Revolution in Social Science Methodology: Possibilities and Pitfalls[1]

Dipak K. Gupta, Department of Political Science
Brian Spitzberg, School of Communication
Ming-Hsian Tsou, Department of Geography
Mark Gawron, Department of Linguistics
Li An, Department of Geography
San Diego State University

Abstract

There is a revolution taking placing in how social scientists research aggregate and synthesize the data and observations that constitute our understanding of social phenomena. These analyses received an unprecedented boost after the invention of high-speed computers. Traditional social science research has depended on statistical analyses of past data or opinion surveys to determine root causes of social and political events. Both of these take a long time to develop and derive results. Yet, in the world of the Internet and the proliferation of social media, world events, from the Arab Spring to the lesser-known social changes are taking place in breathtaking speed. The new methodology for this altered world comes through mining data from the publicly available databases and postings. In the past, where researchers used to worry about the sample size, the methods of data mining yield millions of observations. However, in order to collect and analyze the vast amount of data, we need to go far beyond where social disciplines have traditionally gone. These efforts must not only be imaginative, methodologically sophisticated, but also truly multi-disciplinary.

Introduction

There is a revolution taking placing in how social scientists research aggregate and synthesize the data and observations that constitute our understanding of social phenomena. These analyses received an unprecedented boost after the invention of

[1] Research for this article was supported by "Mapping ideas from Cyberspace to Realspace," funded by National Science Foundation, Division of Computer and Network Systems, NSF Program CDI-Type II Award # 1028177. Opinions expressed are those of the authors and not necessarily those of the National Science Foundation.

high-speed computers. Prior to the 1960s most significant advances in theory building in social sciences were relegated largely to what we may charitably called, "informed speculation." Karl Marx did not have to bother with the problem of falsifiability of his proposed theory (Popper, 1980). John Maynard Keynes advanced his general theory without formally statistically testing his hypotheses; he could not empirically demonstrate the link between aggregate demand and economic activities, the central tenet of his argument. Charles Darwin required months at sea, years more of correspondence and physical examination of specimens, as well as observation and collection of myriad examples to fully formulate the framework of his theory of natural selection. The social movement theorists of the period had to rely primarily of perceived plausibility of their postulates or attempt to offer proof by conducting small case studies. There were a number of significant problems for empirical verification of posited hypotheses, the hallmark of scientific reasoning (Bostrom, 2003; Chow, 1992; Gilman, 1992; Pavitt, 2004). All of these classic achievements of 20th century social science and theory occurred in the absence of computing technology. It is no accident that they were in their nascent forms rich with ideas and poor of data.

To begin with, without the computer, collection of large-scale aggregate data was problematic and extremely time consuming. The first installation of commercial computer, UNIVAC I, took place in 1951 in the U.S. Bureau of Census. Soon thereafter, a similar machine, employed by the CBS was used for the first time to predict the 1952 presidential election (Gupta, 1994). With the help of the new technology, 1 percent of the US population was surveyed and the TV news

organization was able to correctly predict the outcome of the election: the win of Dwight D. Eisenhower.

With a rapid increase in the ability to crunch numbers, strides were made in applied statistics. Multivariate regressions, which were relegated largely to theoretical development, suddenly became commonplace in estimation. As computing capabilities increased exponentially much more sophisticated statistical techniques became commonplace in social sciences. A virtual explosion began to take place in the collection of data on innumerable aspects of life. For instance, today, a nation's development does not need definition solely based on per capita GDP. It is definable much more broadly with multi-faceted Human Development Index.[2] There are cross-national indicators of state failure[3], corruption[4], and even gross national happiness[5].

As noted by Bell, Hey and Szalay (2009), "Today, some areas of science are facing hundred- to thousand-fold increases in data volumes from satellites, telescopes, high throughput instruments, sensor networks, accelerators, and supercomputers, compared to the volumes generated only a decade ago" (p. 1297). The problem is no longer obtaining information, but obtaining clarity in the midst of information overload.

---

[2] http://www.nationmaster.com/graph/eco_hum_dev_ind-economy-human-development-index

[3] http://www.foreignpolicy.com/articles/2011/06/17/2011_failed_states_index_interactive_map_and_rankings

[4] http://www.transparency.org/policy_research/surveys_indices/cpi/2010

[5] http://www.nsf.gov/news/newsmedia/pr111725/pr111725.pdf

With such developments in information retrieval, social science has entered a new era of potential empirical verification and prediction. The invention of computers enhanced the capabilities to analyze quantities of numbers beyond any previous human capabilities (Brier & Hopp, 2011; Johnson, Dunlap & Benoit, 2010).

Need for A New Methodology

Despite these breathtaking developments in data collection and computing capabilities, traditional social science research methodology suffers from some significant shortcomings and challenges (Ramakrishnan & Grama, 1999; Yang & Wu, 2006). These challenges stem less from the constraints of computational capability or methodological sophistication, but from the process by which social science data are collected.

Most social science data are collected through one of two ways. They are either collected from observations of direct indicators in the "real world" (e.g., votes counted in an election, instances of violence in the media, etc.) or sample populations are surveyed to learn about their preferences and attitudes. In either case, the process can be extremely time-consuming. The decennial census data, the most accurate portrayal of life in the nation takes more than a decade to collect, collate, and to publish. Even the quarterly data are lagged by time. Surveys – unless constituted by routine questions such as daily presidential approval ratings – take a long time to design, implement, and to analyze.

While data collection traditionally has taken a long time, the world events in the age of super connectivity and spatial mobility are moving at a faster rate than

anyone could have even imagined. In the past, it has generally taken years for collective movements to take shape. The Civil Rights movement took at least a couple of decades to reach its height. Al-Qaeda, similarly, had to struggle for years before it was recognized on the global stage. Compared to these, the lightning speed with which the so-called "Arab Spring" spread through the Middle East (Elson et al., 2012; Howard & Hussain, 2001; Murphy, 2006; Spitzberg et al., 2012) and the "Occupy Wall Street" (Ladhani, 2011) spilled over most of the developed world is virtually unprecedented (see, e.g., Hwang et al., 2006; Postmes & Brunsting, 2002; Rohlinger & Brown, 2009).

The rapid diffusion of ideas across various social domains is taking place primarily due to two important factors, both of which relate to the advancement in computer technology and electronic media. The first relates to the physical nature of the innovations whereas the second pertains to the psychology of the users. The ease of communication has indeed made efforts to communicate with a large number of people extremely cost effective and efficient (e.g., Andrés et al., 2010; Huang & Chen, 2010). In the past, people needed to have a "megaphone," offices or institutional pulpits that often came at the heavy expenditure of time and money. As a result, those who were in positions of public prominence had a "bully pulpit" from which to address large crowds or whatever constituted the local or national mass media of the day. Others, attempting a grass-root movement had to slog through years of work. In the contemporary technological context, messages can be spread through the web and myriad social media outlets and can put forward ideas that have not been previously vetted through years of socio-political group activities.

Second, as Olson (1968) pointed out, there has traditionally been a psychological impediment in voluntarily organizing large protest movements, known as the "collective action" problem. Some of the collective action problem resulted from the fact that the costs of exposing contrarian viewpoints would fall primarily on those who would dare to take the first step, whereas the benefits, if the common good is procured, would flow to everyone in the community. As a result, the collective action problem would dictate that even when there is a deeply help desire by a large segment of the population for a significant social change, such demand would probably not even be properly articulated (Gupta, 2001; 2008). Compared to contemporary capabilities to give voice to risky beliefs or to evidence physical group affiliations, the risks of having to incur stiff costs would weigh heavily in the minds of the early activists (Watts & Dodds, 2007).

