
[17]http://www.reddit.com/r/reddit.com/comments/86bvy/since_wikileaks_is_obviously_in_bandwidth_trouble/.
[18] In fact, Wikileaks was already on the Australian blacklist for publishing related 'porn' blacklists proposed by Denmark, Thailand and Norway.
[19] See http://news.cnet.com/8301-1023_3-10144413-93.html
[20] About 12 hours after DENIC released their statement, they posted the following on Twitter: "Short update on Wikileaks.de issues: more open questions remaining, situation is still unclear. We will update once we have all information." One poster wrote in response, "no offense, Wikileaks, but i do hope you'll stick to hard facts this time; your/our cause can't afford another case of crying wolf." Following the official Wikileaks response, another follower of the incident wrote on Twitter "Oh, Wikileaks, why do you shoot yourself in the leg with a garbled press release? [translated from original German].
[21] There is one incident that might belie this claim. On March 5, Oscar Kamau Kingara and John Paul Oulo Kenyan human rights activists whose report on Kenyan police assassinations had been leaked to Wikileaks last November, were shot at close range in their car on their way to a meeting with the Kenyan National Commission on Human Rights. It is possible, though of course not established, that Kingara and Oulo themselves might have sent the report to Wikileaks, thus drawing more attention to their findings and resulting in their death.
[22] http://wikileaks.org/wiki/Successes_and_three_near_misses_for_Wikileaks
[23] See http://chnm.gmu.edu/resources/essays/scarcity.php

eurozine

Geert Lovink, Patrice Riemens

Twelve theses on WikiLeaks

Vindictive, politicized, conspiratorial, reckless: one need not agree with WikiLeaks' modus operandi to acknowledge its service to democracy. Geert Lovink and Patrice Riemens see in WikiLeaks indications of a new culture of exposure beyond the traditional politics of openness and transparency.

Thesis 0

"What do I think of WikiLeaks? I think it would be a good idea!" (after Mahatma Gandhi's famous quip on "Western Civilization")

Thesis 1

Disclosures and leaks have been a feature of all eras, however never before has a non–state or non– corporate affiliated group done anything on the scale of what WikiLeaks has managed to do, first with the "collateral murder" video, then the "Afghan War Logs", and now "Cablegate". It looks like we have now reached the moment that the quantitative leap is morphing into a qualitative one. When WikiLeaks hit the mainstream early in 2010, this was not yet the case. In a sense, the "colossal" WikiLeaks disclosures can be explained as the consequence of the dramatic spread of IT use, together with the dramatic drop in its costs, including for the storage of millions of documents. Another contributing factor is the fact that safekeeping state and corporate secrets –– never mind private ones –– has become difficult in an age of instant reproducibility and dissemination. WikiLeaks becomes symbolic for a transformation in the "information society" at large, holding up a mirror of things to come. So while one can look at WikiLeaks as a (political) project and criticize it for its modus operandi, it can also be seen as the "pilot" phase in an evolution towards a far more generalized culture of anarchic exposure, beyond the traditional politics of openness and transparency.

Thesis 2

For better or for worse, WikiLeaks has skyrocketed itself into the realm of high–level international politics. Out of the blue, WikiLeaks has become a full–blown player both on the world scene as well as in the national spheres of some countries. Small player as it is, WikiLeaks, by virtue of its disclosures, appears to be on a par with governments or big corporations (its next target) –– at least in the domain of information gathering and publication. At same time, it is unclear whether this is a permanent feature or a temporary, hype–induced phenomenon –– WikiLeaks appears to believe the former, and that looks more and more likely to be the case. Despite being a puny non–state and non–corporate actor, in its fight against the US government WikiLeaks does

not believe it is punching above its weight –– and is starting to behave accordingly. One might call this the "Talibanization" stage of the postmodern "Flat World" theory, where scales, times and places are declared largely irrelevant. What counts is celebrity momentum and the intense accumulation of media attention. WikiLeaks manages to capture that attention by way of spectacular information hacks, where other parties, especially civil society groups and human rights organizations, are desperately struggling to get their message across. While the latter tend to play by the rules and seek legitimacy from dominant institutions, WikiLeaks' strategy is populist insofar that it taps into public disaffection with mainstream politics. Political legitimacy, for WikiLeaks, is no longer something graciously bestowed by the powers that be. WikiLeaks bypasses this Old World structure of power and instead goes to the source of political legitimacy in today's info–society: the rapturous banality of the spectacle. WikiLeaks brilliantly puts to use the "escape velocity" of IT, using IT to leave IT behind and rudely irrupt the realm of real–world politics.

Thesis 3

In the ongoing saga called "The Decline of the US Empire", WikiLeaks enters the stage as the slayer of a soft target. It would be difficult to imagine it being able to inflict quite same damage to the Russian or Chinese governments, or even to the Singaporean –– not to mention their "corporate" affiliates. In Russia or China, huge cultural and linguistic barriers are at work, not to speak of purely power–related ones, which would need to be surmounted. Vastly different constituencies are also factors there, even if we are speaking about the narrower (and allegedly more global) cultures and agendas of hackers, info–activists and investigative journalists. In that sense, WikiLeaks in its present manifestation remains a typically "western" product and cannot claim to be a truly universal or global undertaking.

Thesis 4

One of the main difficulties with explaining WikiLeaks arises from the fact that it is unclear (also to the WikiLeaks people themselves) whether it sees itself and operates as a content provider or as a simple conduit for leaked data (the impression is that it sees itself as either/or, depending on context and circumstances). This, by the way, has been a common problem ever since media went online en masse and publishing and communications became a service rather than a product. Julian Assange cringes every time he is portrayed as the editor–in–chief of WikiLeaks; yet WikiLeaks says it edits material before publication and claims it checks documents for authenticity with the help of hundreds of volunteer analysts. Content vs. carrier debates of this kind have been going on for decades among media activists, with no clear outcome. Instead of trying to resolve the inconsistency, it might be better to look for fresh approaches and develop new critical concepts for what has become a hybrid publishing practice involving actors far beyond the traditional domain of the professional news media. This might be why Assange and his collaborators refuse to be labelled in terms of "old categories" (journalists, hackers, etc.) and claim to represent a new *Gestalt* on the world information stage.

Thesis 5

The steady decline of investigative journalism caused by diminishing funding is an undeniable fact. Journalism these days amounts to little more than outsourced PR remixing. The continuous acceleration and over–crowding of

the so–called attention economy ensures there is no longer enough room for complicated stories. The corporate owners of mass circulation media are increasingly disinclined to see the workings and the politics of the global neoliberal economy discussed at length. The shift from information to infotainment has been embraced by journalists themselves, making it difficult to publish complex stories. WikiLeaks enters this state of affairs as an outsider, enveloped by the steamy ambiance of "citizen journalism", DIY news reporting in the blogosphere and even faster social media like Twitter. What WikiLeaks anticipates, but so far has been unable to organize, is the "crowd sourcing" of the interpretation of its leaked documents. That work, oddly, is left to the few remaining staff journalists of selected "quality" news media. Later, academics pick up the scraps and spin the stories behind the closed gates of publishing stables. But where is networked critical commentariat? WikiLeaks generates its capacity to inspire irritation at the big end of town precisely because of the transversal and symbiotic relation it holds with establishment media institutions. There's a lesson here for the multitudes –– get out of the ghetto and connect with the Oedipal other. Therein lies the conflictual terrain of the political.

Traditional investigative journalism used to consist of three phases: unearthing facts, crosschecking these and backgrounding them into an understandable discourse. WikiLeaks does the first, claims to do the second, but omits the third completely. This is symptomatic of a particular brand of open access ideology, where content production itself is externalized to unknown entities "out there". The crisis in investigative journalism is neither understood nor recognized. How productive entities are supposed to sustain themselves materially is left in the dark: it is simply presumed that analysis and interpretation will be taken up by the traditional news media. But this is not happening automatically. The saga of the Afghan War Logs and Cablegate demonstrate that WikiLeaks has to approach and negotiate with well–established traditional media to secure sufficient credibility. At the same time, these media outlets prove unable to fully process the material, inevitably filtering the documents according to their own editorial policies.

Thesis 6

WikiLeaks is a typical SPO (Single Person Organization, or "UPO": Unique Personality Organization). This means that the initiative taking, decision–making and execution is largely concentrated in the hands of a single individual. Like small and medium–sized businesses, the founder cannot be voted out, and, unlike many collectives, leadership does not rotate. This is not an uncommon feature within organizations, irrespective of whether they operate in the realm of politics, culture or the "civil society" sector. SPOs are recognizable, exciting, inspiring, and easy to feature in the media. Their sustainability, however, is largely dependent on the actions of their charismatic leader, and their functioning is difficult to reconcile with democratic values. This is also why they are difficult to replicate and do not scale up easily. Sovereign hacker Julian Assange is the identifying figurehead of WikiLeaks, the organization's notoriety and reputation merging with Assange's own. What WikiLeaks does and stands for becomes difficult to distinguish from Assange's rather agitated private life and his somewhat unpolished political opinions.

Thesis 7

WikiLeaks raises the question as to what hackers have in common with secret services, since an elective affinity between the two is unmistakable. The

love–hate relationship goes back to the very beginning of computing. One does not have to be a fan of German media theorist Friedrich Kittler or, for that matter, conspiracy theories, to acknowledge that the computer was born out of the military–industrial complex. From Alan Turing's deciphering of the Nazi Enigma code up to the role played by the first computers in the invention of the atomic bomb, from the cybernetics movement up to the Pentagon's involvement in the creation of the Internet –– the articulation between computational information and the military–industrial complex is well established. Computer scientists and programmers have shaped the information revolution and the culture of openness; but at the same time they have also developed encryption ("crypto"), closing access to data for the non–initiated. What some see as "citizen journalism" others call "info war".

