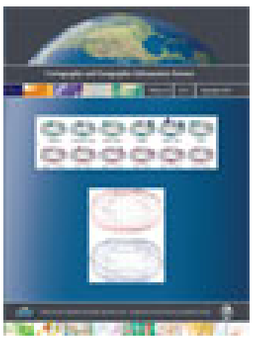


Taylor & Francis
Taylor & Francis Group

Cartography and Geographic Information Science

ISSN: 1523-0406 (Print) 1545-0465 (Online) Journal homepage: http://www.tandfonline.com/loi/tcag20

Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election

Ming-Hsiang Tsou , Jiue-An Yang , Daniel Lusher , Su Han , Brian Spitzberg , Jean Mark Gawron , Dipak Gupta & Li An

To cite this article: Ming-Hsiang Tsou , Jiue-An Yang , Daniel Lusher , Su Han , Brian Spitzberg , Jean Mark Gawron , Dipak Gupta & Li An (2013) Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election, Cartography and Geographic Information Science, 40:4, 337-348, DOI: 10.1080/15230406.2013.799738

To link to this article: http://dx.doi.org/10.1080/15230406.2013.799738

Published online: 30 May 2013.

Submit your article to this journal

Article views: 577

View related articles

Citing articles: 3 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tcag20

Download by: [SDSU San Diego State University] **Date:** 06 November 2015, At: 17:15

Cartography and Geographic Information Science, 2013
Vol. 40, No. 4, 337–348, http://dx.doi.org/10.1080/15230406.2013.799738

Taylor & Francis
Taylor & Francis Group

Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election

Ming-Hsiang Tsou[a]*, Jiue-An Yang[a,b], Daniel Lusher[a], Su Han[a,b], Brian Spitzberg[c], Jean Mark Gawron[d], Dipak Gupta[e] and Li An[a]

[a]*Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4493, USA;* [b]*Joint Doctoral program with University of California, Santa Barbara, CA, USA;* [c]*School of Communication, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4493, USA;* [d]*Department of Linguistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4493, USA;* [e]*Department of Political Science, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4493, USA*

(*Received 30 November 2012; accepted 5 March 2013*)

We introduce a new research framework for analyzing the spatial distribution of web pages and social media (Twitter) messages with related contents, called Visualizing Information Space in Ontological Networks (VISION). This innovative method can facilitate the tracking of ideas and social events disseminated in cyberspace from a spatial-temporal perspective. Thousands of web pages and millions of tweets associated with the same keywords were converted into visualization maps using commercial web search engines (Yahoo application programming interface (API) and Bing API), a social media search engine (Twitter APIs), Internet Protocol (IP) geolocation methods, and Geographic Information Systems (GIS) functions (e.g., kernel density and raster-based map algebra methods). We found that comparing multiple web information landscapes with different keywords or different dates can reveal important spatial patterns and "geospatial fingerprints" for selected keywords. We used the 2012 US Presidential Election candidates as our case study to validate this method. We noticed that the weekly changes of the geographic probability of hosting "Barack Obama" or "Mitt Romney" web pages are highly related to certain major campaign events. Both attention levels and the content of the tweets were deeply impacted by Hurricane Sandy. This new approach may provide a new research direction for studying human thought, human behaviors, and social activities quantitatively.

Keywords: web information landscapes; geospatial fingerprints; social media; Twitter; election

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

Introduction

The spread of ideas in the age of the Internet is a double-edged sword; it can enhance our collective welfare as well as produce forces that can destabilize the world. Traditional approaches to understanding the spread of impacts of ideas or events are based on twentieth century media – such as newsletters, advertisements, physically proximal group meetings, and telephone conversations. Cyberspace (Gibson 1984) (including web pages, social media, and online communities) is a powerful platform for collective social communications, personal networking, and idea exchange. Scientists now can trace, monitor, and analyze the spreads of radical social movements, protests, political campaigns, etc., via social media and weblogs. These research efforts can help us understand the diffusion of innovations (Roger 1962; Hägerstrand 1967; Brown 1981), a dynamic process whereby new concepts, ideas, and technologies spread through our society via cyberspace and digital social networks over time. An *innovation* is "an idea, practice, or object that is perceived as new by an individual or other unit of adoption" (Rogers 2003, 12). When an individual generates a new message, and that message is received and re-sent by others, it reflects a process of communicative innovation adoption. When spread across the potential population of all who might adopt a given message or idea, the diffusion rate, adoption curve shape, and market saturation all reflect aspects of the influence of that particular idea. In this sense, every message that is sent in cyberspace is a potential trace or reflection of an idea (i.e., potential influence), and every re-sent message is a trace of actual influence. The more interconnected certain social networks are, and the more central and durable certain ideas are in their recirculation within those social networks, they can illustrate potential signifiers of social and societal influence. This is not to ignore some of the critiques of traditional diffusion, such as in Blaut (1987). Using users within social networks as the innovation centers, the diffusion assumption of constant centers of innovation disappears, as any number of people can innovate, let the idea spread, and cause another individual to innovate without physical geographic impediments. Using social networks also breaks down the cited exchange of diffusion (trading civilization/modernization for

*Corresponding author. Email: mtsou@mail.sdsu.edu

© 2013 Cartography and Geographic Information Society

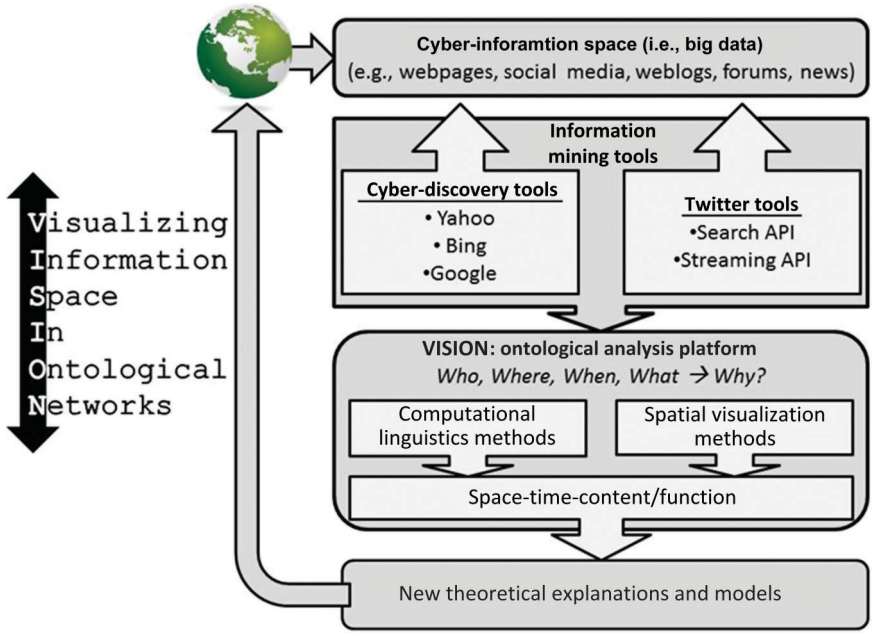


Figure 1. The Visualizing Information Space In Ontological Networks (VISION) framework.

raw materials), as online idea sharing is often close to free (see Blaut 1987 for the cited exchange).