In contrast, many activists feel—albeit often erroneously—a sense of anonymity when computers and electronic media are relied upon for the purposes of political mobilization (Lewis et al., 2008; Rohlinger & Brown, 2009; Saldarini & DeRobertis, 2003). Most people tend to have a barrier between their public pronouncements and their privately held beliefs (Kuran, 1988; Petronio, 2002). People possess different levels of their own personal "threshold point" of tolerance for the status quo. When these thresholds are breached, people come out of their cocoons and join a mass movement. The society may appear to be extremely stable until the day when people would come out and join the revolution. Significant "cascades can only occur when the influence network exhibits a 'critical mass' of

early adopters, … who adopt after they are exposed to a single adopting neighbor" or peer (Watts & Dodds, 2007).

Given that neighbors and peers often engage in false expressions of their actual preferences until a sufficient threshold of peers has adopted the actual preferred position (Kuran, 1989, 1995), radical positions are particularly risky in the context of the status quo. That is why, Kuran argued that it is impossible to predict the demise of established political systems, such as the Soviet Union or the Shah's Iran, Mubarak's Egypt or Assad's Syria (see also: Taleb & Blyth, 2011). Since many users, especially the neophytes, believe that the computer accords greater anonymity than face-to-face interactions or traditional print and video media, it is possible that in the present condition of technological advancement the threshold level has been reduced considerably (Spitzberg et al., 2012). With so many being willing to express their deeply held ambivalence or resistance online, more and more are emboldened to join such voiced causes, thereby forming large political movements almost out of thin air.

"Data Mining" and New Research Methodology

In view of such an altered world, it is clear that the social sciences and policy studies will need, and find, newer and far more sophisticated methods of collecting, extracting, synthesizing, and analyzing data to explain and predict social processes (e.g., Apinar et al., 2006; Bell et al., 2009; Cannataro & Talia, 2003; Lazer et al., 2009; Shaw et al., 2008; Zimbra et al., 2010). Given the lag in collecting and disseminating traditional data, the traditional research method is akin to

astronomers looking at a distant star; the light rays that hit the telescope are light years in the past and say nothing about its current condition. Thus, data mining will be increasingly necessary as an arrow in the arsenals of social science methodology. Human nature may not have accelerated, but the ability of the species to observe it and communicate it have.

The word "data mining" was once a dirty word in the social science research where it was seen as a way of fishing for answers by randomly examining a lot of information without any theoretical background (Gupta, 2010). However, in the field of computer-assisted research, the term, implying collecting information through monitoring the Internet, is finding new respectability (Leetaru, 2012; Ramakrishnan & Grama, 1999; Yang et al., 2006). As a result, the monitoring of the web sites, the Twitter and the like is open up new methods of understanding social interactions (e.g., Crandall et al., 2010; Erickson, 2010; Takhteyev, et al., 2011; Zimbra et al., 2010).

The monitoring of communications on the Internet creates a deluge of data (Bell et al., 2009). In the past, where researchers used to worry about the sample size, the methods of data mining yield millions of observations. Therefore, the first challenge for the researchers is to classify these in a meaningful way so that we can make some sense of our collective mood, positions, and opinions (Altaweel, Alessa, & Kliskey, 2010; Brier & Hopp, 2011; Cao, 2010). For instance, by monitoring the Twitter traffic, a group of researchers claimed to have observed the world's mood swings during the week, the year, and in response to momentous world events (Miller, 2011). Other research has demonstrated that Tweets can predict stock

market fluctuations (Bollen et al., 2011). Such research relies not on the direct observation of moods or people's self-reports about how they feel, but on empirical proxies or surrogates that would imply what individuals might be feeling. Critics abound; information culled from the cyberspace may indeed be perfunctory, misleading, or worse (Brasch, 2005). This, however, is similar to understanding economic achievement of nations by measuring illumination at night from satellite imagery. Or, for instance, the tonnage of trucking can provide an important clue to the future economic activities of a nation. These may not be perfect indexes, but when used judiciously they can be sufficiently good indicators of what we are attempting to measure.

Web sites, Social Media and the Twitter

We can collect publicly available Internet data from various sources. Among the most informative of these sources for social scientific at this time are (a) the contents of world wide web postings, (b) social media, such as open pages of Facebook, and (c) social media such as Twitter, Flickr, Pinterest, etc.[6] There are, however, some important differences among these sources of information (e.g., Papacharissi, 2009).

Are there differences among such media sources in terms of how people reveal their inner thoughts? Some of these differences are more obvious than others (Attrill & Jalil, 2011). For instance, while the Internet allows for expansive explanations of the posters' positions, the Twitter has a strict limitation of 140

[6] R. Kumar, J. Novak, and A. Tomkins (2010) in P.S. Yu, et al. (eds.), *Link Mining: Models, Algorithms, and Applications*, Springer Science+Business Media, LLC.

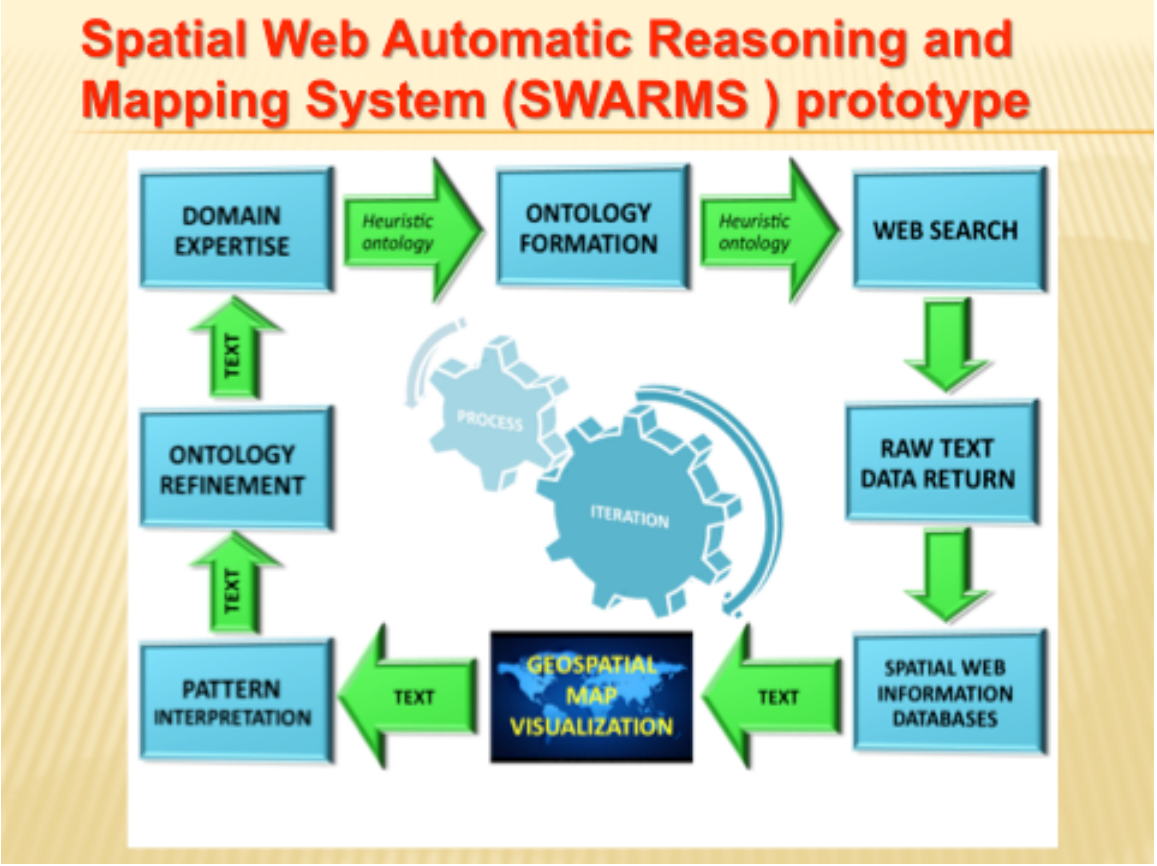
characters. However, while many Internet sites have a "private members' only" part or other restrictive privacy settings, the messages sent by Twitter are totally public, although stored for a relatively short period of time. During this time, the dataset is available to researchers. Twitter's reliance on hash tags allows search programs to trace tweets and followers' re-tweets, which allows researchers to seek insights into social interaction and networking by the users.