WikiLeaks is also an organization deeply shaped by 1980s hacker culture, combined with the political values of techno–libertarianism that emerged in the 1990s. The fact that WikiLeaks was founded –– and to a large extent is still run –– by hard–core geeks is essential to understanding its values and moves. Unfortunately, this comes together with a good dose of the less savoury aspects of hacker culture. Not that idealism, the desire to contribute to making the world a better place, could be denied to WikiLeaks: on the contrary. But this brand of idealism (or, if you prefer, anarchism) is paired with a preference for conspiracies, an elitist attitude and a cult of secrecy (never mind condescension). This is not conducive to collaboration with like–minded people and groups, who are relegated to being the simple consumers of WikiLeaks output. The missionary zeal to enlighten the idiotic masses and "expose" the lies of government, the military and corporations is reminiscent of the well–known (or infamous) media–culture paradigm from the 1950s.

Thesis 8

Lack of commonality with congenial, "another world is possible" movements drives WikiLeaks to seek public attention by way of increasingly spectacular and risky disclosures, thereby gathering a constituency of often wildly enthusiastic, but generally passive supporters. Assange himself has stated that WikiLeaks has deliberately moved away from the "egocentric" blogosphere and assorted social media and nowadays collaborates only with professional journalists and human rights activists. Yet following the nature and quantity of WikiLeaks exposures from its inception up to the present day is eerily reminiscent of watching a firework display, and that includes a "grand finale" in the form of the doomsday–machine pitched, yet–to–be–unleashed "insurance" document (".aes256"). This raises serious doubts about the long–term sustainability of WikiLeaks itself, and possibly also of the WikiLeaks model. WikiLeaks operates with ridiculously small staff –– probably no more than a dozen of people form the core of its operation. While the extent and savviness of WikiLeaks' tech support is proved by its very existence, WikiLeaks' claim to several hundreds of volunteer analysts and experts is unverifiable and, to be frank, barely credible. This is clearly WikiLeaks Achilles' heel, not only from a risk and/or sustainability standpoint, but politically as well –– which is what matters to us here.

Thesis 9

WikiLeaks displays a stunning lack of transparency in its internal organization. Its excuse that "WikiLeaks needs to be completely opaque in order to force others to be totally transparent" amounts, in our opinion, to little more than *Mad* magazine's famous Spy vs. Spy cartoons. You beat the opposition but in a

way that makes you indistinguishable from it. Claiming the moral high ground afterwards is not helpful –– Tony Blair too excelled in that exercise. As WikiLeaks is neither a political collective nor an NGO in the legal sense, and nor, for that matter, a company or part of social movement, we need to discuss what type of organization it is that we are dealing with. Is WikiLeaks a virtual project? After all, it does exist as a (hosted) website with a domain name, which is the bottom line. But does it have a goal beyond the personal ambition of its founder(s)? Is WikiLeaks reproducible? Will we see the rise of national or local chapters that keep the name? What rules of the game will they observe? Should we rather see it as a concept that travels from context to context and that, like a meme, transforms itself in time and space?

Thesis 10

Maybe WikiLeaks will organize itself around its own version of the Internet Engineering Task Force's slogan "rough consensus and running code"? Projects like Wikipedia and Indymedia have both resolved this issue in their own ways, but not without crises, conflicts and splits. A critique such as the one voiced here is not intended to force WikiLeaks into a traditional format; on the contrary, it is to explore whether WikiLeaks (and its future clones, associates, avatars and congenial family members) might stand as a model for new forms of organization and collaboration. The term "organized network" has been coined as a possible term for these formats. Another term has been "tactical media". Still others have used the generic term "internet activism". Perhaps WikiLeaks has other ideas about the direction it wants to take. But where? It is up to WikiLeaks to decide for itself. Up to now, however, we have seen very little by way of an answer, leaving others to raise questions, for example about the legality of WikiLeaks' financial arrangements (*Wall Street Journal*).

We cannot flee the challenge of experimenting with post–representational networks. As ur–blogger Dave Winer wrote about the Apple developers, "it's not that they're ill–intentioned, they're just ill–prepared. More than their users, they live in a Reality Distortion Field, and the people who make the Computer For the Rest of Us have no clue who the rest of us are and what we are doing. But that's okay, there's a solution. Do some research, ask some questions, and listen."

Thesis 11

The widely shared critique of the self–inflicted celebrity cult of Julian Assange invites the formulation of alternatives. Wouldn't it be better to run WikiLeaks as an anonymous collective or "organized network"? Some have expressed the wish to see many websites doing the same work. One group around Daniel Domscheit–Berg, who parted company with Assange in September, is already known to be working on a WikiLeaks clone. What is overlooked in this call for a proliferation of WikiLeaks is the amount of expert knowledge required to run a leak site successfully. Where is the ABC tool–kit of WikiLeaks? There is, perhaps paradoxically, much secrecy involved in this way of making–things–public. Simply downloading a WikiLeaks software kit and getting going is not a realistic option. WikiLeaks is not a plug 'n' play blog application like Wordpress, and the word "Wiki" in its name is really misleading, as Wikipedia's Jimmy Wales has been at pains to stress. Contrary to the collaboration philosophy of Wikipedia, WikiLeaks is a closed shop run with the help of an unknown number of faceless volunteers. One is forced to acknowledge that the know–how necessary to run a facility like WikiLeaks is

pretty arcane. Documents not only need to be received anonymously, but also to be further anonymized before they are released online. They also need to be "edited" before being dispatched to the servers of international news organizations and trusted, influential "papers of record".

WikiLeaks has built up a lot of trust and confidence over the years. Newcomers will need to go through that same, time–consuming process. The principle of WikiLeaks is not to "hack" (into state or corporate networks) but to facilitate insiders based in these large organisations to copy sensitive, confidential data and pass it on to the public domain –– while remaining anonymous. If you are aspiring to become a leak node, you'd better start to get acquainted with processes like OPSEC or operations security, a step–by–step plan which "identifies critical information to determine if friendly actions can be observed by adversary intelligence systems, determines if information obtained by adversaries could be interpreted to be useful to them, and then executes selected measures that eliminate or reduce adversary exploitation of friendly critical information" (Wikipedia). The WikiLeaks slogan says: "courage is contagious". According to experts, people who intend to run a WikiLeaks–type operation need nerves of steel. So before we call for one, ten, many WikiLeaks, let's be clear that those involved run risks. Whistleblower protection is paramount. Another issue is the protection of people mentioned in the leaks. The Afghan Warlogs showed that leaks can also cause "collateral damage". Editing (and eliding) is crucial. Not only OPSEC, also OPETHICS. If publishing is not carried out in a way that is absolutely secure for all concerned, there is a definite risk that the "revolution in journalism" –– and politics –– unleashed by WikiLeaks will be stopped in its tracks.

Thesis 12

We do not think that taking a stand for or against WikiLeaks is what matters most. WikiLeaks is here to stay, until it either scuttles itself or is destroyed by opposing forces. Our point is rather to (try to) assess and ascertain what WikiLeaks can, could –– and maybe even should –– do, and to help formulate how "we" could relate to and interact with WikiLeaks. Despite all its drawbacks, and against all odds, WikiLeaks has rendered a sterling service to the cause of transparency, democracy and openness. As the French would say, if something like it did not exist, it would have to be invented. The quantitative –– and what looks soon to become the qualitative –– turn of information overload is a fact of contemporary life. The glut of disclosable information can only be expected to continue grow –– and exponentially so. To organize and interpret this Himalaya of data is a collective challenge that is clearly out there, whether we give it the name "WikiLeaks" or not.

This is an extended version of an article first published on the nettime mailing list and elsewhere in August 2010

Published 2010–12–07
Original in English
First published in Eurozine (English version); *Frankfurter Rundschau* 07.12.10 (German version)

eurozine

Felix Stalder

Contain this!

Leaks, whistle–blowers and the networked news ecology

WikiLeaks' series of exposés is causing a very different news and informational landscape to emerge. Whilst acknowledging the structural leakiness of networked organisations, Felix Stalder finds deeper reasons for the crisis of information security and the new distribution of investigative journalism.

WikiLeaks is one of the defining stories of the Internet, which means by now, one of the defining stories of the present, period. At least four large–scale trends which permeate our societies as a whole are fused here into an explosive mixture whose fall–out is far from clear. First is a change in the materiality of communication. Communication becomes more extensive, more recorded, and the records become more mobile. Second is a crisis of institutions, particularly in western democracies, where moralistic rhetoric and the ugliness of daily practice are diverging ever more at the very moment when institutional personnel are being encouraged to think more for themselves. Third is the rise of new actors, "super–empowered" individuals, capable of intervening into historical developments at a systemic level. Finally, fourth is a structural transformation of the public sphere (through media consolidation at one pole, and the explosion of non–institutional publishers at the other), to an extent that rivals the one described by Habermas with the rise of mass media at the turn of the twentieth century.