To date, most empirical work on mapping cyberspace has viewed it as only loosely tethered to geospatial coordinates. Summary structural counts of messages or topics, and network linkages of message densities reflect the structure of cyberspace, but say relatively little about the realspaces (referring to the physical world that contains face-to-face communication and idea dispersion) from which such messages originate or terminate. Yet, real people in realspaces are sending and re-sending these messages, and it has long been known that propinquity and proximity significantly influence communication exchanges (see Rainie and Wellman 2012; Yin, Shaw, and Yu 2011). Investigating the correspondence between cyberspace and realspace is not only becoming increasingly possible given current technologies, but the discovery of such correspondences holds substantial promise for understanding the diffusion of ideas through time and space, both real and digital (see Adams 2010a, 2010b). Some, such as Lerman and Ghosh (2010) have been using social networks such as Digg and Twitter in mapping cyberspace related to news stories. Others, such as Paul and Dredze (2011), have been harnessing Twitter in relation to public health; isolating geographic regions related to cyberspace messages. There are also Vieweg et al. (2010) who examined Twitter in relation to natural hazard events.

This article introduces an innovative research framework, called Visualizing Information Space in Ontological Networks (VISION) (http://mappingideas.sdsu.edu). VISION is designed to track spatial patterns of publicly accessible web pages and semi-private social media based upon searching predefined clusters of keywords determined by domain experts (Figure 1). The digital "footprints" of human beings (including social media, web pages, weblogs, and online forums) were traced by our two "information mining" tool sets. The *Cyber-Discovery Tools* (Java-based) were created to collect web pages with associated keywords using Yahoo, Google, and Bing commercial search engine APIs (application programming interfaces). The *Geo-Search-enabled Twitter Tools* (Python-based) are designed to collect tweets associated with different cities, regions, and keywords by using Twitter Search APIs or Streaming APIs. The collected digital footprints were converted into visualization maps and graphs using Geographic Information Systems (GIS) analysis functions and geolocation methods. These visualization maps represent cyberspace information landscapes constructed by a collection set of human ideas and messages. The ontological analysis focuses on the dynamic relationships among space, time, and actual message contents, which may facilitate the creation of new communication models and social science theories in the future (Figure 1).

Following the concepts of diffusion innovation introduced by Rogers (2003), Hägerstrand (1966, 1967), and related works (e.g., Andrés et al. 2010; Elkink 2011; Postmes and Brunsting 2002), this multidisciplinary framework (VISION) demonstrated a new methodology for visualizing and analyzing web pages and social media contents from a spatiotemporal perspective. Our research extends the scope of spatial analysis from physical world phenomena to cyberspace contents. Applications of web

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

information landscapes can be extended to multiple fields including marketing, homeland security public health, and business planning.

In this article, we used the 2012 US Presidential Election as our case study to validate this VISION framework. Thousands of web pages and millions of tweets were geocoded with real world coordinates and represented as cyberspace information landscapes. Three types of comparison methods were demonstrated in this article for the analysis of cyberspace information landscapes:

(1) The weekly dynamic change of web page information landscapes associated with the comparison between "Mitt Romney" and "Barack Obama".
(2) The daily and weekly change of social media (tweets) attention levels associated with "Mitt Romney" vs. "Barack Obama" during the election campaign.
(3) The weekly comparison of word cloud changes (sentiment analysis from tweets) associated with "Mitt Romney" vs. "Barack Obama" before and after the Hurricane Sandy.

Overall, this research demonstrates the validity of our new theoretical framework, VISION, while discussing new insights concerning cyberspace regarding the US Presidential Election. We show the value of investigating the spatial location of relevant web servers and geocoded tweets while establishing a method for using these sources to uncover new ways of relating cyberspace to realspace. From the case study of the election, we also show the levels of web server activity for particular candidates and the Twitter activity related to particular candidates. These levels of activity are then contrasted when a large-scale news event occurs (Hurricane Sandy). The primary goal of this is to evaluate and refine our theoretical framework, VISION, but we have also discussed the implications of the research on the US Presidential Election.

Collecting big data: semi-public web pages and semi-private social media

Our VISION framework focuses on mapping two types of cyberspace communication channels: public channels (mass media) and private channels (personal communication networks) (Figure 2) (Robinson 1976). In traditional communication research, the public channels are TVs, newspapers, radios, etc. The private channels are face-to-face conversations, local community meetings, personal letters, etc. In cyberspace, our VISION framework utilized web search engines to analyze the spread of similar web pages associated with keywords as semi-public channels. Higher ranked web pages are more "public" to users. Lower ranked web pages are less public. On the other hand, we analyzed the spread of tweets associated with keywords by Twitter API as semi-private channels. Most readers of tweets are the friends of Twitter users as "followers". Both communication channels can generate a large volume of data (Big Data), which requires a large data archive and high performance analysis cyberinfrastructure.

The VISION framework also attempts to understand the relationship between cyberspace activity and the events of realspace. In order to frame the distortion of the real world to cyberspace, we use the analogy of the Earth distorted to maps using different projections (Figure 2). As geographers, we can take a projection (which is in essence a specific distortion) and transform it back to the original coordinates. In the same vein, we are using the VISION framework to understand how to transform the cyberspace communications we collect into an accurate picture of the world.

Figure 3 illustrates two examples of communication channels (media) we collected in VISION. Figure 3a displays the web pages ranked by the Yahoo search engine with the keyword "Obama" (representing the semi-public channels) by using the Cyber-Discovery Tools. Figure 3b shows the tweets collected by the Geo-search-enabled

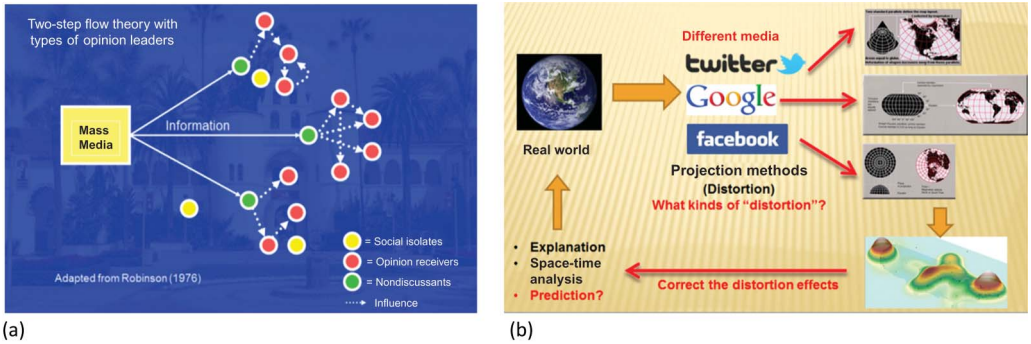


Figure 2. (a, b) The Two types of cyberspace communication channels: public mass media vs. private networks (a, left) and the distortion effects of cyberspace maps by different media (b, right).