Our project introduced a new methodology for web search and web content analysis, called Spatial Web Automatic Reasoning and Mapping System (SWARMS) (http://mappingideas.sdsu.edu), to track ideas, events, and trends disseminated in cyberspace (the web) and social networks. This research method integrates GIScience, computational linguistics, and web search engines to track and analyze web page content identified by clusters of keywords. The search results were mapped with real world coordinates (by geolocating their IP addresses, URLs, or gazetteers). The resulting maps represented web information landscapes consisting of hundreds of populated web pages searched by selected keywords with time stamps. By utilizing Geographic Information Systems (GIS) and visualization methods, researchers can visualize the spread of concepts and the density of related web pages on a real world map over time and space (Figure 1). The new SWARMS prototype can help us visualize and analyze the space-time dimensions of the spread of information, concepts, and ideas posted on the publically-accessible web pages. Hundreds of web pages were geocoded with real world coordinates and represented in the form of web information landscapes (web page density maps). We hope that these web information landscapes can help us monitor the spatial and temporal distribution patterns

of web pages and reveal the nature of significant events, controversial concepts or epidemics. Understanding the diffusion and acquisition patterns of web information landscapes in response to disasters, terrorism, and epidemics has the potential to facilitate intervention and response, and eventually, prevention.

Figure 1.

A basic heuristic flow diagram depicting the use of search and interpretation processes to geo-locate the diffusion of symbolic content on the web



The "Buzz" and the "Narrative"

When we eavesdrop on peoples' electronic conversations, we are in essence looking for the buzz and the narrative. We may define "buzz" as concentration of

postings at a certain location. Thus, we may find that by conducting a search, we discover that there is a high cluster of postings on the location of a certain city, but not on others. In that case, we can attempt to understand the causes of the buzz based on statistical methods by using a number of demographic and other economic, political, or sociological independent variables (Hand, 2000).

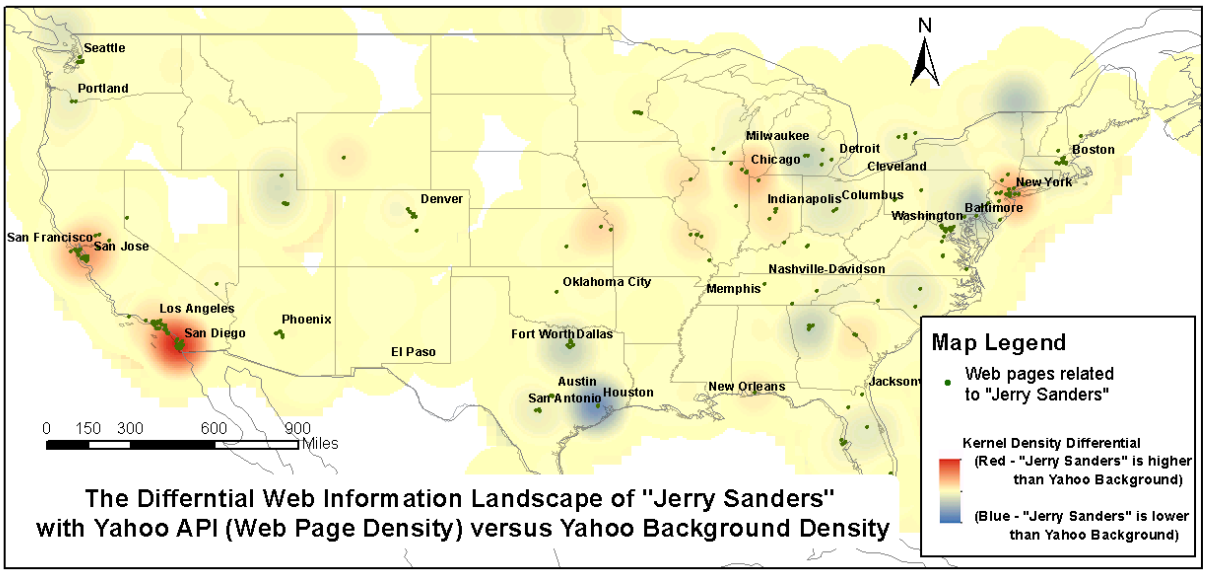
The problems of geolocation of the Internet data are varied (e.g., González et al., 2008; Jones et al., 2008; Song et al., 2010; Takhteyev et al., 2011). The first problem is to know where the poster of the information is actually located. The task of pinpointing the geolocation depends on platform on which it is being done. For instance, if web sites are searched, each web server contains Internet Protocol (IP) address registration information, which can then be retrieved and by converting these into latitude and longitude data, they be plotted on a map (Tsou et al., 2011). A number of commercial outfits provide such services. Unfortunately, simply because an IP address can be located on the world map, it does not necessarily imply that it reflects the true location of the person posting a message. Based on our recent case studies with IP geolocation methods, we can convert 90% web pages into real world locations associated with their web server IPs (Tsou et al., 2011). But only 50% - 60% of web pages are "correctly" mapped with the true locations of the servers. Even with this low percentage of locational accuracy, our maps still show a significant cluster pattern of web page posting in various cities and regions.

To illustrate our points, we used a keyword "Jerry Sanders" (who is the mayor of San Diego) to search Web pages in both Yahoo Search Engine and Bing Search Engine. The Yahoo search engine returned 1000 related web pages related to Jerry Sanders with ranks. The Bing (Microsoft) search engine returned 688 web pages related to Jerry Sanders. We found out that the top 10 results between Yahoo and Bing are similar (five out of ten pages are the same). However, only 40 out of 688 pages (5.8 percent) in Bing search results have identical URLs (Web addresses) comparing to the 1000 Yahoo search results. Most people may not realize that the top 1000 (or 688) web pages from two search engines (Bing and Yahoo) are quite different.

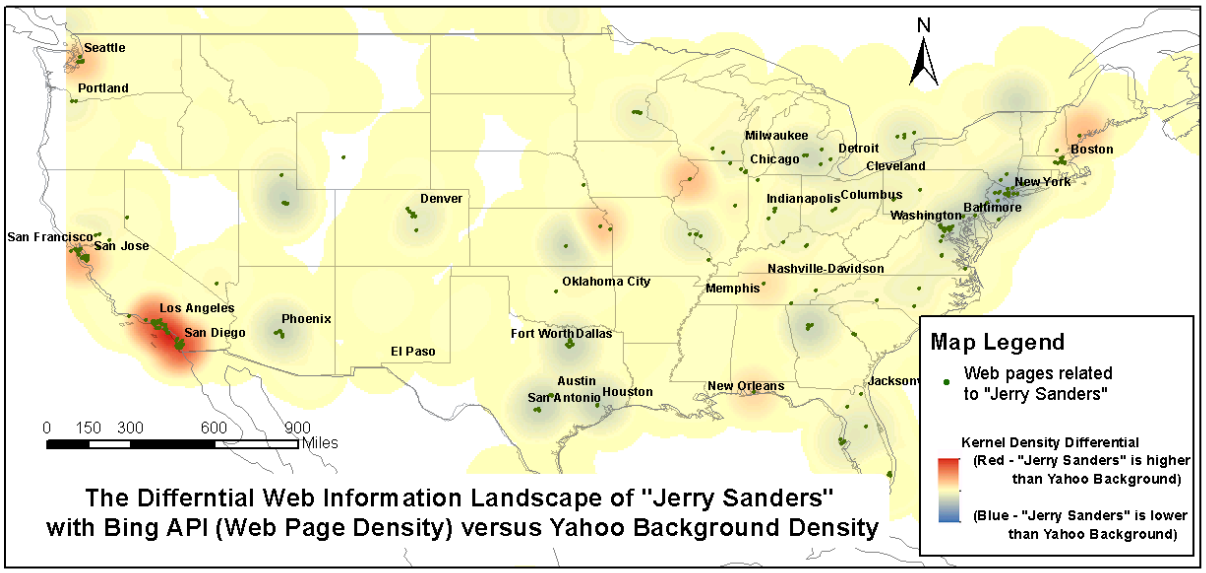
Although the two search engines returned very different web pages using the same keyword, the web page density differential maps we created still display strong spatial patterns for the keywords in both the Yahoo results and Bing results. Figure 2A shows the comparison between the Yahoo background versus Yahoo search results for "Jerry Sanders" (top). Figure 2B shows the Yahoo background versus Bing search results for "Jerry Sanders" (bottom). The regions in red indicate more web page density than average; blue indicates less web pages/density than average.