Leaky containers

Imagine dumping nearly 400 000 paper documents into a dead drop located discreetly on the hard shoulder of a road. Impossible. Now imagine the same thing with digital records on a USB stick, or as an upload from any networked computer. No problem at all. Yet, the material differences between paper and digital records go much further than mere bulk. Digital records are the impulses travelling through the nervous systems of dynamic, distributed organisations of all sizes. They are intended, from the beginning, to circulate with ease. Otherwise such organisations would fall apart and dynamism would grind to a halt. The more flexible and distributed organisations become, the more records they need to produce and the faster these need to circulate. Due to their distributed aspect and the pressure for cross–organisational cooperation, it is increasingly difficult to keep records within particular organisations whose boundaries are blurring anyway. Surveillance researchers such as David Lyon have long been writing about the leakiness of "containers", meaning the tendency for sensitive digital records to cross the boundaries of the institutions which produce them. This leakiness is often driven by commercial considerations (private data being sold), but it happens also out of incompetence (systems being secured insufficiently), or because insiders

deliberately violate organisational policies for their own purposes. Either they are whistle–blowers motivated by conscience, as in the case of WikiLeaks, or individuals selling information for private gain, as in the case of the numerous employees of Swiss banks who recently copied the details of private accounts and sold them to tax authorities across Europe. Within certain organisation such as banks and the military, virtually everything is classified and large number of people have access to this data, not least mid–level staff who handle the streams of raw data such as individuals' records produced as part of daily procedure.

This basic data processing needs to be efficient, that is, data access and sharing has to be possible. It cannot be restricted by too much red tape, overly stringent security clearance requirements or the too strict compartmentalisation of the data into distinct sets that cannot be connected. After all, this inability to connect data located in different bureaucratic domains was one of the main criticisms coming out the enquires into the 9/11 attacks. There is an inherent paradox. Vast streams of classified records need to flow freely in order to sustain complex, distributed and time–sensitive operations. Yet, since the information is classified, it needs to flow within strict boundaries which cannot be clearly defined on a general level (after all, you never know what needs to get connected with what in advance), and it needs to flow through many, many hands. This creates the techno–organisational preconditions for massive amounts of information to leak out.

WikiLeaks, on the other side of the equation, created a custom–made infrastructure to receive these torrents of records. More than a decade after the heady discussions by cypherpunks who dreamed of total anonymity through full encryption, WikiLeaks managed for the first time to create an effective infrastructure for anonymous communication. Rather than relying purely on technology, they built social intelligence (filtering, editorial control) into the system in order to encourage only one type of anonymous speech –– whistle–blowing –– while insulating themselves from the usual criticisms of anonymous communication (child–porn trafficking and the like). Yet, the transformation of the materiality of records and the new infrastructures only create possibilities, and cannot single–handedly explain why certain containers are actually very leaky while others are not.

Institutions adrift

Is it a coincidence that so far the vast majority of WikiLeaks' material has originated from within institutions in democratic systems? I think not. In its rhetoric, Western politics is becoming ever more moralising. Tony Blair was the undisputed master in this discipline. He could speak passionately about "humanitarian wars" which were supposed to advance human rights. Afghanistan was to prosper under the warm attention of allied forces, following decades of neglect and civil war. This time, the invasion was going to develop the country, rebuild infrastructures, liberate women, give children hope and whatnot. The Iraq war –– once the weapons of mass destruction turned out to be imaginary –– was about liberating the Iraqi people from despotism, bringing democracy to the Middle East and ushering in a new era of peace, rule of law and commercial opportunities. All in all, these were just wars, wars we wanted to fight, wars soldiers could be proud of fighting. To some degree, there is always a gap between political rhetoric and practice, particularly in times of war. Yet, there is a qualitative difference now. Western political systems seem to have lost their ability to construct overarching historical narratives that would justify and give meaning to their actions and

make sense of the ugliness that is part of any war. Since the end of the Cold War, politics can no longer be said to pursue a historical project creating a void which has been papered over by empty moralising.

However, if a superficial morality is all that is left, then the encounter with the brutal day–to–day operations of the battle field is unmediated and corrosive. The moral rationale for going to war quickly dissolves under the actual experience of war and what's left is a cynical machinery run amok. It can no longer generate any lasting and positive identification from its protagonists. In some way, a similar lack of identification can be seen within corporations, as evidenced in the leaks from Swiss banks. With neoliberal ideology dominant, employees are told over and over not to expect anything from the company, that their job is continually in danger and that if they do not perform according to targets they can be replaced at a moment's notice. There is no greater narrative than the next quarter and generalised insecurity.

This emptying out of institutions takes place in the context of a general transformation of work away from strict hierarchies and the unquestioning execution of commands towards a more involved style of cognitive labour. Thus, people are told to engage more fully with their work, to become more creative, more self–reliant, more entrepreneurial. Simply following orders without investing one's creativity and personality is no longer enough. Thus, there is a second internal contradiction. People are asked to identify personally with organisations who can either no longer carry historical projects worthy of major sacrifices or expressly regard their employees as nothing but expendable, short–term resources. This, I think, creates the cognitive dissonance that justifies, perhaps even demands, the leaker to violate procedure and actively damage the organisation of which he, or she, has been at some point a well–acculturated member (this is the difference to the spy). This dissonance creates the motivational energy to move from the potential to the actual.

Super–empowered

There is a vast amount of infrastructure –– transportation, communication, financing, production –– openly available that, until recently, was only accessible to very large organisations. It now takes relatively little –– a few dedicated, knowledgeable people – to connect these pieces into a powerful platform from which to act. Military strategists have been talking about 'super–empowered individuals' by which they mean someone who

> is autonomously capable of creating a cascading event, [...] a "system perturbation"; a disruption of system function and invalidation of existing rule sets to at least the national but more likely the global scale. The key requirements to become "superempowered" are comprehension of a complex system's connectivty and operation; access to critical network hubs; possession of a force that can be leveraged against the structure of the system and a wilingness to use it.[1]

There are a number real weaknesses to this concept, not least that it has thus far been exclusively applied to terrorism and that it reduces structural dynamics to individual actions. Nevertheless, it can be useful insofar as it highlights how complex, networked systems which might be generally relatively stable, posses critical nodes ("systempunkt" in the strange parlance of military strategists) which in case of failure that can cause cascading effects

through the entire systems.[2] It also highlights how individuals, or more likely, small groups, can affect these systems disproportionately if they manage to interfere with these critical nodes. Thus, individuals, supported by small, networked organisations, can now intervene in social dynamics at a systemic level, for the better or worse.

This picture fits WikiLeaks, organised around one charismatic individual, very well. It is both its strength and its weakness. Its strength because it has been able to trigger large–scale events quickly and cheaply. If WikiLeaks had required multi–million dollar investment upfront, it would not have been able to get off the ground. Yet, it is also its key weakness, since it remains so strongly centred around a single person. Many of the issues that are typical of small groups organised by a charismatic leader seem to affect WikiLeaks as well, such as authoritarianism, lack of internal procedure, dangers of burnout and internal and external attacks on the credibility of that single person (if not worse). Such charismatic leadership is often unstable and one must suspect that all of the issues –– positive because of the super–empowerment, as well as negative because of the pressures baring down on it –– are multiplied to an unprecedented scale in the case of WikiLeaks and its leader, Julian Assange. It's hard to imagine how this can be sustainable.

A new public sphere

The public sphere as an arena for political discourse and a counter–balance to the state has been in decline for a very long time. While it's unclear when this decline started –– Habermas puts it at the beginning of the twentieth century –– it's fairly obvious that it has accelerated since the 1980s, following waves of deregulation and consolidation in the media business. Political and economic pressures led to an increase in the amount of "soft news", people stories, and commentary and to a decrease in the investigative reporting being conducted. That's a well–known story. At the same time governments have learned to play the game of access and leaks. Journalists are being skilfully fed with insider information and become increasingly dependent on having access to the centres of power. The embedded journalists at the beginning of the Second Gulf War were the most blatant example of this development. For both sides, this is a good arrangement. For the media, this is much faster and cheaper than doing its own research and for the government, it helps in controlling the story, not only by feeding the information (and even providing "experts" to be interviewed on TV), but also by threatening to withdraw access from journalists and media who don't tow the line. Last but not least, the legal protections of journalism are effectively weakened, in part because challenges are launched more aggressively, in part because commercial media view critical reporting through the eyes of their risk–averse legal and accounting departments.

In the face of the evident crisis of news media, there has been much hope that the Internet –– the blogosphere and citizen journalism –– would be able to replace the old, obsolete structures. On the whole, this has not happened, which is not surprising since new media never simply replace old media. What we can see, however, is a slow, structural transformation of the public sphere in which the old news media is complemented by new actors, designed to address the weaknesses of the mainstream media while making use of its core capacity to bring stories to lots of people. All in all, the process of investigative journalism is reorganised and, one can only hope, reinvigorated.

In a news ecology, the traditional news media remains the most important delivery channel for news. It knows best how to package and deliver news effectively. The legal risk associated with publishing sensitive information, however, is outsourced, to WikiLeaks in extreme cases, or to blogs and other operators without assets in normal circumstances. At the same time, there are new sources of funding and investigative journalism outside the main stream media. In the US, Pro–Publica, with philanthropic money from the Knights Foundation, was established in late 2007 as an "independent, non–profit newsroom that produces investigative journalism in the public interest", because "many news organisations have increasingly come to see it as a luxury."[3] In April 2010, the Bureau of Investigative Journalism was launched in London with a very similar aim, funded by the Potter Foundation it's also a not–for–profit.[4] Both units are partnering with traditional news media, print and television, which will carry the stories they investigate. In addition, new collaborative infrastructures, such as DocumentCloud, "an index of primary source documents and a tool for annotating, organizing and publishing them" is providing the infrastructure to cope with very large amounts of materials efficiently across newsrooms and organisational boundaries. The various elements that make up the process of investigative reporting, (protecting the source, doing the time–consuming work of gathering and making sense of information, providing the tools for handling the material, and delivering the story to the public at large), are no longer performed by a single organisation but by a networked set of organisations, mostly dedicated to performing only one of them well, and all based on different economic models but still working together to move the story into the public sphere. Within this new ecology, WikiLeaks is the actor taking on the most risk, which leaves the others relatively free to act within an otherwise highly constrained environment.