(a)

Rank	Search Engine	Keyword	Search_Date	URL	Title	Latitude	Longitude	ZipCode
1	Yahoo	"Obama"	11/1/2012	http://www.barackobama.com/	Barack Obama	42.36259	−71.0843	2142
2	Yahoo	"Obama"	11/1/2012	http://en.wikipedia.org/wiki/Barack_Obama	Barack Obama - Wikipedia, the free encyclopedia	37.78979	−122.394	94105
3	Yahoo	"Obama"	11/1/2012	http://www.whitehouse.gov/	The White House, President Barack Obama	39.6237	−104.874	80111
4	Yahoo	"Obama"	11/1/2012	http://www.biography.com/people/barack-o	Barack Obama Biography - Facts, Birthday, Life Sto	33.7542	−118.202	90802
5	Yahoo	"Obama"	11/1/2012	http://en.wikipedia.org/wiki/President_Obar	Presidency of Barack Obama - Wikipedia, the free	37.78979	−122.394	94105
6	Yahoo	"Obama"	11/1/2012	http://www.usatoday.com/story/theoval/201	Poll: Obama ahead in Iowa, Wis., N.H.	33.7542	−118.202	90802
7	Yahoo	"Obama"	11/1/2012	http://www.barackobama.com/plans	Plans â€" Barack Obama	42.36259	−71.0843	2142
8	Yahoo	"Obama"	11/1/2012	http://news.yahoo.com/blogs/ticket/obama-	Obama and Romney make young girl cry | The Tick	37.4249	−122.007	94089
9	Yahoo	"Obama"	11/1/2012	http://www.reuters.com/article/2012/11/01/	Obama, Republican Christie tour storm-hit New Je	38.64999	−90.5334	63017
10	Yahoo	"Obama"	11/1/2012	http://www.latimes.com/news/politics/la-pn	Obama edge over Romney on healthcare issues sh	33.7542	−118.202	90802
11	Yahoo	"Obama"	11/1/2012	http://www.whitehouse.gov/administration/	President Barack Obama | The White House	39.6237	−104.874	80111
12	Yahoo	"Obama"	11/1/2012	http://politicalticker.blogs.cnn.com/2012/10/	CNN Poll: Romney 50% â€" Obama 49% in Florida â	33.0347	−96.8134	75093
13	Yahoo	"Obama"	11/1/2012	http://online.wsj.com/article/SB10001424052	Obama, Romney Trade Jabs - WSJ.com	40.3929	−74.5412	8852
14	Yahoo	"Obama"	11/1/2012	https://store.barackobama.com/	Obama for America | 2012 | Store | Homepage [of	42.36259	−71.0843	2142

(b)

tweet_datetime_local	tweet_keyword	text	user_id	lang	location	geo
11/1/2012 23:58	Barack OR Obama	So It took #Sandy for Mayor Bloomberg to endorse Obama!	258679025	en		40.826612, −73.95013
11/1/2012 23:59	Barack OR Obama	Mitt Romney legit had a good sounding plan, but Obama is getting me back	60469411	en		40.800537, −73.941136
11/1/2012 23:49	Barack OR Obama	@[redacted]r I don't think either. The Gov. has handled the disaster/ Obama	19356396	en		39.094894, −76.747587
11/1/2012 23:46	Barack OR Obama	I'm jealous of all you kids back home that got to see Obama speak earlier. Like	99205316	en		37.7799, −122.451923
11/1/2012 23:55	Barack OR Obama	that picture of obama Hanging &lt&lt&lt No .	495476978	en		30.201557, −81.596475
11/1/2012 23:53	Barack OR Obama	RT @[redacted]: Why Shepard endorses Obama http://t.co/fJaABNEI #obeygi	18301388	en	iPhone: 38.132610,−87.640678	
11/1/2012 23:52	Barack OR Obama	Donald Trump's Response To President Obama Not Accepting The $5 Million O	155355668	en	Ã?T: 40.721319,−73.75971	
11/1/2012 23:54	Barack OR Obama	RT @[redacted]n: Nate Silver's latest Electoral College forecast has	200441285	en	Washington, DC	
11/1/2012 23:56	Barack OR Obama	RT @[redacted]: Why I'm voting for Barack Obama: We are all in this together	30673652	en	Irvine, Ca.	
11/1/2012 23:56	Barack OR Obama	Plz flw @[redacted]a: Barack Obama vs. Mitt Romney Presidential Debate --	430787376	en	Hollywood, California, USA	
11/1/2012 23:56	Barack OR Obama	American Crossroads is running ads on my TV saying Obama would cut military	204158284	en	Washington, D.C.	
11/1/2012 23:55	Barack OR Obama	@Starbucks Your CEO endorses Obama? Guess I'll be endorsing Dunkin Donu	28800014	en	New York	
11/1/2012 23:54	Barack OR Obama	.@BarackObama and President Clinton are coming to Bristow. Sign up here to	37708786	en	Baltimore, MD	
11/1/2012 23:54	Barack OR Obama	Obama's US Auto Industry Better Off Than Four Years Ago: Cars - Businesswee	15703884	en	Austin, TX USA	

Figure 3. Two types of communication channels (media) collected by VISION. (a) Web page search results (by Yahoo Engine with keyword "*Obama*"). (b) Tweets search results (by Twitter API with keyword "*Barack OR Obama*").

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

Twitter Tools with keyword "Barack OR Obama" (representing the semi-private channels).

The following sections will explain the methods and tools used for collecting web page search results and geolocation-based tweets.

Collecting and mapping web pages

Most web search engines rely on web crawlers (or web robots) to collect and index web page content into a centralized database. Web crawlers are dynamic network programs designed for collecting and duplicating targeted website contents (remotely) into web index databases. The crawler can perform very comprehensive web page indexing tasks for web search engines (Brin and Page 1998). After the creation of web page index databases, the next step is to decide the ranking of hits based on specific keywords. Different search engines have adopted different ranking algorithms and methods. Current Google search APIs can only be used to retrieve up to 64 web pages from Google search engine each time. Therefore, the VISION prototype requires the use of Yahoo and Bing search engines APIs because they provide up to 1000 web pages from their APIs in a single keyword search.

After retrieving the ranked web pages from Yahoo and Bing search engine APIs, the next step is to find the geolocation of Internet Protocol (IP) addresses associated with each web page. Most geolocation operations are performed by sending requests to a WHOIS database server or to commercial IP geolocation databases, such as MaxMind or IPPage. The WHOIS database server stores hundreds of thousands web server IP addresses, domain names, and associated registration information. WHOIS databases are maintained by Regional Internet Registries (RIR). Each Internet Service Provider (ISP) has to register its web servers to RIR in order to get an assigned IP addresses for their servers and web applications. Therefore, researchers can use the WHOIS protocol to query registrant information for specific domain names or IP addresses. For example, the registrant of the "SDSU. EDU" server is "San Diego State University, 5500 Campanile Drive, San Diego, CA 92182". When an IP address is converted to a geolocation, there is one potential problem: location accuracy. Some Internet machines link to proxy servers in order to protect their geolocations and privacy. The original website IP addresses could be replaced by the proxy servers, and the geolocations of these machines might be incorrect (Svantesson 2005). Due to the limitation of current geolocation technology, we cannot guarantee the 100% successful conversion rate for all geolocation procedures. However, other research has shown relatively high rates IP address geolocation,

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

with estimates of spatial accuracy of 62% to 73% within 40 km for MaxMind databases (Shavitt and Zilberman 2010). This can be seen in Figure 3 as the third record, www.whitehouse.gov, would typically be considered to be in Washington, DC, but is placed in our system as being located in Denver, CO. This is a problem, but published research claims this is not typical and we are in the process of examining this issue on a much larger scale within our research.

In the VISION framework, we developed the Cyber-Discovery Tools, which combine the web search engine APIs from multiple search engines and IP Address Lookup Service from the MaxMind database (using the free version). The Cyber-Discovery Tools can automatically generate ranked web pages (with URLs) associated with keywords with geocoded coordinates. We used the tools to search two keywords, "Obama" and "Romney" weekly from 18 December 2011 to 7 November 2012 in both Yahoo and Bing search engines. We collected over 45 weeks of datasets, but missed 2 weeks of data due to our server errors. After the 45 weeks, we collected over 44,200 web pages related to the "Romney" keyword and another 44,200 web pages for the "Obama" keyword. This article only reveals a small portion of our collected web page datasets.