Figure 2. Comparison of web information differential landscape: Yahoo API results (top), Bing API results (bottom) (standardized by the Yahoo background map).

A).



B).



The web density maps created by Yahoo API (top one) show a hotspot for "Jerry Sanders" in San Diego, California. The density map created by Bing API (bottom one) also shows a hotspot for both Los Angeles and San Diego even though only 5.8 percent of the Bing web page records are identical to the Yahoo API results (Figure 2).

In the course of our analysis, we also observed that different keywords may have markedly different temporal change rates. We tested the search of “Osama Bin Laden” on 04 May 2011, two days after he was killed in Pakistan. Figure 3 illustrates the dramatic change of "Osama Bin Laden" search results on Day 1 (625 URLs out of 1000 are different between 04 May 2011 and 05 May 2011) and Day 2 (793 URLs are different between 04 May 2011 and 06 May 2011) compared to the slower change rate for “burn Koran” on Day 1 (38 URLs out of 1000 are different between 03 April 2011 and 04 April 2011) and Day 2 (67 URLs are different between 03 April 2011 and 05 April 2011).

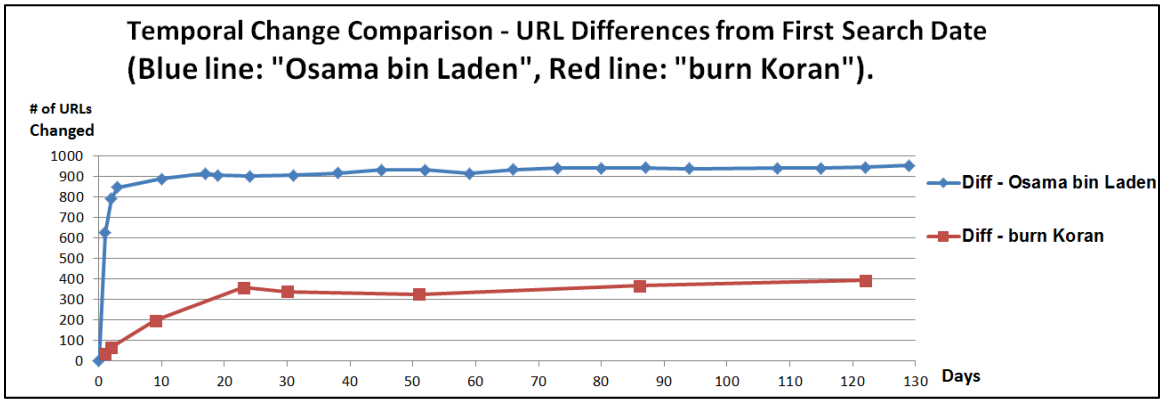


Figure 3. The temporal change comparison of Yahoo API search results for different dates with different keywords (“Osama bin Laden” and “burn Koran”).

The SWARMS framework can be used to query keywords in multiple languages (e.g., Chinese, Arabic, Spanish, or Japanese) and be used in multiple web search engines. In our early tests, we only used English in our keyword search. Different languages may create significantly different web information landscapes. Figure 4 illustrates

the global distribution pattern of the "Osama bin Laden" keyword search in three different languages (English, Chinese (simplified), and Arabic). The global distributions of web pages about "Osama bin Laden" are quite different between the three maps. Further language-specific analysis will be required for understanding the meaning of these spatial pattern language variations.

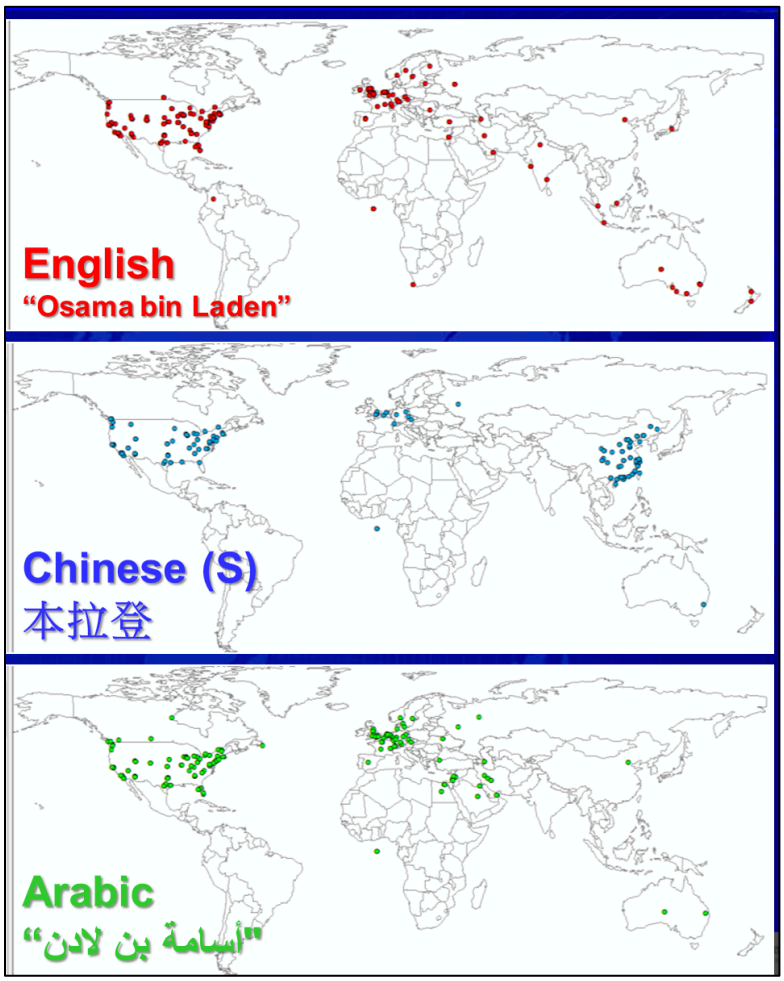


Figure 4. The global distribution patterns of the keyword search "Osama bin Laden" in three different languages (English, Chinese (simplified), and Arabic).