In a way, this changes the character of the final product, the news story, as well. It brings traditional reporting –– where source material is usually kept unpublished –– and blogging, where source material is usually linked to, closer together. Since WikiLeaks publishes the material anyway, many of the newspapers that turn its records into stories do that as well (rather than only quoting a sentence or two). On the whole, this makes the stories more transparent and, frankly, more interesting to read. As far as one can tell already, stories written with DocumentCloud tend to be similar. Whether this amounts to "scientific journalism" as Julian Assange hopes, journalism that publishes its source information the same way that scientists publish raw data and research methods, remains to be seen. But the combination of having access to a highly edited story as well as to sprawling source material could be very powerful.

Fall–out

It's very hard to assess the fall–out from WikiLeaks, since there are so many variables at play. It's pretty safe to say leaking will continue to be an important method of informational politics quite aside from the fate of WikiLeaks. What changed with WikiLeaks is the scale of the leaks –– both in terms of mass and sensitivity. Rather than playing politics as usual, WikiLeaks is capable of interfering within it and setting its own agenda. But what can it accomplish? The most modest goal stated by Assange is raising the "secrecy tax". As he wrote a few years ago in an essay on "The Non–linear Effects of Leaks on Unjust Systems of Governance",

> the more secretive or unjust an organisation is, the more leaks induce fear and paranoia in its leadership and planning coterie.

> This must result in minimization of efficient internal communications mechanisms (an increase in cognitive "secrecy tax") and consequent system–wide cognitive decline.[5]

The more an organisation has to protect against leaks, the more the internal contradiction between the requirement to share information (to operate efficiently) and that of controlling information (to keep it secret) will become prevalent and negatively affect its capacity to carry out its mission. Assange's objectives are likely to be realised in this more narrow respect, but it is unclear whether the "tax" will be high enough to limit the power of organisations such as the US military, or whether it will simply need to invest more resources to carry on doing the same thing as before.

Beyond this, much will depend on how long WikiLeaks can operate. The pressures that bear down on it are tremendous and its institutional base seems relatively feeble, despite, or more likely, because of super–empowerment. How far the media partnerships with the new ecology of journalism will support it, is also far from clear. *The New York Times*, for example, is playing it both ways. It is selectively working with WikiLeaks, but toning down its coverage.[6] Its avoidance of the term "torture" has become so strenuous that even the unpolitical blog BoingBoing mocked it by scripting "The New York Times Torture Euphemism Generator!".[7] At the same time, the NYT is actively participating in the global smear campaign against Assange.[8] The other mainstream media in the US are more openly hostile and are continuing their spin. Fox News claimed that the Iraq War Log contained information about weapons of mass destruction and one of its commentators demanded that WikiLeaks activists be declared "enemy combatants" and called for "non–judicial action", meaning targeted killings, against them.[9] As long as the (US) media remains so dependent on insider access to power (to receive the officially leaked information), their willingness to engage fully with the material published by WikiLeaks will be limited. Their engagement, however, will be critical since the interpretation and the political consequences of the leaks will not depend on the facts alone.

---

1 'The Super Empowered Individual', Zenpundit, 28 October 2006, http://zenpundit.blogspot.com/2006/10/super–empowered–individual–man–is.html

2 John Robb, "The Systempunkt", Global Guerillas, 19 December 2004, http://globalguerrillas.typepad.com/globalguerrillas/2004/12/the_systempunkt.html

3 http://www.propublica.org/about/

4 http://thebureauinvestigates.com/

5 http://cryptome.org/0002/ja–conspiracies.pdf

6 See Glenn Greenwald, 'NYT v. the World: WikiLeaks coverage', Salon, 25 October 2010, http://www.salon.com/news/media_criticism/index.html?story=/opinion/greenwald/2010/10/25/nyt

7 http://www.boingboing.net/2010/10/22/torture.html

8 John F. Burns and Ravi Somaiya, 'WikiLeaks Founder on the Run, Trailed by Notoriety', 23 October 2010, http://www.nytimes.com/2010/10/24/world/24assange.html?hp

9 Stephen C. Webster, "Fox News editorial: WikiLeaks employees should be declared 'enemy combatants'", 25 October 2010, http://www.rawstory.com/rs/2010/10/fox–news–editorial–wikileaks–employees–declared–enemy–combatants/

Published 2010–11–29
Original in English
Contribution by Mute
First published in www.metamute.org

The Blast Shack

22 December 2010

We asked Bruce Sterling (who spoke at Webstock '09) for his take on Wikileaks.

The Wikileaks Cablegate scandal is the most exciting and interesting hacker scandal ever. I rather commonly write about such things, and I'm surrounded by online acquaintances who take a burning interest in every little jot and tittle of this ongoing saga. So it's going to take me a while to explain why this highly newsworthy event fills me with such a chilly, deadening sense of Edgar Allen Poe melancholia.

But it sure does.

Part of this dull, icy feeling, I think, must be the agonizing slowness with which this has happened. At last — at long last — the homemade nitroglycerin in the old cypherpunks blast shack has gone off. Those "cypherpunks," of all people.

Way back in 1992, a brainy American hacker called Timothy C. May made up a sci-fi tinged idea that he called "The Crypto Anarchist Manifesto." This exciting screed — I read it at the time, and boy was it ever cool — was all about anonymity, and encryption, and the Internet, and all about how wacky data-obsessed subversives could get up to all kinds of globalized mischief without any fear of repercussion from the blinkered authorities. If you were of a certain technoculture bent in the early 1990s, you had to love a thing like that.

As Tim blithely remarked to his fellow encryption enthusiasts, "The State will of course try to slow or halt the spread of this technology, citing national security concerns, use of the technology by drug dealers and tax evaders, and fears of societal disintegration. Many of these concerns will be valid; crypto anarchy will allow national secrets to be traded freely," and then Tim started getting really interesting. Later, May described an institution called "BlackNet" which might conceivably carry out these aims.

Nothing much ever happened with Tim May's imaginary BlackNet. It was the kind of out-there concept that science fiction writers like to put in novels. Because BlackNet was clever, and fun to think about, and it made impossible things seem plausible, and it was fantastic and also quite titillating. So it was the kind of farfetched but provocative issue that ought to be properly raised within a sci-fi public discourse. Because, you know, that would allow plenty of time to contemplate the approaching trainwreck and perhaps do something practical about it.

Nobody did much of anything practical. For nigh on twenty long years, nothing happened with the BlackNet notion, for good or ill. Why? Because thinking hard and eagerly about encryption involves a certain mental composition which is alien to normal public life. Crypto guys — (and the cypherpunks were all crypto guys, mostly well-educated, mathematically gifted middle-aged guys in Silicon Valley careers) — are geeks. They're harmless geeks, they're not radical politicians or dashing international

crime figures.

Cypherpunks were visionary Californians from the WIRED magazine circle. In their personal lives, they were as meek and low-key as any average code-cracking spook who works for the National Security Agency. These American spooks from Fort Meade are shy and retiring people, by their nature. In theory, the NSA could create every kind of flaming scandalous mayhem with their giant Echelon spy system — but in practice, they would much rather sit there gently reading other people's email.

One minute's thought would reveal that a vast, opaque electronic spy outfit like the National Security Agency is exceedingly dangerous to democracy. Really, it is. The NSA clearly violates all kinds of elementary principles of constitutional design. The NSA is the very antithesis of transparency, and accountability, and free elections, and free expression, and separation of powers — in other words, the NSA is a kind of giant, grown-up, anti-Wikileaks. And it always has been. And we're used to that. We pay no mind.

The NSA, this crypto empire, is a long-lasting fact on the ground that we've all informally agreed not to get too concerned about. Even foreign victims of the NSA's machinations can't seem to get properly worked-up about its capacities and intrigues. The NSA has been around since 1947. It's a little younger than the A-Bomb, and we don't fuss much about that now, either.

The geeks who man the NSA don't look much like Julian Assange, because they have college degrees, shorter haircuts, better health insurance and far fewer stamps in their passports. But the sources of their power are pretty much identical to his. They use computers and they get their mitts on info that doesn't much wanna be free.

Every rare once in a while, the secretive and discreet NSA surfaces in public life and does something reprehensible, such as defeating American federal computer-security initiatives so that they can continue to eavesdrop at will. But the NSA never becomes any big flaming Wikileaks scandal. Why? Because, unlike their wannabe colleagues at Wikileaks, the apparatchiks of the NSA are not in the scandal business. They just placidly sit at the console, reading everybody's diplomatic cables.

This is their function. The NSA is an eavesdropping outfit. Cracking the communications of other governments is its reason for being. The NSA are not unique entities in the shadows of our planet's political landscape. Every organized government gives that a try. It's a geopolitical fact, although it's not too discreet to dwell on it.

You can walk to most any major embassy in any major city in the world, and you can see that it is festooned with wiry heaps of electronic spying equipment. Don't take any pictures of the roofs of embassies, as they grace our public skylines. Guards will emerge to repress you.