Collecting and mapping social media messages (tweets)

Social media (such as Twitter and Facebook) are powerful communication platforms for idea exchange, breaking news, personal networking, political opinions, and collective actions. By using smartphones, personal computers, and mobile devices, people can communicate and coordinate their activities geospatially, and to a significant degree, to accomplish these social communication functions in near-real time. The rich information available in social media can now be monitored, traced, and analyzed in ways that may assist researchers understanding of various diffusion processes, human behaviors, and the collective moods around the world (Newsam 2010; Perreault and Ruths 2011; Golder and Macy 2011; Lee, Wakamiya, and Sumiya 2011).

Twitter is a popular online micro-blogging service established in 2006. Users can write and broadcast short messages (restricted to 140 characters) to their "followers" in Twitter. These short messages are called "tweets", which are searchable by keywords, authors, and hashtags (#). Twitter has over 140 million active users in 2012 and generates over 340 million tweets daily (Twitter 2012). The age demographics of Twitter are slaneted toward the youth, with users aged 18–24 averaging nearly two and a half times as many hours on social media as users aged 65 and above (Nielsen 2012). Scientists can analyze this huge collection of tweets and their content to conduct both qualitative and quantitative analysis of social communication. This new approach provides an unprecedented opportunity to research social networks and human communication (Miller 2011; Stefanidis, Crooks, and Radzikowski 2011).

While privacy may be an issue for some, the location tracking service of Twitter is an "opt-in" service, meaning that users must allow Twitter to track their locations as opposed to this being a default setting. With deep questions about the privacy of the "geoweb" in general, a future discussion would be needed to further evaluate the ethics of using geographic information science and large databases with social media (Elwood and Leszczynski 2011). Privacy can also be an issue in data accuracy, as those who value privacy may not use accurate information or may not update information. Web demographics consider this a sincere concern, and some sites use an "opt-out" style that requires the user to remove their information as a balance between database integrity and privacy protection (Chow 2013).

Our research team developed the Geo-search-enabled Twitter Tools utilizing the official Twitter search APIs. The Python programs can retrieve tweets by using keywords and by defining searchable spatial range. Search results including *user names, user ID, tweet text content, created_time, and spatial locations* were saved into Structured Query Language (SQL) database and exported to excel files for analysis and visualization purposes. The spatial locations of tweets were tagged by the Twitter API automatically. We performed searches using two candidates' full name "Barack OR Obama" and "Mitt OR Romney" to capture tweets mentioning the two candidates in full name or first/last name only. Regarding the study area, we selected the top 30 US cities (by population) and set up a spatial range to cover major metropolitan areas without overlapping each other. The center of each city was defined using the GeoNames map centers and the spatial radius was set as 17 miles from each city center. The span of 17 miles was selected to cover the metropolitan areas of our 30 cities without overlapping nearby cities such as Washington DC and Baltimore. We compared the weekly and daily numbers of tweets collected by each candidate keywords (as the "attention" level index) to the poll data and the final election results. These comparison results are highlighted in the later section of "Geolocation-Based Tweet Analysis".

Visualizing the dynamic change of web page information landscapes

There are various spatial analysis methods applicable for mapping web page search results, such as Thiessen (Voronoi) polygons, Inverse Distance Weighting, or simple Kriging. We selected the kernel density methods because the kernel density method reflects the "probability" concept of IP geolocations: the contents of web pages are more likely to be associated with the geolocation of IP addresses. For example, the content of San Diego State University (SDSU) web page is more likely to be

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

associated with the actual geolocation of SDSU server's IP address, which is registered as "5500 Campanile Drive, San Diego, California" in the WHOIS database. In addition, many points (web pages) overlap (with the same server IP addresses, or geolocation coordinates). The kernel density method can better represent the "density" of points in the overlap situation.

In our design, the ranking numbers of web page search results were considered as the "popularity" or the "population" in the kernel density algorithm. A higher ranked web page is more "popular" and has a higher probability value compared to a lower ranked web page. Therefore, we converted the ranking numbers into the population parameter. After we created the kernel density maps of web pages associated with various keywords, we found that higher density areas of web page IP geolocations are associated with major US cities with bigger population, such as New York and Los Angeles. This indicates that the density (or geographic probability) of web pages may be closely related to the size of city populations.

Our next step was to calculate the differences between two different keyword maps, such as "Mitt Romney" vs. "Barack Obama". A raster-based map algebra tool from ArcGIS was used with the following formula:

Differential Value = (Keyword-A/Maximum-Kernel-Value-of-Keyword-A) – (Keyword-B/Maximum-Kernel-Value-of-Keyword-B)

The differential information landscape map illustrates important *geospatial fingerprints* hidden in the text-based web search results depending on the context of selected keywords. In this article, geospatial fingerprints are defined as *the unique spatial patterns (e.g., clusters) of web information landscapes associated with different keywords or concepts*. One important aspect of the creation of information landscapes is the selection of the kernel density threshold (radius). We used 2 map units (around 100 miles) to reflect the average size of US cities (including suburban areas). We noticed that changing threshold distances adopted in kernel density operations can result in drastically different spatial patterns and relationships at various map scales. The spatial scale dependency reflects the nature of geospatial fingerprints and the spatial characteristics of web information landscapes.

Figure 4 illustrates two weekly web page information landscapes with the differential value between "Romney" and "Obama". The red color areas have relative higher probability of hosting "Romney" related web pages comparing to the probability of hosting "Obama" web pages based on their web server IP address geolocations. The blue color areas have relative higher probability of hosting "Obama" web pages comparing to the probability of hosting "Romney" web pages. The changes in color intensity is not related to an overall increase in collected web pages as there were 505 web pages collected for week 37 and 488 web pages collected for week 38.

The color patterns in the differential value maps illustrated some interesting "signals" or "geospatial fingerprints" about the two keywords ("Romney" and "Obama"). In week 38, the red color areas have significantly increased comparing to the previous week (week 37). This change may be related to the Republican National Convention in Tampa, Florida on August 27–30, 2012. In the Figure 4b (week 38), Salt Lake City in Utah has very high probability of hosting "Romney" web pages. This might be related to his previous political connection to the Salt Lake City and his religious preferences. On the other hand, Chicago, Illinois shows the higher probability of hosting web pages related to "Obama" due to his political connection (as his former chief of staff is the city mayor, and Obama was the US Senator from Illinois). However, there are some patterns which are difficult to explain, such as the blue areas in Denver. One interesting observation is that the Convention was hold in Tampa, Florida, but the major hot zones (red color areas) are not in the same location. The changes in visual patterns represent the temporal changes of the "geospatial fingerprint". Depending on the date of collection, the "fingerprint" may fluctuate. However, we are primarily concerned with the most significant probabilities and the most significant changes within these fingerprints. Early on, we have been using visual techniques to examine changes that may relate to major news events, but plan on using more sophisticated correlation techniques as we refine the process to better understand the temporal changes.

In the following few months, our weekly comparisons showed some similar observations that the dynamic changes of web information landscapes are closely related to the real world events, such as the 11 September tragedy in Benghazi (the killing of the US Ambassador) and the Second Presidential Debate on October 17, 2012. The completed series of web information landscapes can be accessed from our project website: http://mappingideas.sdsu.edu/mapshowcase/election/webpage/election3.html.