Semantic sensitivity of search: Two Cases

Climate Change vs. Global Warming

In constructing a new methodology for social science research, we must be aware of the semantic sensitivity of our search. A small variation in the spelling or the keyword can provide radically different results. For instance, we can search by using keywords "Koran" or "Quran." We can see the former yields websites that are less religious or even anti-Islamic sites. These sites are located mostly in the Western countries. In contrast, a search by "Quran" provides us with more Islamic sites. Similarly, by adding a single term, we can change the charter of our search. Thus, we can search the name al-Awlaki, a firebrand American born Yemeni cleric, who was killed by US drone attack, or we can add the honorific title to his name and discover that "Sheikh al-Awlaki" returns a completely different sets of sites and postings. Finally, our research demonstrates that while the term "Global Warming" brings in a lot of postings by the skeptics, we find a different crop of people using the term "Climate Change" despite that many people are using them interchangeably. In the content analysis of all relevant web search results (on two dates in June and September 2011) for the keyword "climate change," all websites are classified as *con-climate change* if these websites hold that climate change is not happening or not a serious problem, or (even if it exists or is a serious problem) it arises from natural reasons and humans are not responsible to it. On the other hand, all websites are classified as *pro-climate change* if they consider climate change is happening, posing a serious problem to humans, and it arises primarily from

anthropogenic reasons (e.g., emission of carbon dioxide) and humans are responsible to take actions and mitigate the negative effects associated with climate change. The websites are classified as *neutral* if they do not take obvious slant to either side, take no obvious pro or con position, or use qualified terms and little obviously evaluative or judgmental terms to represent positions, policies, policy-makers, researchers, or theories. The same content-based classification was applied to the search results for "global warming" on similar dates. It was found that there is a much larger fraction of people who are con-global warming than those who are con-climate change, suggesting high semantic sensitivity between the two terms. In the above climate change/global example, our regression analyses have resulted in R-squares ranging from 30% to 65% when the overall interest (internet traffic), the con or pro attitudes, or the net effects (pro attitudes subtracted by con attitudes) on climate change or global warming were examined as dependent variables. A set of biophysical (e.g., drought, precipitation), demographic (e.g., age, ethnicity, education), and socioeconomic (e.g., median household income, commuting time between home and work) independent variables were found to significantly affect interest or attitudes on climate change (An et al. 2012). Such findings confirmed our hypotheses that interest on climate change at a site is influenced by the climate pattern at the site as well as by who they (i.e. the dwellers of the site) are (in terms of age, sex, and ethnicity), where they live (in terms of local climate, population density, and commuting time), and what socioeconomic status they are associated with (in terms of income and education). Such findings would be very useful when

used to customize our limited resources to those most relevant areas and achieve the highest potential effectiveness (An et al. 2012).

White Supremacist Movement

This immense variability in the relation between language and results is both a blessing and a curse. It means on the one hand that when we know the right search terms we have very fine-grained control over the kinds of text collected by our search. On the other, if our assumptions about the connection of the language and the text type are off, the search results may be nothing like a representative sample of the kinds of texts we think we are collecting. For example, the term "ZOG" (used by white as an acronym for "Zionist Occupational Government") turns out to be a very poor term if one is searching for white militant texts: (a) it turns up a lot of anti-militant or news sites discussing the phenomenon or problem of white militancy (for example, the Southern Poverty Law Center or Wikipedia); (b) it's highly ambiguous; for example, it's the name of a San Francisco hot dog store and Zydeco music bands); (c) it seems to be falling out of fashion as a designation for the ruling elite power structure among white militants themselves.

The consequence of this is that when the goals of a study have been defined independently of language we must take great care to assure ourselves that the language we choose for a search matches those goals. Mismatches between term and goal such as inn the "ZOG" example will yield overwhelming numbers of false positives. More subtly, our terms may build in biases we are unaware of, so that the results omit significant components of the group or phenomenon under study. This

happens quite naturally following the splintering or splitting of an idea or group in multiple directions. For example, the sheer increase in the number of radical Islamic organizations recruiting U.S.-born members means a smaller percentage will owe any allegiance to al-Awlaki, entailing that the "Sheikh al-Awlaki" search alluded to above will be less successful in netting a representative sampling of the outputs of such groups. Instead, a host of more minor figures must be discovered and a variety of searches using different names must be used to track this spreading phenomenon. As the idea spreads it changes not its demographics, but also its linguistic expression.

The general solution to this problem is to define our objects of study independently of the language we use for search, to collect seed data, and to discover multiple mutually reinforcing linguistic indicators. This provides a robust solution to the problem of false positives because we can use such indicators to build reliable classifiers of the text types (Gawron et al 2012). It also provides an approach to the problem of false negatives, because while texts of interest may lack certain indicators (such as "Sheikh al-Awlaki"), there will be a whole network of such indicators. While no text of interest will contain all the indicators, the hypothesis is that all of them will contain some, and all the texts can be linked together by family resemblances among the indicators they do contain. The group identification work described in the next section provides an example of such modeling, using "group membership markers" as the linguistic indicators.

Going Beyond Sentiment analysis

In this new methodological revolution, sentiment analyses become crucial in understanding the evolving narrative (Fortunati, 2009; Li & Wu, 2010). They can provide invaluable information regarding variety of topics from product placement (e.g., "AT&T network is slower and has more dropped calls") to the outcome of an election, where each contestant wants to pin a narrative around an opponent (e.g., "Romney is an out-of-touch billionaire" or "Obama is a socialist"). Yet, when it comes to collective actions in terms of a political movement, we may dig a bit deeper into human motivation.

The "rational actor" hypothesis, arguably the most widely accepted assumption in the social sciences, tries to explain patterns of human behavior as the natural result of individuals acting in their own individual interest. The idea is an organizing principle in disciplines as diverse as economics, artificial intelligence, psychology, and linguistics. Yet the growing field of social psychology starting, *inter alia*, with the seminal work of Tajfel (1978, 1981; see also: Brown, 2000; Hogg et al., 1995; Korte, 2006; Turner & Reynolds, 2001) is busily accumulating evidence of the importance of groups and group in our decision-making process (Watts & Dodds, 2007). These research efforts clearly demonstrate that our decisions are heavily influenced by the group(s) in which we claim membership (McGlone & Giles, 2011; McFarland & Pals, 2005; Mullen et al., 2001; Reid & Giles, 2005). Group and collective identities can supersede individual identity; people often embark upon courses of action detrimental to their personal economic well-being, liberty, and life

itself. This possibility is strongest in groups in which the sense of self-identification is strongest. The key implication is that *individual and group identities are manifest in symbolic expression in electronic texts, and such texts are systematically searchable and reliably identifiable en masse* (Gawron et al., 2012).

A reasonable starting assumption is that intense group identification requires a clear articulation not only of who "we" are, but also who "they" are— the outsiders, the other, the "out group", often, the enemies — an articulation that is central to all large-scale collective action from nationalism (Anderson, 2003) to terrorism (Gupta, 2008). In some cases a group is defined by a pre-existing language, but in most cases it is not; whether it is or not, an essential part of the process of dividing us from them is developing a group sublanguage. This may have a complex array of linguistic components, ranging from phonological to syntactic features, but an essential part of it is evaluative language referring to us and to them, as well as language referring to properties of us and properties of them (Little et al., 2003).

For well-established groups with a longer history the language includes a complex set of references to heroes, leaders, victims, and artists, as well as to subgroups, key events, key dates, and key writings and key works of art, including music and games. Although group formation requires identification of "us" and "them," the mobilization of a large number of people for collective action requires a third factor: a clear articulation of an impending existential threat (Gupta 2008). This is not an unintuitive result. Behavioral research by likes of Kahenman, Slovic and Tversky (1982) demonstrates the dominance of prospective losses over gains in

the evaluation of uncertain futures. In brief, the prospect of loss of what is currently possessed is a far more potent motivator than the prospect of gains (Kahneman & Tversky, 1979; Novemsky & Kahneman, 2005). As a result, from political extremism to electoral politics, fear tactics are a winning strategy. In accord with this idea, a well-conceived "us-versus-them" analysis will target language articulating threats as well as language referring to the enemy.

For example, the in-group for white militant and hate groups is members of the white race, whereas the out-group or enemies are the non-white population, including Jews and Catholics for some groups, but significantly, also a group of white people who are perceived as traitors to the race. The general existential threat is the degradation and pollution of pure white stock, but there are many more specific instantiations because degradation has many potential dimensions and strategies. A semantic ontology can be developed, relying on domain expertise, incorporating elements of the militant group argot and linguistic iconography referring to us-versus-them (Chau & Xu, 2007). The ontology is further expanded to include properties and products of us-versus-them and to existential threats to "us." The hypothesis is that the elements of the us-versus-them language are strong markers of group identity. Moreover, the us-versus-them language is largely learned, with more experienced speakers using it more fluently and more frequently. Communicators who control a significant subset of this language are likely to be well-established in the group. Identifying a significant set of such markers in a text provides strong *prima facie* evidence of core group membership, such as high degrees of militan cy. Subsequent research that directly inspects the websites

extracted from the search processes can help to validate the accuracy of such attributions.