Now, Tim May and his imaginary BlackNet were the sci-fi extrapolation version of the NSA. A sort of inside-out, hippiefied NSA. Crypto people were always keenly aware of the NSA, for the NSA were the people who harassed them for munitions violations and struggled to suppress their academic publications. Creating a BlackNet is like having a pet, desktop NSA. Except, that instead of being a

vast, federally-supported nest of supercomputers under a hill in Maryland, it's a creaky, homemade, zero-budget social-network site for disaffected geeks.

But who cared about that wild notion? Why would that amateurish effort ever matter to real-life people? It's like comparing a mighty IBM mainframe to some cranky Apple computer made inside a California garage. Yes, it's almost that hard to imagine.

So Wikileaks is a manifestation of something that has been growing all around us, for decades, with volcanic inexorability. The NSA is the world's most public unknown secret agency. And for four years now, its twisted sister Wikileaks has been the world's most blatant, most publicly praised, encrypted underground site.

Wikileaks is "underground" in the way that the NSA is "covert"; not because it's inherently obscure, but because it's discreetly not spoken about.

The NSA is "discreet," so, somehow, people tolerate it. Wikileaks is "transparent," like a cardboard blast shack full of kitchen-sink nitroglycerine in a vacant lot.

That is how we come to the dismal saga of Wikileaks and its ongoing Cablegate affair, which is a melancholy business, all in all. The scale of it is so big that every weirdo involved immediately becomes a larger-than-life figure. But they're not innately heroic. They're just living, mortal human beings, the kind of geeky, quirky, cyberculture loons that I run into every day. And man, are they ever going to pay.

Now we must contemplate Bradley Manning, because he was the first to immolate himself. Private Manning was a young American, a hacker-in-uniform, bored silly while doing scarcely necessary scutwork on a military computer system in Iraq. Private Manning had dozens of reasons for becoming what computer-security professionals call the "internal threat."

His war made no sense on its face, because it was carried out in a headlong pursuit of imaginary engines of mass destruction. The military occupation of Iraq was endless. Manning, a tender-hearted geek, was overlooked and put-upon by his superiors. Although he worked around the clock, he had nothing of any particular military consequence to do.

It did not occur to his superiors that a bored soldier in a poorly secured computer system would download hundreds of thousands of diplomatic cables. Because, well, why? They're very boring. Soldiers never read them. The malefactor has no use for them. They're not particularly secret. They've got nothing much to do with his war. He knows his way around the machinery, but Bradley Manning is not any kind of blackhat programming genius.

Instead, he's very like Jerome Kerviel, that obscure French stock trader who stole 5 billion euros without making one dime for himself. Jerome Kerviel, just like Bradley Manning, was a bored, resentful, lower-echelon guy in a dead end, who discovered some awesome capacities in his system that his bosses never knew it had. It makes so little sense to behave like Kerviel and Manning that their threat can't be imagined. A weird hack like that is self-defeating, and it's sure to bring terrible repercussions to the transgressor. But then the sad and sordid days grind on and on; and that blindly

potent machinery is just sitting there. Sitting there, tempting the user.

Bradley Manning believes the sci-fi legendry of the underground. He thinks that he can leak a quarter of a million secret cables, protect himself with neat-o cryptography, and, magically, never be found out. So Manning does this, and at first he gets away with it, but, still possessed by the malaise that haunts his soul, he has to brag about his misdeed, and confess himself to a hacker confidante who immediately ships him to the authorities.

No hacker story is more common than this. The ingenuity poured into the machinery is meaningless. The personal connections are treacherous. Welcome to the real world.

So Private Manning, cypherpunk, is immediately toast.

No army can permit this kind of behavior and remain a functional army; so Manning is in solitary confinement and he is going to be court-martialled. With more political awareness, he might have made himself a public martyr to his conscience; but he lacks political awareness. He has only his black-hat hacker awareness, which is all about committing awesome voyeuristic acts of computer intrusion and imagining you can get away with that when it really matters to people.

The guy preferred his hacker identity to his sworn fidelity to the uniform of a superpower. The shear-forces there are beyond his comprehension.

The reason this upsets me is that I know so many people just like Bradley Manning. Because I used to meet and write about hackers, “crackers,” “darkside hackers,” “computer underground” types. They are a subculture, but once you get used to their many eccentricities, there is nothing particularly remote or mysterious or romantic about them. They are banal. Bradley Manning is a young, mildly brainy, unworldly American guy who probably would have been pretty much okay if he’d been left alone to skateboard, read comic books and listen to techno music.

Instead, Bradley had to leak all over the third rail. Through historical circumstance, he’s become a miserable symbolic point-man for a global war on terror. He doesn’t much deserve that role. He’s got about as much to do with the political aspects of his war as Monica Lewinsky did with the lasting sexual mania that afflicts the American Republic.

That is so dispiriting and ugly. As a novelist, I never think of Monica Lewinsky, that once-everyday young woman, without a sense of dread at the freakish, occult fate that overtook her. Imagine what it must be like, to wake up being her, to face the inevitability of being That Woman. Monica, too, transgressed in apparent safety and then she had the utter foolishness to brag to a lethal enemy, a trusted confidante who ran a tape machine and who brought her a mediated circus of hells. The titillation of that massive, shattering scandal has faded now. But think of the quotidian daily horror of being Monica Lewinsky, and that should take a bite from the soul.

Bradley Manning now shares that exciting, oh my God, Monica Lewinsky, tortured media-freak condition. This mild little nobody has become super-famous, and in his lonely military brig, screenless and without a computer, he’s strictly confined and, no doubt, he’s horribly bored. I don’t want to

condone or condemn the acts of Bradley Manning. Because legions of people are gonna do that for me, until we're all good and sick of it, and then some. I don't have the heart to make this transgressor into some hockey-puck for an ideological struggle. I sit here and I gloomily contemplate his all-too-modern situation with a sense of Sartrean nausea.

Commonly, the authorities don't much like to crush apple-cheeked white-guy hackers like Bradley Manning. It's hard to charge hackers with crimes, even when they gleefully commit them, because it's hard to find prosecutors and judges willing to bone up on the drudgery of understanding what they did. But they've pretty much got to make a purée out of this guy, because of massive pressure from the gravely embarrassed authorities. Even though Bradley lacks the look and feel of any conventional criminal; wrong race, wrong zipcode, wrong set of motives.

Bradley's gonna become a "spy" whose "espionage" consisted of making the activities of a democratic government visible to its voting population. With the New York Times publishing the fruits of his misdeeds. Some set of American prosecutorial lawyers is confronting this crooked legal hairpin right now. I feel sorry for them.

Then there is Julian Assange, who is a pure-dye underground computer hacker. Julian doesn't break into systems at the moment, but he's not an "ex-hacker," he's the silver-plated real deal, the true avant-garde. Julian is a child of the underground hacker milieu, the digital-native as twenty-first century cypherpunk. As far as I can figure, Julian has never found any other line of work that bore any interest for him.

Through dint of years of cunning effort, Assange has worked himself into a position where his "computer crimes" are mainly political. They're probably not even crimes. They are "leaks." Leaks are nothing special. They are tidbits from the powerful that every journalist gets on occasion, like crumbs of fishfood on the top of the media tank.

Only, this time, thanks to Manning, Assange has brought in a massive truckload of media fishfood. It's not just some titillating, scandalous, floating crumbs. There's a quarter of a million of them. He's become the one-man global McDonald's of leaks.

Ever the detail-freak, Assange in fact hasn't shipped all the cables he received from Manning. Instead, he cunningly encrypted the cables and distributed them worldwide to thousands of fellow-travellers. This stunt sounds technically impressive, although it isn't. It's pretty easy to do, and nobody but a cypherpunk would think that it made any big difference to anybody. It's part and parcel of Assange's other characteristic activities, such as his inability to pack books inside a box while leaving any empty space.

While others stare in awe at Assange's many otherworldly aspects — his hairstyle, his neatness, his too-precise speech, his post-national life out of a laptop bag — I can recognize him as pure triple-A outsider geek. Man, I know a thousand modern weirdos like that, and every single one of them seems to be on my Twitter stream screaming support for Assange because they can recognize him as a brother and a class ally. They are in holy awe of him because, for the first time, their mostly-imaginary and lastingly resentful underclass has landed a serious blow in a public arena. Julian

Assange has hacked a superpower.

He didn't just insult the captain of the global football team; he put spycams in the locker room. He showed the striped-pants set without their pants. This a massively embarrassing act of technical voyeurism. It's like Monica and her stains and kneepads, only even more so.

Now, I wish I could say that I feel some human pity for Julian Assange, in the way I do for the hapless, one-shot Bradley Manning, but I can't possibly say that. Pity is not the right response, because Assange has carefully built this role for himself. He did it with all the minute concentration of some geek assembling a Rubik's Cube.

In that regard, one's hat should be off to him. He's had forty years to learn what he was doing. He's not some miserabilist semi-captive like the uniformed Bradley Manning. He's a darkside player out to stick it to the Man. The guy has surrounded himself with the cream of the computer underground, wily old rascals like Rop Gonggrijp and the fearsome Teutonic minions of the Chaos Computer Club.

Assange has had many long, and no doubt insanely detailed, policy discussions with all his closest allies, about every aspect of his means, motives and opportunities. And he did what he did with fierce resolve.