Analyzing and mapping geolocation-based tweets

In addition to analyzing the weekly changes of web page information landscapes of two presidential candidates, our project utilized the Geo-search-enabled Twitter Tools to collect over 16,751,331 tweets using keywords (related to the two candidates' names) from the selected 30 US cities from 25 June 2012 to 5 November 2012. We realized that there are many "noises", "errors", "biases", and "distortion" within these collected tweets. For example, over 30% of tweets are RT (retweets) and over 20% of tweets are generated by "robots" or media tools (based on our preliminary analysis). It should also be noted that geolocated tweets are a small percentage of all tweets; research by Hale, Gaffney, and Graham (2012) states it as low as

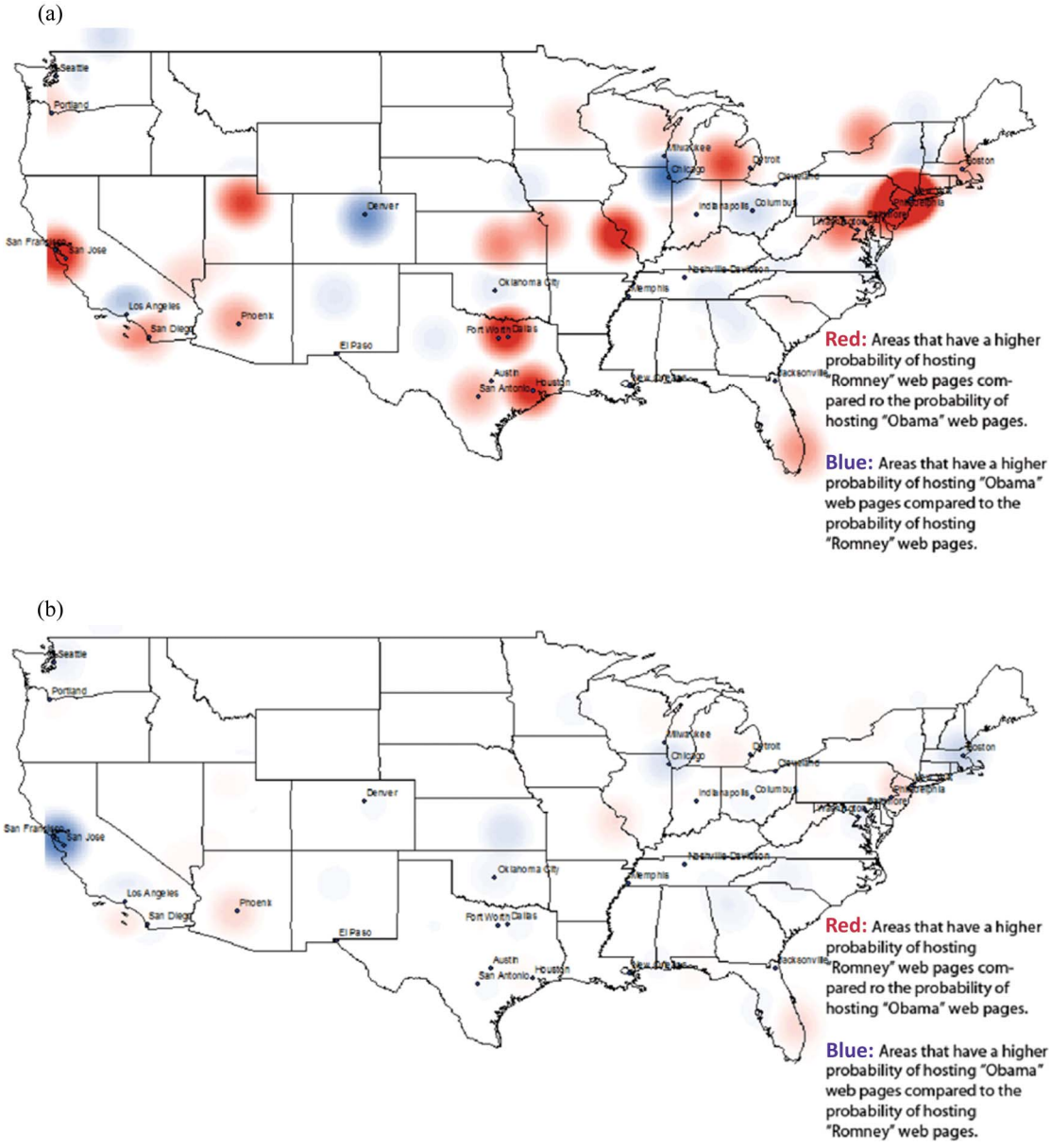


Figure 4. The weekly change of web information landscapes (comparing "Romney" web page density probability vs. "Obama" web page density probability from Week 37 to Week 38). (a) 26 August 2012 (Week 37) "Romney" web pages vs. "Obama" web pages. (b) 3 September 2012 (Week 38) "Romney" web pages vs. "Obama" web pages.

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

0.7% while Takhteyev, Gruzd, and Wellman (2012) present evidence that the ratio is as high as 6%.

The higher number of tweets associated with keywords may not indicate the supporting rate or true popularity. In fact, there are multiple reasons that a user may retweet something, ranging from sharing an idea, starting a conversation, or to entertain the user's followers (Boyd, Golder, and Lotan 2010). We use the term "attention levels" rather than "popularity" to indicate the higher numbers of tweets associated with each candidate. This is not to ignore the differences between social contagion (social influence) and homophily (simple sharing of traits). While we are looking for the diffusion of ideas in a pattern that reflects social contagion, it cannot be discounted that we are observing homophily, and are conscious of the uncertainty in contagion (Miller 2011, 1815). Another source of potential bias in our tweet analysis is that we only collect tweets from the 17 mile radius of major US cities, where the urban population profile may prefer the Democratic candidate. While there could be an issue

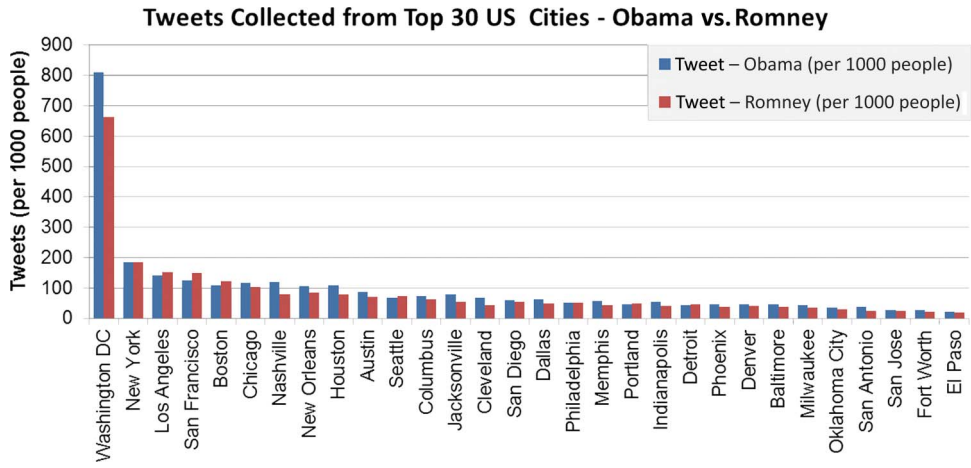


Figure 5. Accumulated tweets (per 1000 people) in top 30 US cities from June 25 to November 05, 2012 (total number of collected tweets: 16,751,331). The order of the cities is based on the total number of tweets per 1000 people for each city, starting with the highest number of tweets per 1000 people.

related to only collecting urban tweets, it could have been worse to add in the rural tweets directly. The difference in the frequency of tweets from urban vs. rural areas creates what is essentially a different scale for the different regions. There are too few tweets from rural areas to apply it to this research. The following analysis of tweets are only based on the original numbers of tweets without any filtering or cleaning processes due to the limitation of time and resources. Surprisingly, the raw data tweets are still highly related to some events and changes in the real world.