Another example of group identification and its social implications is entailed in the analysis of social movements. From a diffusion of innovations perspective (Compeau et al., 2007; Elkink, 2011; Meade & Islam, 2006; Rogers, 1983; Rogers & Kincaid, 1981; Vishwanath & Chen, 2011; Young, 2009), ideas are a form of diffusion innovation, and as such, can be understood in terms of their adoption curves within populations (Earl, 2010; Marquette, 1981). Social movements represent collective efforts to diffuse a particular vector of ideas, beliefs and values. Given the increasing use and reliance upon various electronic media for the mass diffusion of ideas (e.g., Carty, 2010; Diani, 2000; Earl, 2010; Stein, 2009; Strodthoff, 1985; Van Laer, 2010), the diffusion of social movements may be potentially mapped in almost real time, if the semantic contents (i.e., ontology) of such ideas can be discriminated from the background of other ideas. The key elements of this process of ontology development, search processes, and subsequent iterations of interpretation and refinement.

Preliminary investigations of variants of the phase “Arab Spring” in English and Arabic have demonstrated an ability to reveal sensitive changes in the diffusion of democratic concepts throughout various geographic regions in the Middle East (Spitzberg et al., 2012; see also Etling, et al., 2010; Howard & Hussein, 2011). Eventually, correlation of such patterns over time and space, carefully intersected with the communication strategies employed by the protesters (Van Laer & Van Aelst, 2010) and the thresholds of communication activity achieved, may well

provide reasonably predictive models that can differentiate stalled from successful revolutions, policy or authority shifts.

Political suppression and Issues of Civil Liberty

As this new methodology in social science research develops, it opens up new possibilities along with possible danger of its misuse. For obvious reasons, law enforcement authorities from all over the world have taken notice of the potential power of data-mining. There are risks that the most powerful processes of data mining may "become the exclusive domain of private companies and government agencies" (Lazer et al., 2009, p. 721). The power with which such media can be increasingly used to geo-locate individual and group activity with spatial referents (Erickson, 2010; Tillema et al., 2010) portend dark possibilities of abuse.

Apart from the Arab Spring, which is changing the political landscape of the entire Arab/Muslim world and the "Occupy Wall Street" movement, changing the course of national discussion about poverty and income inequality, social media and twitter have also helped spawn "flash mobs" with quick and instantly spreading violence among several of the world's capitals (Massaro & Mullaney, 2011). This has sounded alarm bells, particularly in the authoritarian nations. The Egyptian authorities realized the power of the social media a bit too late. By the time they attempted to pull the plug on the Internet traffic, the damage had already been done. Therefore, the surveillance of the Internet traffic and their regulation have become an essential part of the authoritarian nations' law enforcement apparatus, where they not only try to control the flow of the information, but also aim at suppressing legitimate political dissent.

The Infancy of Research and the Need

The world is wired, wireless, and increasingly linked (Barabási, 2002; Dodds et al., 2003; Watts, 2003) and therefore, increasingly interdependent in complex ways (Barabási, 2010). People's behaviors are already revealing far-reaching insights into their everyday patterns of spatial (González et al., 2008; Jones et al., 2008; Song et al., 2010) and communicative (Lewis et al., 2011; Suh et al., 2010; Walther & Bazarova, 2008; Watts et al., 2002) organization. Theorists are increasingly understanding that *cyberspace* is beginning to map into, onto, and through *realspace* (see, e.g., Adams, 2010; Breese, 2011), fulfilling Hägerstrand's (1965; Gale, 1978) envisioning of an interdisciplinary theory of time-space geography for the cartography of human behavior.

In order for us to collect and analyze the vast amount of data, we need to go far beyond what social disciplines have traditionally gone. These efforts must not only be imaginative, sophisticated, but also truly multi-disciplinary. The early post-war development in social science methodology saw the destruction of he academic tower of Babel, where each branch became happy developing its own language, assumptions, and methodologies. Today's research challenges require joint efforts from not only from all branches of social sciences, it requires technical knowledge from such diverse disciplines as computer science, network engineering, mathematics, and medicine, to name a few.

Social science research based data mining is still at its infancy. As it develops its possibilities are truly enormous and much of it are yet to be realized. On the one

hand, the spread of Internet technology is empowering common people from all over the world it is also being viewed with extreme suspicion by others. There is no question about the fact that the new technology is ushering in a new era of social science research; in order for us to fully harness its power from disaster preparedness and mitigation of threats of pandemics, as social scientists we must be cognizant of its abilities for good and evil.

References

Adams, Paul C. (2010) A taxonomy for communication geography. *Progress in Human Geography, 35*(1): 37-57.

Alexander, R. J. (2009). *Framing discourse on the environment: A critical discourse approach.* New York, NY: Routledge.

Altaweel, M. R., Alessa, L. N., & Kliskey, A. D. (2010). Visualizing situational data: Applying information fusion for detecting social-ecological events. *Social Science Computer Review*, *28*(4), 497-514. doi:10.1177/0894439309360837

An, L., and D. G. Brown. 2008. Survival analysis in land-change science: integrating with GIScience to address temporal complexities. Annals of Association of American Geographers 98(2): 323-344.

An, L., D. G. Brown, J. I. Nassauer, and B. Low. 2011. Variations in development of exurban residential landscapes: timing, location, and driving forces, Journal of Land Use Science 6(1): 13-32.

An, L., M. Tsou, S. Wandersee, D.K. Gupta, B. Spitzberg, and J.M. Gawron. 2012. Is climate change a myth? Evidence from cyberspace and realspace. Presented at the Symposium "Web Surveillance: Fighting Terrorism and Infectious Diseases", annual meeting of American Association for the Advancement of Science (AAAS), February 17-21, 2012, Vancouver, Canada.

An, L., Tsou, M-T., Wandersee, S., Gupta, D. K., Spitzberg, B. H., & Gawron, M. (2012, February). *Who Is concerned about climate change? Evidence from space-time analysis.* Paper submitted to the American Association of Geographers Conference, New York, NY.

Andrés, Luis, Cuberes, David, Diouf, Mame, & Serebrisky, Tomás. (2010). The diffusion of the internet: A cross-country analysis. *Telecommunications Policy, 34*, 323-340.

Arpinar, I. Budak, Sheth, Amit, & Ramakrishnan, Cartic, Usery, Lynn, Azami, Molly, & Kwan, Mei-Po. (2006). Geospatial ontology development and semantic analytics. *Transactions in GIS, 10*(4), 551-575.

Attrill, A., & Jalil, R. (2011). Revealing only the superficial me: Exploring categorical self-disclosure online. *Computers In Human Behavior*, *27*(5), 1634-1642. doi:10.1016/j.chb.2011.02.001

Barabási, Albert-László. (2002). *Linked: The new science of networks*. Cambridge, MA: Perseus.

Barabási, Albert-László. (2010). *Bursts: The hidden pattern behind everything we do.* New York, NY: Dutton.

Barrett, C., Bisset, K., Leidig, J., Marathe, A., & Marathe, M. (2010). An Integrated Modeling Environment to Study the Coevolution of Networks, Individual Behavior, and Epidemics. *AI Magazine*, *31*(1), 75-87.

Bell, Gordon, Hey, Tony, & Szalay, Alex (2009). Beyond the data deluge. *Science*, *323*, 1297-1298.

Bollen, Johan, Mao, Huina, & Zeng, Xiaojun. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*, 1-8.

Bostrom, R. N. (2003). Theories, data, and communication research. *Communication Monographs, 70*, 275-294.