Furthermore, and not as any accident, Assange has managed to alienate everyone who knew him best. All his friends think he's nuts. I'm not too thrilled to see that happen. That's not a great sign in a consciousness-raising, power-to-the-people, radical political-leader type. Most successful dissidents have serious people skills and are way into revolutionary camaraderie and a charismatic sense of righteousness. They're into kissing babies, waving bloody shirts, and keeping hope alive. Not this chilly, eldritch guy. He's a bright, good-looking man who — let's face it — can't get next to women without provoking clumsy havoc and a bitter and lasting resentment. That's half the human race that's beyond his comprehension there, and I rather surmise that, from his stern point of view, it was sure to be all their fault.

Assange was in prison for a while lately, and his best friend in the prison was his Mom. That seems rather typical of him. Obviously Julian knew he was going to prison; a child would know it. He's been putting on his Solzhenitsyn clothes and combing his forelock for that role for ages now. I'm a little surprised that he didn't have a more organized prison-support committee, because he's a convicted computer criminal who's been through this wringer before. Maybe he figures he'll reap more glory if he's martyred all alone.

I rather doubt the authorities are any happier to have him in prison. They pretty much gotta feed him into their legal wringer somehow, but a botched Assange show-trial could do colossal damage. There's every likelihood that the guy could get off. He could walk into an American court and come out smelling of roses. It's the kind of show-trial judo every repressive government fears.

It's not just about him and the burning urge to punish him; it's about the public risks to the reputation of the USA. The superpower hypocrisy here is gonna be hard to bear. The USA loves to read other people's diplomatic cables. They dote on doing it. If Assange had happened to out the

cable-library of some outlaw pariah state, say, Paraguay or North Korea, the US State Department would be heaping lilies at his feet. They'd be a little upset about his violation of the strict proprieties, but they'd also take keen satisfaction in the hilarious comeuppance of minor powers that shouldn't be messing with computers, unlike the grandiose, high-tech USA.

Unfortunately for the US State Department, they clearly shouldn't have been messing with computers, either. In setting up their SIPRnet, they were trying to grab the advantages of rapid, silo-free, networked communication while preserving the hierarchical proprieties of official confidentiality. That's the real issue, that's the big modern problem; national governments and global computer networks don't mix any more. It's like trying to eat a very private birthday cake while also distributing it. That scheme is just not working. And that failure has a face now, and that's Julian Assange.

Assange didn't liberate the dreadful secrets of North Korea, not because the North Koreans lack computers, but because that isn't a cheap and easy thing that half-a-dozen zealots can do. But the principle of it, the logic of doing it, is the same. Everybody wants everybody else's national government to leak. Every state wants to see the diplomatic cables of every other state. It will bend heaven and earth to get them. It's just, that sacred activity is not supposed to be privatized, or, worse yet, made into the no-profit, shareable, have-at-it fodder for a network society, as if global diplomacy were so many mp3s. Now the US State Department has walked down the thorny road to hell that was first paved by the music industry. Rock and roll, baby.

Now, in strict point of fact, Assange didn't blandly pirate the massive hoard of cables from the US State Department. Instead, he was busily "redacting" and minutely obeying the proprieties of his political cover in the major surviving paper dailies. Kind of a nifty feat of social-engineering there; but he's like a poacher who machine-gunned a herd of wise old elephants and then went to the temple to assume the robes of a kosher butcher. That is a world-class hoax.

Assange is no more a "journalist" than he is a crypto mathematician. He's a darkside hacker who is a self-appointed, self-anointed, self-educated global dissident. He's a one-man Polish Solidarity, waiting for the population to accrete around his stirring propaganda of the deed. And they are accreting; not all of 'em, but, well, it doesn't take all of them.

Julian Assange doesn't want to be in power; he has no people skills at all, and nobody's ever gonna make him President Vaclav Havel. He's certainly not in it for the money, because he wouldn't know what to do with the cash; he lives out of a backpack, and his daily routine is probably sixteen hours online. He's not gonna get better Google searches by spending more on his banned MasterCard. I don't even think Assange is all that big on ego; I know authors and architects, so I've seen much worse than Julian in that regard. He's just what he is; he's something we don't yet have words for.

He's a different, modern type of serious troublemaker. He's certainly not a "terrorist," because nobody is scared and no one got injured. He's not a "spy," because nobody spies by revealing the doings of a government to its own civil population. He is orthogonal. He's asymmetrical. He panics people in power and he makes them look stupid. And I feel sorry for them. But sorrier for the rest of us.

Julian Assange's extremely weird version of dissident "living in truth" doesn't bear much relationship

to the way that public life has ever been arranged. It does, however, align very closely to what we've done to ourselves by inventing and spreading the Internet. If the Internet was walking around in public, it would look and act a lot like Julian Assange. The Internet is about his age, and it doesn't have any more care for the delicacies of profit, propriety and hierarchy than he does.

So Julian is heading for a modern legal netherworld, the slammer, the electronic parole cuff, whatever; you can bet there will be surveillance of some kind wherever he goes, to go along with the FREE ASSANGE stencils and xeroxed flyers that are gonna spring up in every coffee-bar, favela and university on the planet. A guy as personally hampered and sociopathic as Julian may in fact thrive in an inhuman situation like this. Unlike a lot of keyboard-hammering geeks, he's a serious reader and a pretty good writer, with a jailhouse-lawyer facility for pointing out weaknesses in the logic of his opponents, and boy are they ever. Weak, that is. They are pathetically weak.

Diplomats have become weak in the way that musicians are weak. Musicians naturally want people to pay real money for music, but if you press them on it, they'll sadly admit that they don't buy any music themselves. Because, well, they're in the business, so why should they? And the same goes for diplomats and discreet secrets.

The one grand certainty about the consumers of Cablegate is that diplomats are gonna be reading those stolen cables. Not hackers: diplomats. Hackers bore easily, and they won't be able to stand the discourse of intelligent trained professionals discussing real-life foreign affairs.

American diplomats are gonna read those stolen cables, though, because they were supposed to read them anyway, even though they didn't. Now, they've got to read them, with great care, because they might get blindsided otherwise by some wisecrack that they typed up years ago.

And, of course, every intelligence agency and every diplomat from every non-American agency on Earth is gonna fire up computers and pore over those things. To see what American diplomacy really thought about them, or to see if they were ignored (which is worse), and to see how the grownups ran what was basically a foreign-service news agency that the rest of us were always forbidden to see.

This stark fact makes them all into hackers. Yes, just like Julian. They're all indebted to Julian for this grim thing that he did, and as they sit there hunched over their keyboards, drooling over their stolen goodies, they're all, without exception, implicated in his doings. Assange is never gonna become a diplomat, but he's arranged it so that diplomats henceforth are gonna be a whole lot more like Assange. They'll behave just like him. They receive the goods just like he did, semi-surreptitiously. They may be wearing an ascot and striped pants, but they've got that hacker hunch in their necks and they're staring into the glowing screen.

And I don't much like that situation. It doesn't make me feel better. I feel sorry for them and what it does to their values, to their self-esteem. If there's one single watchword, one central virtue, of the diplomatic life, it's "discretion." Not "transparency." Diplomatic discretion. Discretion is why diplomats do not say transparent things to foreigners. When diplomats tell foreigners what they really think, war results.

Diplomats are people who speak from nation to nation. They personify nations, and nations are brutal, savage, feral entities. Diplomats used to have something in the way of an international community, until the Americans decided to unilaterally abandon that in pursuit of Bradley Manning's oil war. Now nations are so badly off that they can't even get it together to coherently tackle heroin, hydrogen bombs, global warming and financial collapse. Not to mention the Internet.

The world has lousy diplomacy now. It's dysfunctional. The world corps diplomatique are weak, really weak, and the US diplomatic corps, which used to be the senior and best-engineered outfit there, is rattling around bottled-up in blast-proofed bunkers. It's scary how weak and useless they are.

US diplomats used to know what to do with dissidents in other nations. If they were communists they got briskly repressed, but if they had anything like a free-market outlook, then US diplomats had a whole arsenal of gentle and supportive measures; Radio Free Europe, publication in the West, awards, foreign travel, flattery, moral support; discreet things, in a word, but exceedingly useful things. Now they're harassing Julian by turning those tools backwards.

For a US diplomat, Assange is like some digitized nightmare-reversal of a kindly Cold War analog dissident. He read the dissident playbook and he downloaded it as a textfile; but, in fact, Julian doesn't care about the USA. It's just another obnoxious national entity. He happens to be more or less Australian, and he's no great enemy of America. If he'd had the chance to leak Australian cables he would have leapt on that with the alacrity he did on Kenya. Of course, when Assange did it to that meager little Kenya, all the grown-ups thought that was groovy; he had to hack a superpower in order to touch the third rail.

But the American diplomatic corps, and all it thinks it represents, is just collateral damage between Assange and his goal. He aspires to his transparent crypto-utopia in the way George Bush aspired to imaginary weapons of mass destruction. And the American diplomatic corps are so many Iraqis in that crusade. They're the civilian casualties.

As a novelist, you gotta like the deep and dark irony here. As somebody attempting to live on a troubled world… I dunno. It makes one want to call up the Red Cross and volunteer to fund planetary tranquilizers.

I've met some American diplomats; not as many as I've met hackers, but a few. Like hackers, diplomats are very intelligent people; unlike hackers, they are not naturally sociopathic. Instead, they have to be trained that way in the national interest. I feel sorry for their plight. I can enter into the shame and bitterness that afflicts them now.

