Citation: Tsou, M-H., Lusher, D., Yang, J-A., Gupta, D., Gawron, J. M., Spitzberg, B. H., An, L., & Wandersee, S. (2012, September). 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 Sarah Battersby edited, *AutoCarto International Symposium on Automated Cartography Proceedings* (Columbus, OH): Mt. Pleasant, South Carolina, Cartography and Geographic Information Society.

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

Ming-Hsiang Tsou, Daniel Lusher, Jiue-An Yang, Dipak Gupta, Jean Mark Gawron, Brian Spitzberg, Li An, Sarah Wandersee

ABSTRACT: We introduce a new research framework for analyzing the spatial distribution of web pages and social media (Twitter) with related contents, called Spatial Web Automatic Reasoning and Mapping System (SWARMS). This new method can facilitate the tracking of ideas and social events disseminated in cyberspace. Hundreds of web pages and thousands of tweets associated with the same keywords were converted into visualization maps using commercial web search engines (Yahoo API and Bing API), a social media search engine (Twitter API), IP geolocation methods, and GIS functions (e.g., kernel density and raster-based map algebra methods). During our early tests, 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 U.S. Presidential Election and the Republican candidates as our case study to validate this method. There are many interesting patterns found in our comparisons. For example, we noticed that the geographic probability of hosting “Mitt Romney” related web pages in Salt Lake City is much higher compared to the probability of hosting “Newt Gingrich” web pages in the same city. We hope that this new approach may provide a new research direction for studying human thought, web content, and social activities.

Keywords: web information landscapes, geographic probability, geospatial fingerprints, web search engines, social media.

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 20th century media—such as newsletters, advertisements, physically proximal group meetings, and telephone conversations. Cyberspace (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. These research efforts can help us understand the diffusion of innovations (Roger 1962), a dynamic process whereby new concepts, ideas, and technologies spread through our society via cyberspace and digital social networks over time.

This paper introduces a new research method, called the Spatial Web Automatic Reasoning and Mapping System (SWARMS) (http://mappingideas.sdsu.edu). SWARMS is designed to track spatial patterns of publically-accessible web pages and semi-private social media (such as Twitter or Facebook) based upon searching predefined clusters of keywords determined by domain experts (Figure 1). Web pages and tweets associated with the same keywords were converted into visualization maps using GIS analysis functions and geolocation methods.

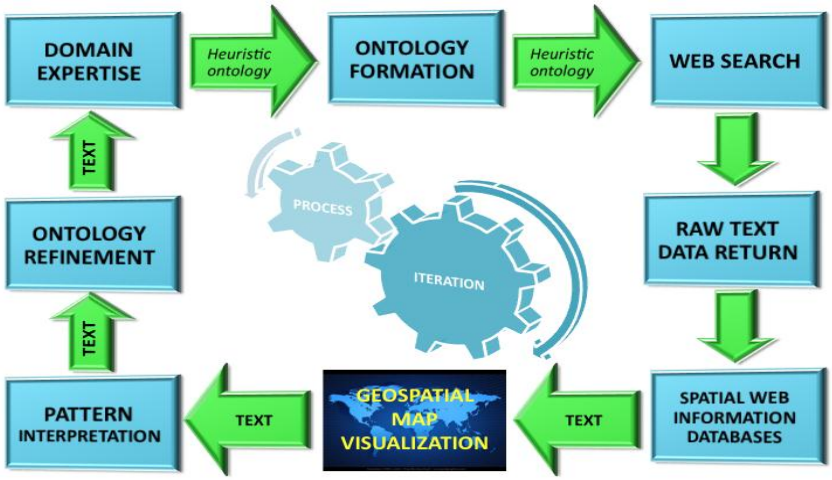

Figure 1. The Spatial Web Automatic Reasoning and Mapping System (SWARMS) framework.

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 websites. Hundreds of web pages were geocoded with real world coordinates and represented in the form of web information landscapes. These web information landscapes (maps) can help us monitor the spatial and temporal distribution patterns of web pages that coincide to reveal the nature of significant events, radical concepts or epidemics. Understanding the diffusion and acquisition patterns of web information in

response to disasters, terrorism, and epidemics may significantly facilitate intervention responses, and eventually, prevention responses.