Figure 5 illustrated the normalized numbers of tweets by the 17 miles radius population in each city and the comparison of tweet "attention levels" (represented by the numbers of tweets per 1000 people) between the two candidates (Obama vs. Romney). Washington DC has highest ratio of tweets per 1000 people comparing to other cities.

Figure 6 illustrated the pie chart maps to compare the changes of tweet attention levels before and after Hurricane Sandy. Hurricane Sandy was a devastating storm causing severe damages to the US East Coast in October 2012, a week before the 2012 Presidential Election. This event created a significant change of tweet attention levels between the two candidates based our tweet collection (Figure 6). The size of circle indicates the total numbers of tweets divided by the city population. Bigger circles mean more people submitted tweets in the city during that day. Comparing Figure 6a and b, the circles of East Coast cities (Washington DC, New York, and Boston) have increased significantly after Hurricane Sandy. The attention levels between "Obama" and "Romney" also changed in these cities. For example, in New York City, the attention percentage of Romney has changed from 56% (October 24, 2012) to 34% (November 01, 2012). Overall, nine of the thirty cities changed from a majority Romney attention percentage to a majority Obama attention percentage, and many others increased the Obama percentage compared to the Romney percentage.

Figure 7 illustrates the daily comparison of the total numbers of tweets between the two candidates (combining all 30 US cities) from October 28 to November 4, 2012. We found that the changes of tweet attention levels between two candidates are similar to other official polls regarding the Presidential Election. "Obama" keyword's attention level became higher than "Romney" keyword after October 31 when Hurricane Sandy caused significant damages in the New York City. But, the gap between the two candidates' attention levels became smaller in the last 2 days (November 3 and 4, 2012).

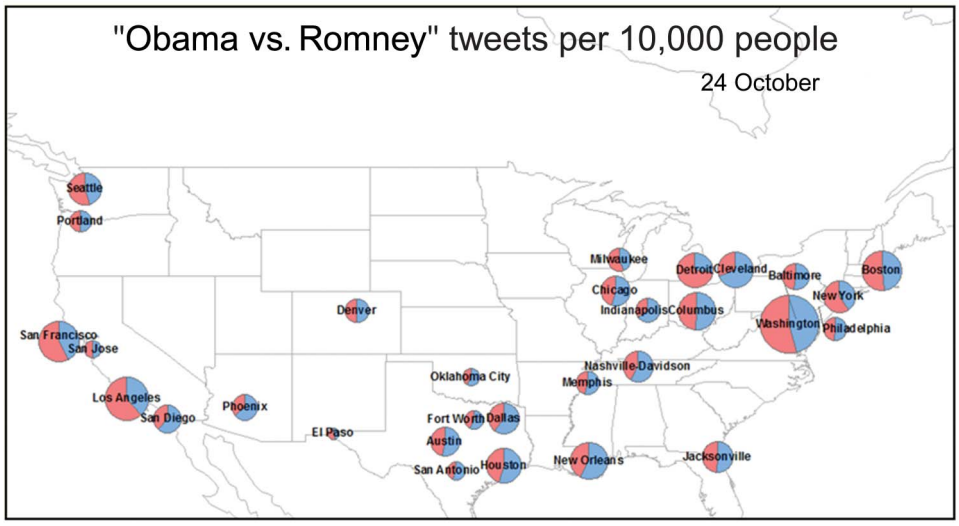
Top vocabulary items in weekly tweets

In addition to the tweet attention level analysis, we also conducted sentiment analysis by calculating the frequency of vocabulary items mentioned in tweets. To reveal the trending topics of these tweets, our research team developed a Python script (called *Vocab*) which reads millions of tweets from our excel files and extracts the most frequent vocabulary items used within a week or a month. A long list of "stopwords" are fed to the *Vocab* to recognize and ignore the common words people use in sentences. The output of *Vocab* is the top 800 most frequent vocabulary items and the separated counts of how many times each word shows up. *Vocab* is also programmed to easily integrate with R, a free statistical software, for result visualizations. Word clouds were created by using the "wordcloud" library of R. The sizes of words in the word cloud are based on their

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

(a)



(b)

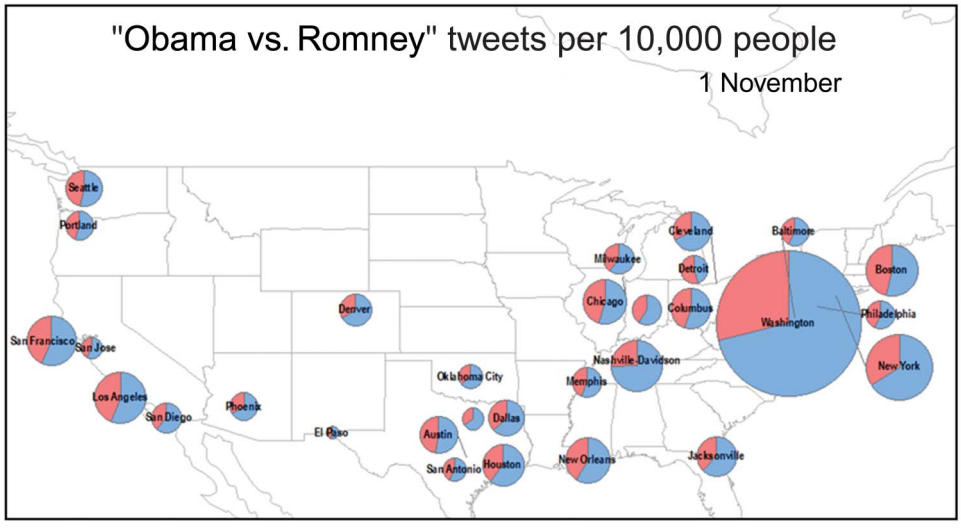


Figure 6. The comparison of tweet attention levels between "Romney" and "Obama" before and after the Hurricane Sandy. (a) The tweet attention levels between Obama and Romney on 24 October 2012 (before the Hurricane Sandy). Red: "Romney" related tweets, Blue: "Obama" related tweets. (b) The tweet attention levels between Obama and Romney on 1 November 2012 (after the Hurricane Sandy). Red: "Romney" related tweets, Blue: "Obama" related tweets.

frequencies from weekly aggregated tweets and trending topics could be seen from the larger words in the word clouds (Figure 8). In the word clouds, we excluded the keywords of candidates' names, such as "Obama", "Barack", "Mitt", and "Romney", because these keywords are always the highest ranked keywords in our collected tweets (being that they must be in the tweet to collect it).

Figure 8 illustrated four different word clouds, created by extracting the most frequent vocabulary items from tweets 1 week before Hurricane Sandy and 1 week after Hurricane Sandy. The two left-side clouds are the vocabulary from Obama related tweets and the two right-side clouds are from Romney related tweets.