The cables that Assange leaked have, to date, generally revealed rather eloquent, linguistically gifted American functionaries with a keen sensitivity to the feelings of aliens. So it's no wonder they were of dwindling relevance and their political masters paid no attention to their counsels. You don't have to be a citizen of this wracked and threadbare superpower — (you might, for instance, be from New Zealand) — in order to sense the pervasive melancholy of an empire in decline. There's a House of Usher feeling there. Too many prematurely buried bodies.

For diplomats, a massive computer leak is not the kind of sunlight that chases away corrupt misbehavior; it's more like some dreadful shift in the planetary atmosphere that causes ultraviolet light to peel their skin away. They're not gonna die from being sunburned in public without their pants on; Bill Clinton survived that ordeal, Silvio Berlusconi just survived it (again). No scandal lasts forever; people do get bored. Generally, you can just brazen it out and wait for the public to find a fresher outrage. Except.

It's the damage to the institutions that is spooky and disheartening; after the Lewinsky eruption, every American politician lives in permanent terror of a sex-outing. That's "transparency," too; it's the kind of ghastly sex-transparency that Julian himself is stuck crotch-deep in. The politics of personal destruction hasn't made the Americans into a frank and erotically cheerful people. On the contrary, the US today is like some creepy house of incest divided against itself in a civil cold war. "Transparency" can have nasty aspects; obvious, yet denied; spoken, but spoken in whispers. Very Edgar Allen Poe.

That's our condition. It's a comedy to those who think and a tragedy to those who feel, but it's not a comedy that the planet's general cultural situation is so clearly getting worse. As I sit here moping over Julian Assange, I'd love to pretend that this is just me in a personal bad mood; in the way that befuddled American pundits like to pretend that Julian is some kind of unique, demonic figure. He isn't. If he ever was, he sure as hell isn't now, as "Indoleaks," "Balkanleaks" and "Brusselsleaks" spring up like so many filesharing whackamoles. Of course the Internet bedroom legions see him, admire him, and aspire to be like him — and they will. How could they not?

Even though, as major political players go, Julian Assange seems remarkably deprived of sympathetic qualities. Most saintly leaders of the oppressed masses, most wannabe martyrs, are all keen to kiss-up to the public. But not our Julian; clearly, he doesn't lack for lust and burning resentment, but that kind of gregarious, sweaty political tactility is beneath his dignity. He's extremely intelligent, but, as a political, social and moral actor, he's the kind of guy who gets depressed by the happiness of the stupid.

I don't say these cruel things about Julian Assange because I feel distant from him, but, on the contrary, because I feel close to him. I don't doubt the two of us would have a lot to talk about. I know hordes of men like him; it's just that they are programmers, mathematicians, potheads and science fiction fans instead of fiercely committed guys who aspire to topple the international order and replace it with subversive wikipedians.

The chances of that ending well are about ten thousand to one. And I don't doubt Assange knows that. This is the kind of guy who once wrote an encryption program called "Rubberhose," because he had it figured that the cops would beat his password out of him, and he needed some code-based way to finesse his own human frailty. Hey, neat hack there, pal.

So, well, that's the general situation with this particular scandal. I could go on about it, but I'm trying to pace myself. This knotty situation is not gonna "blow over," because it's been building since 1993 and maybe even 1947. "Transparency" and "discretion" are virtues, but they are virtues that clash. The

international order and the global Internet are not best pals. They never were, and now that's obvious.

The data held by states is gonna get easier to steal, not harder to steal; the Chinese are all over Indian computers, the Indians are all over Pakistani computers, and the Russian cybermafia is brazenly hosting wikileaks.info because that's where the underground goes to the mattresses. It is a godawful mess. This is gonna get worse before it gets better, and it's gonna get worse for a long time. Like leaks in a house where the pipes froze.

Well... every once in a while, a situation that's one-in-a-thousand is met by a guy who is one in a million. It may be that Assange is, somehow, up to this situation. Maybe he's gonna grow in stature by the massive trouble he has caused. Saints, martyrs, dissidents and freaks are always wild-cards, but sometimes they're the only ones who can clear the general air. Sometimes they become the catalyst for historical events that somehow had to happen. They don't have to be nice guys; that's not the point. Julian Assange did this; he direly wanted it to happen. He planned it in nitpicky, obsessive detail. Here it is; a planetary hack.

I don't have a lot of cheery hope to offer about his all-too-compelling gesture, but I dare to hope he's everything he thinks he is, and much, much, more.

Bruce Sterling

Name (required)

Mail (will not be published) (required)

Website

Original URL:

http://www.webstock.org.nz/blog/2010/the-blast-shack/

London Review of Books

Good Manners in the Age of WikiLeaks

Slavoj Žižek

In one of the diplomatic cables released by WikiLeaks Putin and Medvedev are compared to Batman and Robin. It's a useful analogy: isn't Julian Assange, WikiLeaks's organiser, a real-life counterpart to the Joker in Christopher Nolan's *The Dark Knight*? In the film, the district attorney, Harvey Dent, an obsessive vigilante who is corrupted and himself commits murders, is killed by Batman. Batman and his friend police commissioner Gordon realise that the city's morale would suffer if Dent's murders were made public, so plot to preserve his image by holding Batman responsible for the killings. The film's take-home message is that lying is necessary to sustain public morale: only a lie can redeem us. No wonder the only figure of truth in the film is the Joker, its supreme villain. He makes it clear that his attacks on Gotham City will stop when Batman takes off his mask and reveals his true identity; to prevent this disclosure and protect Batman, Dent tells the press that he is Batman – another lie. In order to entrap the Joker, Gordon fakes his own death – yet another lie.

The Joker wants to disclose the truth beneath the mask, convinced that this will destroy the social order. What shall we call him? A terrorist? *The Dark Knight* is effectively a new version of those classic westerns *Fort Apache* and *The Man Who Shot Liberty Valance*, which show that, in order to civilise the Wild West, the lie has to be elevated into truth: civilisation, in other words, must be grounded on a lie. The film has been extraordinarily popular. The question is why, at this precise moment, is there this renewed need for a lie to maintain the social system?

Consider too the renewed popularity of Leo Strauss: the aspect of his political thought that is so relevant today is his elitist notion of democracy, the idea of the 'necessary lie'. Elites should rule, aware of the actual state of things (the materialist logic of power), and feed the people fables to keep them happy in their blessed ignorance. For Strauss, Socrates was guilty as charged: philosophy *is* a threat to society. Questioning the gods and the ethos of the city undermines the citizens' loyalty, and thus the basis of normal social life. Yet philosophy is also the highest, the worthiest, of human endeavours. The solution proposed was that philosophers keep their teachings secret, as in fact they did, passing them on by writing 'between the lines'. The true, hidden message contained in the 'great tradition' of philosophy

from Plato to Hobbes and Locke is that there are no gods, that morality is merely prejudice, and that society is not grounded in nature.

So far, the WikiLeaks story has been represented as a struggle between WikiLeaks and the US empire: is the publishing of confidential US state documents an act in support of the freedom of information, of the people's right to know, or is it a terrorist act that poses a threat to stable international relations? But what if this isn't the real issue? What if the crucial ideological and political battle is going on within WikiLeaks itself: between the radical act of publishing secret state documents and the way this act has been reinscribed into the hegemonic ideologico-political field by, among others, WikiLeaks itself?

This reinscription does not primarily concern 'corporate collusion', i.e. the deal WikiLeaks made with five big newspapers, giving them the exclusive right selectively to publish the documents. Much more important is the conspiratorial mode of WikiLeaks: a 'good' secret group attacking a 'bad' one in the form of the US State Department. According to this way of seeing things, the enemy is those US diplomats who conceal the truth, manipulate the public and humiliate their allies in the ruthless pursuit of their own interests. 'Power' is held by the bad guys at the top, and is not conceived as something that permeates the entire social body, determining how we work, think and consume. WikiLeaks itself got the taste of this dispersion of power when Mastercard, Visa, PayPal and Bank of America joined forces with the state to sabotage it. The price one pays for engaging in the conspiratorial mode is to be treated according to its logic. (No wonder theories abound about who is 'really' behind WikiLeaks – the CIA?)

The conspiratorial mode is supplemented by its apparent opposite, the liberal appropriation of WikiLeaks as another chapter in the glorious history of the struggle for the 'free flow of information' and the 'citizens' right to know'. This view reduces WikiLeaks to a radical case of 'investigative journalism'. Here, we are only a small step away from the ideology of such Hollywood blockbusters as *All the President's Men* and *The Pelican Brief*, in which a couple of ordinary guys discover a scandal which reaches up to the president, forcing him to step down. Corruption is shown to reach the very top, yet the ideology of such works resides in their upbeat final message: what a great country ours must be, when a couple of ordinary guys like you and me can bring down the president, the mightiest man on Earth!

The ultimate show of power on the part of the ruling ideology is to allow what appears to be powerful criticism. There is no lack of anti-capitalism today. We are overloaded with critiques of the horrors of capitalism: books, in-depth investigative journalism and TV documentaries expose the companies that are ruthlessly polluting our environment, the corrupt bankers who continue to receive fat bonuses while their banks are rescued by public money, the sweatshops in which children work as slaves etc. However, there is a catch: what isn't questioned in these critiques is the democratic-liberal framing of the fight against these excesses. The (explicit or implied) goal is to democratise capitalism, to extend democratic control to the economy by means of media pressure, parliamentary inquiries, harsher laws, honest police investigations and so on. But the institutional set-up of the (bourgeois) democratic state is never

questioned. This remains sacrosanct even to the most radical forms of 'ethical anti-capitalism' (the Porto Allegre forum, the Seattle movement etc).