Brasch, W. M. (2005). Fool's Gold in the Nation's Data-Mining Programs. *Social Science Computer Review*, *23*(4), 401-428. doi:10.1177/0894439305278869

Breese, Elizabeth Butler. (2011). Mapping the variety of public spheres. *Communication Theory, 21*, 130-149.

Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity: International Journal Of Methodology*, *45*(1), 103-128. doi:10.1007/s11135-010-9350-8

Brown, Rupert. (2000). Social identity theory: Past achievements, current problems and future challenges. *European Journal of Social Psychology, 30*, 745-778.

Cannataro, M., & Talia,D. (2003). The knowledge grid. *Communications of the ACM, 46*(1), 89-93.

Cao, L. (2010). In-depth behavior understanding and use: The behavior informatics approach. *Information Sciences*, *180*(17), 3067-3085. doi:10.1016/j.ins.2010.03.025

Carrasco, Juan Antonio, Hogan, Bernie, Wellman, Barry, & Miller, Eric J. (2008). Agency in social actdivity interactions: The role of social networks in time and space. *Tijdschrift voor Economische en Sociale Geografie, 99* (5), 562-583.

Carty, V. (2010). New information communication technologies and grassroots mobilization. *Information, Communication and Society,* 13(2): 155-173.

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies,* 65(1), 57-70. doi:10.1016/j.ijhcs.2006.08.009

Chow, S.L. (1992). Acceptance of a theory: Justification or rhetoric? *Journal for the Theory of Social Behaviour, 22*, 447-474.

Copmpeau, Deborah R., Meisher, Darren B., & Higgins, Christopher A. (2007). From prediction to explination: Reconceptualizing and extending the perceived characteristics of innovating. *Journal of the Association for Information Systems, 8*(8), 409-439.

Corman, Stephen R., Kuhn, T, McPhee, Robert D., & Dooley, KJ (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research,* 28(2), 157-206.

Crandall, David J., Backstrom, Lars, Cosley, Dan, Suri, Siddharth, Huttenlocher, Daniel & Kleinberg, Jon. (2010) Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences, 107*(52): 22436-22441.

Diani, M. (2000). Social movement networks: Virtual and real. *Information, Communication and Society,* 3(3): 386-401.

Dodds, Peter Sheridan, Muhamad, Roby & Watts, Duncan J. (2003) An experimental study of search in global social networks. *Science, 301*(5634), 827-829.

Earl, J. (2010). The dynamics of protest-related diffusion on the web. *Information, Communication and Society,* 13(2): 209-225.

Elkink, Johan A. (2011). The international diffusion of democracy. *Comparative Political Studies, 44*(12), 1651-1674.

Elson, Sara Beth, Yeung, Douglas, Roshan, Parisa, Bohandy, S. R., & Nader, Alireza. (2012). *Using social media to gauge Iranian public opinion and mood after the 2009 election.* Santa Monica, CA: Rand/National Security Research Division.

Erickson, Ingrid. (2010) Geography and community: New forms of interaction among people and places. *American Behavioral Scientist, 53*(8), 1194-1207.

Etling, B., Kelly, .J, Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere: Politics and dissent online. *New Media and Society,* 12(8), 1225-1243.

Fortunati, Leopold. (2009). Old and new media, old emotion. In L. Fortunati & J. Vincent (Eds.), *Electronic emotion: The mediationof emotion via information and communication technologies* (pp. 35-62). Bern, Switzerland: Peter Lang.

Gale, Stephen. (1972) Some formal properties of Hägerstrand's model of spatial interactions. *Journal of Regional Science, 12*(2), 199-217.

Gawron, J. M., Gupta, D., Stephens, K., Tsou, M-H., Spitzberg, B. H., & Li, A. (2012, June). *Using group membership markers for group identification in web texts*. Paper presented at the Sixth International AAAI Conference on Weblogs and Social Media Conference, Dublin, Ireland.

Gergel, S.E, M. G. Turner (Editors). 2001. Learning Landscape Ecology. Springer: New York.

Getis, A., and D. A. Griffith. 2002. Comparative Spatial Filtering in Regression Analysis. Geographical Analysis 34 (2):130-140.

Gilman, D. (1992). What's a theory to do...with seeing? or some empirical considerations for observation and theory. *British Journal for the Philosophy of Science, 43,* 287-309.

González, Marta C., Hidalgo, César A., & Barabási, Albert-Lászlo. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779-782.

Gupta, Dipak K. (1994) Decisions By The Numbers. Englewood Cliffs, N.J.: Prentice Hall.

Gupta, Dipak K. (2001) *Path to Collective Madness: A Study In Social Order and Political Pathology*. Westport, CT: Praeger.

Gupta, Dipak K. (2008) *Understanding Terrorism and Political Violence: Lifecycle of Birth, Growth, Transformation, and Demise*. London: Routledge.

Hägerstrand, Torsten. (1966) Aspects of the spatial structure of social communication and the diffusion of information. *Papers in Regional Science, 16*(1), 27-42.

Hägerstrand, Torsten. (1970) What about people in regional science? *Papers in Regional Science, 24*(1), 7-24.

Hand, D. J. (2000). Data Mining: New Challenges for Statisticians. *Social Science Computer Review*, *18*(4), 442.

Ho, Y., Chung, Y., & Lau, K. (2010). Unfolding large-scale marketing data. *International Journal Of Research In Marketing*, doi:10.1016/j.ijresmar.2009.12.009

Hogg, Michael A., Terry, Deborah J., & White, Katherine M. (1995). A tale of two theories: A critical comparison of identity theory with social identity theory. *Social Psychology Quarterly, 58*, 255-269.

Howard, Philip N., & Hussain, Muzammil M. (2011) The upheavals in Egypt and Tunisia: The role of digital media. *Journal of Democracy, 22*(3): 35-48.

Huang, Chun-Yao, & Chen, Hau-Ning. (2010). Global digital divide: A dynamic analysis based on the Bass model. *Journal of Public Policy & Marketing, 29*(2), 248-264.

Hwang, Hyunseo, Schmierbach, M., Paek, Hye-Jin, Zuniga, de Homero Gil, & Shah, Dhavan. (2006). Media dissociation, internet use, and antiwar political participation: A case study of political dissent and action against the war in Iraq. *Mass Communication & Society, 9*(4), 461-483.

Johnson, B. D., Dunlap, E., & Benoit, E. (2010). Organizing "mountains of words" for data analysis, both qualitative and quantitative. *Substance Use & Misuse*, *45*(5), 648-670. doi:10.3109/10826081003594757

Jones, Quenton, Grandhi, Sukeshini A., Karam, Samer, Whittaker, Steve, Zhou, Changqing & Terveen, Loren. (2007) Geographic 'place' and 'community information' preferences. *Computer Supported Cooperative Work, 17*(2-3), 137-167.

Korte, Russell F. (2006). A review of social identity theory with implications for training and development. *Journal of European Industrial Training, 31*, 166-130.

Ladhani, N. (2011, Winter). Occupy social media. *Social Policy,* 83. Available: http://www.socialpolicy.org/index.php/component/content/article/4-latest-issue/503-occupy-social-media

Lazer, David, Pentland, Alex, Adamic, Lada, Aral, Sinan, Barabási, Albert-László, et al. (2009). Computational social science. *Science, 323*, 721-723.

Leetaru, Kalev Hannes. (2012). *Data mining methods for the content analyst: An introduction to the computational analysis of content*. New York, NY: Routledge.

Lewis, Kevin, Gonzalez, Marco, & Kaufman, Jason. (2011). Social selection and peer influence in an online social network. *PNAS Early Edition*. Available: www.pnas.org/cgi/doi/10.1073/pnas.1109739109.

Lewis, Kevin, Kaufman, Jason,& Christakis, Nicholas. (2008). The taste for privacy: An analysis of college student privacy settings in an online network. *Journal of Computer-Mediated Communication, 14*, 79-100.