For both of the candidates, the before-Sandy tweet vocabulary was centered around the final presidential debate (held on 22 October 2012). This can be seen from the top result in each cloud, "debate", and from the other high results, such as "foreign" and "policy". However, immediately after Hurricane Sandy, the vocabulary shifted to weather and relief-related

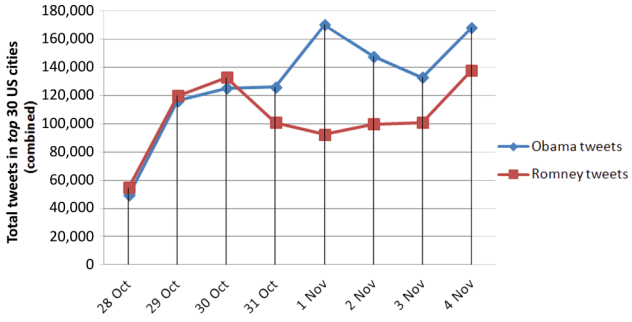


Figure 7. The daily comparison of the total numbers of tweets between the two candidates (combining all 30 US cities) from 28 October 2012 to 4 November 2012.

topics. Top vocabulary terms for Obama were "sandy", "hurricane". For Romney, the top terms were "fema" (for the Federal Emergency Management Agency), "sandy", and "campaign". In 1 week, the national focus related to each candidate switched from the debates, which had been significant for most of the month of October, to Hurricane Sandy, and the aftermath of the destruction. The attention on Romney tended to focus on critical statements regarding FEMA from a Republican primary debate on 13 June 2011. Many criticized Romney's view on the elimination of the organization in light of the need for relief after Hurricane Sandy.

Figure 9 displays the word clouds as raw vocabulary lists, and the stark difference between the 2 weeks can be seen. The term "sandy" is in the top three for both candidates after Hurricane Sandy, while "debate", the former top result, has disappeared from the top 25 results of both candidates. There are also storm-related terms throughout the post-Sandy list, such as "storm" and "disaster". The term "christie", referring to New Jersey Governor Chris Christie, rose to ninth in the list post-Sandy, probably due to complementary comments made by Christie in the aftermath of the storm. These comments were significant as Christie was known as an outspoken opponent of Obama before the storm.

Summary and future research

This research demonstrated a research framework, VISION, for tracking and analyzing the spatial content of social media (Twitter) and web pages, visualized the

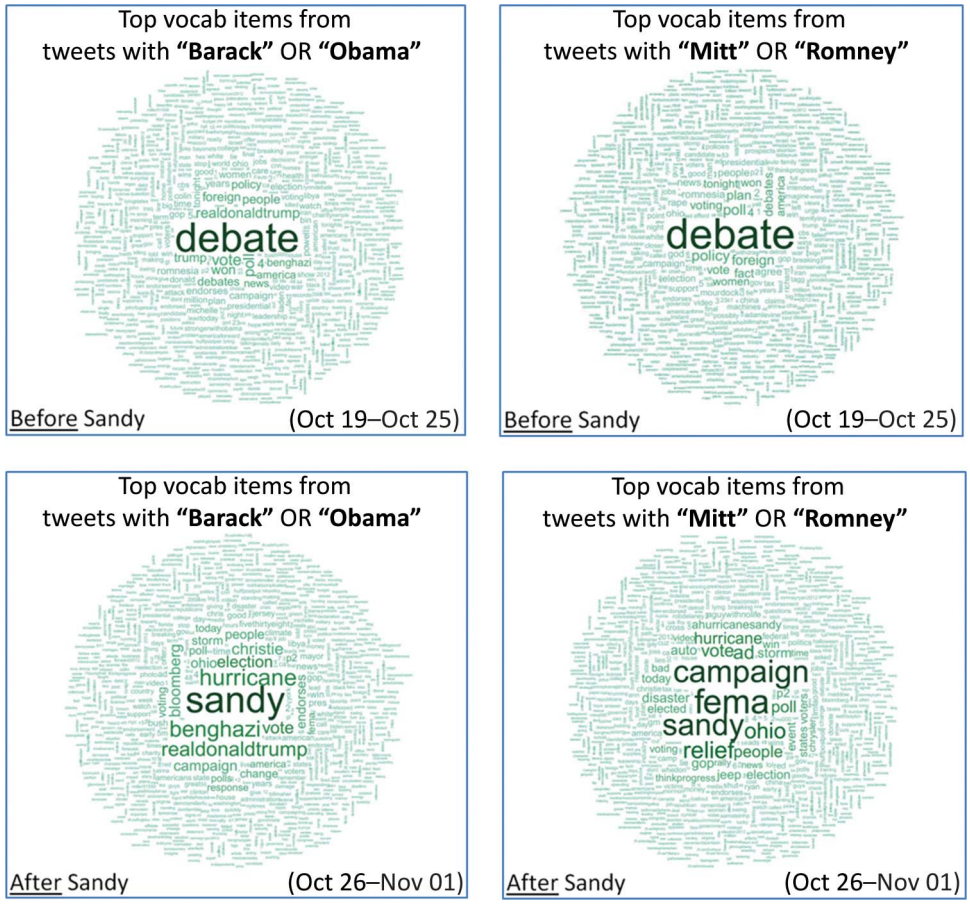


Figure 8. The comparison of word clouds between Obama related tweets and Romney related tweets (before and after Hurricane Sandy).

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

19–25 October 2012	
Obama-pre-Sandy	Counts
debate	111709
vote	39814
poll	37225
realdonaldtrump	33297
won	31673
people	28492
foreign	28147
trump	27968
4	27676
policy	26479
debates	26337
news	23398
america	22861
benghazi	22624
years	21852
endorses	20793
tonight	20704
campaign	20517
women	20112
romnesia	19949
election	19848
time	19527

26 October–1 November 2012	
Obama-post-Sandy	Counts
sandy	65884
hurricane	38607
benghazi	38087
realdonaldtrump	33787
vote	27199
election	26401
christie	25650
bloomberg	24960
campaign	22204
ohio	21225
people	20095
endorses	19188
storm	18252
poll	17575
change	17127
voting	16790
today	14766
america	14441
fema	14076
polls	13691
p2	13404
response	13358

19 October–25 October 2012	
Romney-pre-Sandy	Counts
debate	106698
policy	31394
poll	31291
foreign	30034
plan	27251
debates	25992
voting	25818
vote	25440
fact	24051
women	23725
4	23538
america	23134
won	21918
tonight	21478
2	20713
agree	19716
people	19712
romnesia	19207
news	19037
rape	18856
god	17725
ohio	17637

26 October–1 November 2012	
Romney-post-Sandy	Counts
fema	45806
campaign	41799
sandy	40311
relief	30340
ohio	27780
ad	23796
vote	22564
hurricane	20201
people	19945
poll	19709
storm	15930
disaster	15919
auto	15614
gop	15520
jeep	15159
election	14476
ahurricanesandy	13972
p2	13654
elected	13434
event	13393
today	13349
states	12897

Figure 9. The top 25 vocabulary list of the tweets between Obama and Romney (before Hurricane Sandy and after Hurricane Sandy).

dynamic comparison of web information landscapes, and examined the correlation between the popularity of candidates on Twitter and the actual election results. Social media and web information landscapes have much potential to be applied in the election campaign or poll analysis. But we need to develop more comprehensive data analysis methods and data cleaning algorithms to reduce the noises and errors in social media data.