WikiLeaks cannot be seen in the same way. There has been, from the outset, something about its activities that goes way beyond liberal conceptions of the free flow of information. We shouldn't look for this excess at the level of content. The only surprising thing about the WikiLeaks revelations is that they contain no surprises. Didn't we learn exactly what we expected to learn? The real disturbance was at the level of appearances: we can no longer pretend we don't know what everyone knows we know. This is the paradox of public space: even if everyone knows an unpleasant fact, saying it in public changes everything. One of the first measures taken by the new Bolshevik government in 1918 was to make public the entire corpus of tsarist secret diplomacy, all the secret agreements, the secret clauses of public agreements etc. There too the target was the entire functioning of the state apparatuses of power.

What WikiLeaks threatens is the formal functioning of power. The true targets here weren't the dirty details and the individuals responsible for them; not those in power, in other words, so much as power itself, its structure. We shouldn't forget that power comprises not only institutions and their rules, but also legitimate ('normal') ways of challenging it (an independent press, NGOs etc) – as the Indian academic Saroj Giri put it, WikiLeaks 'challenged power by challenging the normal channels of challenging power and revealing the truth'.[*] The aim of the WikiLeaks revelations was not just to embarrass those in power but to lead us to mobilise ourselves to bring about a different functioning of power that might reach beyond the limits of representative democracy.

However, it is a mistake to assume that revealing the entirety of what has been secret will liberate us. The premise is wrong. Truth liberates, yes, but not *this* truth. Of course one cannot trust the façade, the official documents, but neither do we find truth in the gossip shared behind that façade. Appearance, the public face, is never a simple hypocrisy. E.L. Doctorow once remarked that appearances are all we have, so we should treat them with great care. We are often told that privacy is disappearing, that the most intimate secrets are open to public probing. But the reality is the opposite: what is effectively disappearing is public space, with its attendant dignity. Cases abound in our daily lives in which not telling all is the proper thing to do. In *Baisers volés*, Delphine Seyrig explains to her young lover the difference between politeness and tact: 'Imagine you inadvertently enter a bathroom where a woman is standing naked under the shower. Politeness requires that you quickly close the door and say, "Pardon, Madame!", whereas tact would be to quickly close the door and say: "Pardon, Monsieur!"' It is only in the second case, by pretending not to have seen enough even to make out the sex of the person under the shower, that one displays true tact.

A supreme case of tact in politics is the secret meeting between Alvaro Cunhal, the leader of the Portuguese Communist Party, and Ernesto Melo Antunes, a pro-democracy member of the army grouping responsible for the coup that overthrew the Salazar regime in 1974. The situation was extremely tense: on one side, the Communist Party was ready to start the real

socialist revolution, taking over factories and land (arms had already been distributed to the people); on the other, conservatives and liberals were ready to stop the revolution by any means, including the intervention of the army. Antunes and Cunhal made a deal without stating it: there was no agreement between them – on the face of things, they did nothing but disagree – but they left the meeting with an understanding that the Communists would not start a revolution, thereby allowing a 'normal' democratic state to come about, and that the anti-socialist military would not outlaw the Communist Party, but accept it as a key element in the democratic process. One could claim that this discreet meeting saved Portugal from civil war. And the participants maintained their discretion even in retrospect. When asked about the meeting (by a journalist friend of mine), Cunhal said that he would confirm it took place only if Antunes didn't deny it – if Antunes did deny it, then it never took place. Antunes for his part listened silently as my friend told him what Cunhal had said. Thus, by not denying it, he met Cunhal's condition and implicitly confirmed it. This is how gentlemen of the left act in politics.

So far as one can reconstruct the events today, it appears that the happy outcome of the Cuban Missile Crisis, too, was managed through tact, the polite rituals of pretended ignorance. Kennedy's stroke of genius was to pretend that a letter had not arrived, a stratagem that worked only because the sender (Khrushchev) went along with it. On 26 October 1962, Khrushchev sent a letter to Kennedy confirming an offer previously made through intermediaries: the Soviet Union would remove its missiles from Cuba if the US issued a pledge not to invade the island. The next day, however, before the US had answered, another, harsher letter arrived from Khrushchev, adding more conditions. At 8.05 p.m. that day, Kennedy's response to Khrushchev was delivered. He accepted Khrushchev's 26 October proposal, acting as if the 27 October letter didn't exist. On 28 October, Kennedy received a third letter from Khrushchev agreeing to the deal. In such moments, when everything is at stake, appearances, politeness, the awareness that one is 'playing a game', matter more than ever.

However, this is only one – misleading – side of the story. There are moments – moments of crisis for the hegemonic discourse – when one should take the risk of provoking the disintegration of appearances. Such a moment was described by the young Marx in 1843. In 'Contribution to the Critique of Hegel's Philosophy of Law', he diagnosed the decay of the German ancien regime in the 1830s and 1840s as a farcical repetition of the tragic fall of the French ancien regime. The French regime was tragic 'as long as it believed and had to believe in its own justification'. The German regime 'only imagines that it believes in itself and demands that the world imagine the same thing. If it believed in its own *essence*, would it … seek refuge in hypocrisy and sophism? The modern ancien regime is rather only the *comedian* of a world order whose *true heroes* are dead.' In such a situation, shame is a weapon: 'The actual pressure must be made more pressing by adding to it consciousness of pressure, the shame must be made more shameful by publicising it.'

This is precisely our situation today: we face the shameless cynicism of a global order whose agents only imagine that they believe in their ideas of democracy, human rights and so on.

Through actions like the WikiLeaks disclosures, the shame – our shame for tolerating such power over us – is made more shameful by being publicised. When the US intervenes in Iraq to bring secular democracy, and the result is the strengthening of religious fundamentalism and a much stronger Iran, this is not the tragic mistake of a sincere agent, but the case of a cynical trickster being beaten at his own game.

[*] 'WikiLeaks beyond WikiLeaks?', www.metamute.org/en/articles/WikiLeaks_beyond_WikiLeaks.

Vol. 33 No. 2 · 20 January 2011 » Slavoj Žižek » Good Manners in the Age of WikiLeaks (print version)
pages 9-10 | 2314 words

Letters

Vol. 33 No. 3 · 3 February 2011

From Laurie Edmundson

The WikiLeaks revelations, like the attacks of 11 September, were one of those spectacular assaults on the symbols of power anarchists used to call the 'propaganda of the deed'. But, also like 9/11, WikiLeaks's info-guerrilla raid has unleashed such a complex chain of effects that it's no longer clear what the organisation intended to achieve – or whether those intentions even matter. Slavoj Zizek argues that its aim was 'to lead us to mobilise ourselves to bring about a different functioning of power that might reach beyond the limits of representative democracy' (*LRB*, 20 January). An intriguing speculation, but can one speak of a single 'aim' when hundreds of thousands of diplomatic cables are released, revealing the dirty laundry of most of the world's governments, not just those of the 'US empire'? And who is meant by 'we'?

Even if Julian Assange hoped to strike at the hegemon, it's worth noting that the Americans don't always come off so badly: US diplomatic cables certainly reveal a fair measure of hypocrisy, but they also show a highly competent foreign service, informed, insightful and capable of the occasional flash of humour. Might this be one reason why some autocrats – Muammar al-Gaddafi of Libya, for example – see WikiLeaks as a sinister American (or Israeli) conspiracy? Perhaps not surprisingly, the most dramatic effects of the WikiLeaks revelations have been felt not in the 'representative democracies' beyond whose 'limits' Zizek urges us to act, but in those countries where people would be grateful to enjoy a bit of democratic representation. One of the most fateful memos was written by the US ambassador in Tunis, describing the beachfront villa of former President Ben Ali's son-in-law, who decorated his home with Roman columns and frescoes, kept a pet tiger called Pasha and served his guests ice cream flown in from Saint-Tropez. The Tunisian uprising wasn't detonated by WikiLeaks, of course, but it didn't hurt, and the uprising is, at its core, an old-fashioned struggle for

representative democracy and transparency in a country that, for the last 50 years, has known only secrecy and dictatorship.

Laurie Edmundson
Chicago

From Gillian de Veras

Slavoj Žižek's article on Wikileaks provided a welcome counterpoint to the lionisation, in some sections of the press, of Julian Assange as some sort of champion of free speech. In fact he is endangering free speech. According to the terms of the Vienna Convention on Diplomatic Relations, to which almost every country in the world adheres, diplomats meeting in private, or communicating with their ministries during foreign postings, rely absolutely on recipients' respect for the security classification they have given their missives. If they suspect that their words will shortly be trumpeted in public by the likes of Assange, the whole machinery of international diplomacy will break down. Žižek's examples of the overthrow of the Salazar regime in Portugal in 1974 and the Cuban Missile Crisis in 1962 vividly show what dire consequences were avoided at those times by the use of diplomatic tact.

Gillian de Veras
London SE25

Vol. 33 No. 4 · 17 February 2011

From David Auerbach

Slavoj Žižek isn't quite right that WikiLeaks made a deal with 'five big newspapers, giving them the exclusive right selectively to publish the documents' (*LRB*, 20 January). WikiLeaks negotiated with four newspapers: the *Guardian*, *Der Spiegel*, *Le Figaro* and *El País*. The *Guardian* in turn leaked the documents to the *New York Times*. WikiLeaks also retains the right to publish the documents; it is not exclusively the right of the newspapers.

David Auerbach
Brooklyn, New York

^ Top