Li, N and Wu, DD (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems,* 48(2), 354-368.

Little, Miles, Jordens, Christopher F. C., & Sayers, Emma-Jane. (2003). Discourse communities and the discourse of experience. *Health: An interdisciplinary Journal for the Social Study of Health, Illness and Medicine, 7*(1), 73-86.

Marquette, J.F. (1981). A logistic diffusion model of political mobilization. *Political Behavior, 3*, 7-30.

Massaro, Vanessa A., & Mullaney, Emma Gaalaas. (2011). The war on teenage terrorists. *City, 15* (5), 591-604.

McFarland, Daniel, & Pals, Heili. (2005). Motives and contexts of identity change: A case for network effects. *Social Psychology Quarterly, 68*, 289-315.

McKenna, K. Y. A., & Bargh, J. A. (2000). Plan 9 from cyberspace: The implications of the internet for personality and social psychology. *Personality and Social Psychology Review, 4*(1), 47-75.

Meade, N., & Islam, T. (2006) Modeling and forecasting the diffusion of innovation – A 25-year review. *International Journal of Forecasting, 22*, 519-525. doi:10.1016/j.ijforecast.2006.01.005

Mullen, B., Migdal, M. J., & Hewstone, M. (2001). Crossed categorization versus simple categorization and intergroup evaluations: A meta-analysis. *European Journal of Social Psychology, 31*, 721-736.

Murphy, Emma C. (2006) Agency and space: The political impact of information technologies in the Gulf Arab states. *Third World Quarterly, 27*, 1059-1083. DOI: 10.1080/01436590600850376

Nebot, V., & Berlanga, R. (2012). Finding association rules in semantic web data. *Knowledge-Based Systems*, *25*(1), 51-62. doi:10.1016/j.knosys.2011.05.009

Novemsky, Nathan, & Kahneman, Daniel. (2005). The boundaries of loss aversion. *Journal of Marketing Research, 42*(2), 119-128.

Olson, Mancur (1964) *The Logic of Collective Action*. Cambridge, Mass.: Harvard University Press.

Papacharissi, Z. (2009). The virtual geographies of social networks: A comparative analysis of Facebook, LinedIn and ASmallWorld. *New Media and Society, 11*, 199-220.

Pavitt, C. (2004). Theory-data interaction from the standpoint of scientific realism: A reaction to Bostrom. *Communication Monographs, 71*, 333-342.

Petronio, S. (2002). *Boundaries of privacy: Dialectics of disclosure.* Albany, NY: State University of New York Press.

Popper, K. (1980). Science: Conjectures and refutations. In E. D. Klemke, R. Hollinger, & A. D. Kline (Eds.), *Introductory readings in the philosophy of science* (pp. 19-34). Buffalo, NY: Prometheus.

Postmes, Tom, & Brunsting, S. (2002). Collective action in the age of the internet: Mass communication and online mobilization. *Social Science Computer Review, 20* (3)290-301.

Ramakrishnan, N., & Grama, A. Y. (1999, August). Data mining: From serendipity to science. *Computer,* , 34-37.

Reid, Scott A., & Giles, Howard. (2005). Intergroup relations: Its linguistic and communicative parameters. *Group Processes & Intergroup Relations, 8* (3), 211-214. DOI: 10.1177/1368430205053938

Rogers, Everett. M. (2003) *Diffusion of innovations* (5th ed.). New York: Free Press.

Rogers, Everett. M., & Kincaid, D. L. (1981) *Communication networks: Toward a new paradigm for research*. New York: Free Press.

Rohlinger, D. A., & Brown, J. (2009). Democracy, action, and the Internet after 9/11. *American Behavioral Scientist, 53(1)*, 133-150.

Saldarini, Robert A., & DeRobertis, Eugene M. (2003). The impact of technology induced anonymity on communications and ethics: New challenges for IT pedagogy. *Journal of Information Technology Impact, 3*(1), 3-10.

Shaw, Shih-Lung, Yu, Hongbo, & Bombom, Leonard S. (2008). A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS, 12*(4), 425-441.

Song, C, Qu, Z, Blumm, N and Barabási, A-L (2010) Limits of predictability in human mobility. *Science, 327*, 1018-1021. DOI: 10.1126/science.1177170

Spitzberg, B. H., Tsou, M-H., An, L., Gupta, D. K., & Gawron, J. M. (2012, May*). The map is not which territory?: Speculating on the geo-spatial diffusion of ideas in the Arab Spring of 2011*. Paper presented at the International Communication Association Conference, Phoenix, AZ.

Stein, L. (2009). Social movement web use in theory and practice: A content analysis of US movement websites. *New Media and Society, 11*(5): 749-771. DOI: 10.1177/1461444809105350

Strodthoff, G. G., Hawkins, R. P., & Schoenfeld, A. C. (1985). Media roles in a social movement: A model of ideology diffusion. *Journal of Communication, 35*, 134-153.

Suh, Bongwon, Hong, Lichan, Pirolli, Peter, & Chi, Ed. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. Second International IEEE International Conference on Social Computing, 177-184.

Takhteyev, Yuri, Gruzd, Anatoliy, & Wellman, Barry. (2011). Geography of Twitter networks. DOI: 10.1016/j.socnet.2011.05.006

Taleb, N. N., & Blyth, M. (2011). The black swan of Cairo: How suppressing volatility makes the world less predictable and more dangerous. *Foreign Affairs, 90*(3), 33-39.

Tillema, Taede, Dijst, M., & Schwanen, Tim. (2010). Face-to-face and electronic communications in maintaining social networks: The influence of geographical and relational distance and of information content. *New Media & Society, 12*(6), 965-983.

Tsou, M-H., An, L., Wandersee, S., Kim, I-H., Spitzberg, B. H., Gupta, D., Gawron, J. M., Smith, J., Lee, T-H. (2011). Mapping ideas from cyberspace to realspace: Visualizing hidden geospatial fingerprints on web information landscapes. *Annals of the Association of American Geographers***.**

Turner, John C., & Reynolds, Katherin J. (2001). The social identity perspective in intergroup relations: Theories, themes, and controversies. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp.133-152). Oxford, England: Blackwell.

Van Laer, J. (2010). Activists online and offline: The internet as an information channel for protest demonstrations. *Mobilization: An International Journal, 15*, 347-366.

Van Laer, J., & Van Aelst, P. (2010). Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication and Society, 13*, 1146-1171. DOI: 10.1080/1369118100368307

Vishwanath, Arun, & Chen, Hao. (2011). Towards a comprehensive understanding of the innovation-decision process. *The diffusion of innovations: A communication science perspective* (pp. 9-32). New York, NY: Peter Lang.

Walther, Joseph B. & Bazarova, Natalya N. (2008) Validation and application of electronic propinquity theory to computer-mediated communication in groups. *Communication Research, 35*, 622-645.

Watts, Duncan J. (2003). *Six degrees: The science of a connected age.* New York, NY: W. W. Norton.

Watts, Duncan J. (2004) The "new" science of networks. *Annual Review of Sociology, 30*, 243-270. DOI: 10.1146/annurev.soc.30.020404.104342

Watts, Duncan J., & Dodds, Peter S. (2007) Influentials, networks, and public opinion formation. *Journal of Consumer Research, 34*, 441-458.

Watts, Duncan J., Dodds, Peter S., & Newman, M. E. J. (2002). Identity and search in social networks. *Science, 296*, 1302-1305.

Yang, Qiang, & Wu, Xindong (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making, 5* 4), 597–604.

Young, H. Peyton. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review, 99*(5), 1899-1924.

Zimbra, D, Abbasi, A and Chen, H (2010) A cyber-archaeology approach to social movement research: Framework and case study. *Journal of Computer-Mediated Communication, 16*, 48-70. DOI: 10.1111/j.1083-6101.2010.01531.x