By tracking and analyzing the contents of tweets and web pages, researchers might be able to reveal important social contexts of specific events (such as presidential elections) and understand the temporal and spatial relationships among these short messages and human behaviors (Tsou et al. 2012). The digitization of social media and web pages may be able to provide massive data and facilitate the emergence of a data-driven computational social science (Lazer et al. 2009). Analyzing the spatial and temporal dynamics of "collective thinking of human beings" in social media and web pages could lead to improved comprehension of the factors behind those ideas, events, and the manifold human behaviors that result, which is important in reducing misunderstandings and strategizing how to address controversies and conflicts in the world.

Acknowledgments

This article is based upon work supported by the National Science Foundation under Grant No. 1028177, project titled "CDI-Type II: Mapping Cyberspace to Realspace: Visualizing and Understanding the Spatiotemporal Dynamics of Global Diffusion of Ideas and the Semantic Web". Any opinions, findings, and conclusions or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Adams, P. C. 2010a. *Geographies of Media and Communication: A Critical Introduction*. Malden, MA: Wiley-Blackwell.

Adams, P. C. 2010b. "A Taxonomy for Communication Geography." *Progress in Human Geography* 35 (1): 37–57.

Andrés, L., D. Cuberes, M. Diouf, and T. Serebrisky. 2010. "The Diffusion of the Internet: A Cross-Country Analysis." *Telecommunications Policy* 34: 323–340.

Blaut, J. M. 1987. "Diffusionism: A Uniformitarian Critique." *Annals of the Association of American Geographers* 77 (1): 30–47.

Boyd, D., S. Golder, and G. Lotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." In *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS-43)*, Honolulu, 1-10, CD-ROM, January 2010. Los Alamitos, CA: IEEE Computer Society.

Brin, S., and L. Page. 1998. "Anatomy of a Large-Scale Hypertextual Web Search Engine." In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, QLD, April 14–18, 107–117.

Brown, L. 1981. *Innovation Diffusion: A New Perspective*. London: Methuen.

Chow, T. E. 2013. "We Know Who You Are and We Know Where You Live: A Research Agenda for Web Demographics." In *Crowdsourcing Geographic Knowledge*, edited by D. Sui, S. Elwood, and M. Goodchild, 265–285. Dordrecht: Springer.

Elkink, J. A. 2011. "The International Diffusion of Democracy." *Comparative Political Studies* 44 (12): 1651–1674.

Elwood, S., and A. Leszczynski. 2011. "Privacy, Reconsidered: New Representations, Data Practices, and the Geoweb." *Geoforum* 42 (1): 6–15.

Gibson, W. 1984. *Neuromancer*, 69. New York: Ace Books.

Golder, S. A., and M. W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333: 1878–1881.

Hägerstrand, T. 1966. "Aspects of the Spatial Structure of Social Communication and the Diffusion of Information." *Papers in Regional Science* 16 (1): 27–42.

Hägerstrand, T. 1967. *Innovation Diffusion as a Spatial Process*. Chicago, IL: The University of Chicago Press.

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015

Hale, S., D. Gaffney, and M. Graham. 2012. "Where in the World Are You? Geolocation and Language Identification in Twitter." Working Paper. Accessed March 03, 2013. http://www.geospace.co.uk/files/icwsm_paper2.pdf

Lazer, D., A. Pentland, L. Adamic, S. Aral, A. L. Baraba´si, D. Brewer, and M. Van Alstyne. 2009. "Computational Social Science." *Science* 323: 721–723.

Lee, R., S. Wakamiya, and K. Sumiya. 2011. "Discovery of Unusual Regional Social Activities Using Geo-Tagged Microblogs." *World Wide Web* 14: 321–349.

Lerman, K., and R. Ghosh. 2010. "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks." In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Palo Alto, CA, May 23–26. Washington, DC: George Washington University.

Miller, G. 2011. "Social Scientists Wade Into the Tweet Stream." *Science* 333: 1814–1815.

Newsam, S. 2010. "Crowdsourcing What Is Where: Community-Contributed Photos as Volunteered Geographic Information." *IEEE Multimedia: Special Issue on Mining Community-Contributed Multimedia* 17 (4): 36–45.

Nielsen. 2012. *State of the Media: The Social Media Report 2012.* Accessed March 03, 2013. http://www.nielsen.com/us/en/reports/2012/state-of-the-media-the-social-media-report-2012.html

Paul, M., and M. Dredze. 2011. "You Are What You Tweet: Analyzing Twitter for Public Health." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, July 17–21.

Perreault, M., and D. Ruths. 2011. "The Effect of Mobile Platforms on Twitter Content Generation." In *Proceedings of the International Conference on Social Media and Weblogs*, Barcelona, July 17–21.

Postmes, T., and S. Brunsting. 2002. "Collective Action in the Age of the Internet: Mass Communication and Online Mobilization." *Social Science Computer Review* 20 (3): 290–301.

Rainie, L., and B. Wellman. 2012. *Networked: The New Social Operating System*. Cambridge, MA: MIT Press.

Robinson, J. 1976. "Interpersonal Influence in Election Campaigns: Two-Step Flow Hypotheses." *The Public Opinion Quarterly* 40: 304–319.

Rogers, E. M. 1962. *Diffusion of Innovations*. Glencoe: Free Press.

Rogers, E. M. 2003. *Diffusion of Innovations*. 5th ed. New York: Free Press.

Shavitt, Y., and N. Zilberman. 2010. "A Study of Geolocation Databases." Accessed March 03, 2013. http://arxiv.org/abs/1005.5674

Stefanidis, A., A. Crooks, and J. Radzikowski. 2011. "Harvesting Ambient Geospatial Information from Social Media Feeds." *GeoJournal* 78 (2): 1–20.

Svantesson, D. J. B. 2005. "Geo-Identification: Now They Know Where You Live." *Privacy Law & Policy Reporter* 11 (6): 171–174.

Takhteyev, Y., A. Gruzd, and B. Wellman. 2012. "Geography of Twitter Networks." *Social Networks* 34: 73–81.

Tsou, M.-H., D. Lusher, J.-A. Yang, D. Gupta, J. M. Gawron, B. H. Spitzberg, L. An, and S. Wandersee. 2012. "Mapping Social Activities and Concepts with Social Media (Twitter) and Web Search Engines (Yahoo and Bing): A Case Study in 2012 U.S. Presidential Election." In *The Proceedings of AutoCarto 2012 International Symposium*, edited by S. Battersby, Columbus, OH, September 16–18. Mt. Pleasant, QLD: Cartography and Geographic Information Society.

Twitter. 2012. *Twitter Turns Six*. Twitter Blog. Accessed May 1, 2012. http://blog.twitter.com/2012/03/twitter-turns-six.html

Vieweg, S., A. L. Hughes, K. Starbird, and L. Palen. 2010. "Microblogging during Two Natural Hazards Events: What Twitter may Contribute to Situational Awareness." In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, 1079–1088. Atlanta, GA: ACM.

Yin, L., S.-L. Shaw, and H. Yu. 2011. "Potential Effects of ICT on Face-to-Face Meeting Opportunities: A GIS-Based Time-Geographic Exploratory Approach." *Journal of Transport Geography* 19: 422–433. doi:10.1016/j.jrrangeo.2010. 09.007.

Downloaded by [SDSU San Diego State University] at 17:15 06 November 2015