This multidisciplinary approach can demonstrate a new methodology for visualizing and analyzing the Web and social media contents from a spatial perspective. Our new approach may be used for a variety of national security topics, including breaking news, epidemics, regional conflicts, and the spread of involvement in radical religious and other violent hate groups. Our research extends the scope of spatial analysis from physical world phenomena to cyberspace contents. Applications of web information landscapes can be extended to many areas including marketing, homeland security, public health, public policy making, and business. For example, we can compare the spatial distribution of popular web pages containing "iPhone" or "Android" keywords to study the regional preferences of smart phone platforms. By analyzing the spatial-temporal changes of keywords "suicide" and "whooping cough", we may identify the trend of these events or diseases and create new public health policies to mitigate the damages.

In this paper, we used the 2012 U.S. Presidential Election and the Republican candidates as our case study to validate this method. Three types of comparison methods were applied for the analysis of web page information landscapes:

1. Comparison of two maps with "Mitt Romney" versus "Newt Gingrich" (both are 2012 Republican Presidential Candidates).
2. Comparison of two maps with "Mitt Romney" versus "Rick Santorum" (both are 2012 Republican Presidential Candidates).
3. Comparison of the temporal changes of maps with a single keyword. For example, we can compare the web page density of "Mitt Romney" keyword on January 04, 2012 versus "Mitt Romney" keyword on January 10, 2012.

Two Types of Communications: Semi-Public Web Pages and Semi-Private Social Media

Following the concepts of diffusion innovation by Rogers (1962) and Hägerstrand (1967), the SWARMS prototype focuses on mapping two types of 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 SWARMS prototype utilized web search engines (Yahoo and Bing) 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". Occasionally, some celebrities may have over millions of followers and their tweets become more public than private messages.

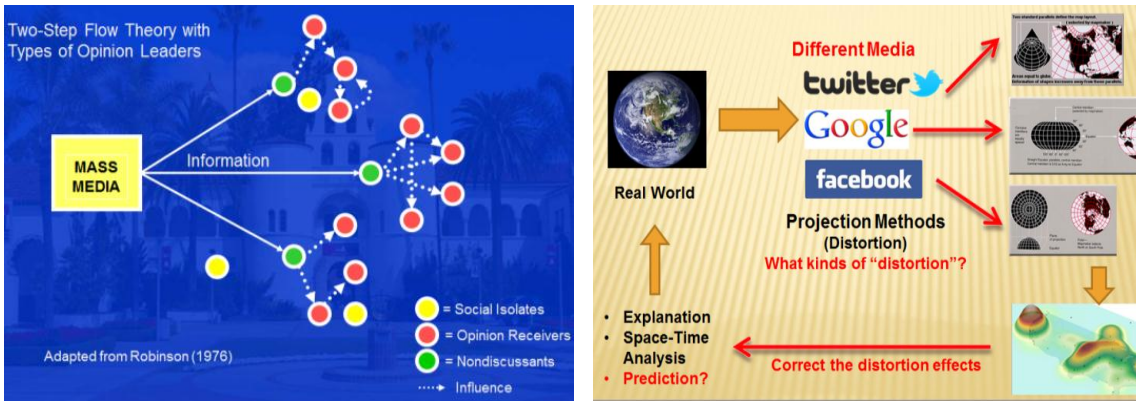

Figure 2. The Two types of communication channels: public mass media vs. private networks (left) and the distortion effects of cyberspace maps by different media (right).

Figure 3 illustrates two types of communication channels (media) we collected in SWARMS. The top one is the web pages ranked by Yahoo search engines with keyword "Mitt Romney" (representing the semi-public channels). The bottom one is the tweets collected by Twitter API and python scripts (representing the semi-private channels).

a) Web page search results (by Yahoo Engine with keyword "Mitt Romney".

	Rank	Search Engine	Keyword	Saerch_Da	URL	Title	Type	Latitude	Longitude	ZipCode
2	1	Yahoo	Mitt+Romney""	2012-01-10	http://www.mittromney.com/	Mitt Romney for President | Mitt Romney<		47.5839	-122.2995	98144
3	2	Yahoo	Mitt+Romney""	2012-01-10	http://en.wikipedia.org/wiki/Mitt_	Mitt Romney - Wikipedia, the free encycloped		37.7898	-122.3942	94105
4	3	Yahoo	Mitt+Romney""	2012-01-10	http://www.telegraph.co.uk/nev	New Hampshire primary: Mitt Romney says h		33.9642	-118.0266	
5	4	Yahoo	Mitt+Romney""	2012-01-10	http://www.washingtonpost.co	Mitt Romney comes under attack after win in		33.7542	-118.2019	90802
6	5	Yahoo	Mitt+Romney""	2012-01-10	http://en.wikipedia.org/wiki/Mitt_	Mitt Romney presidential campaign, 2008 - W		37.7898	-122.3942	94105
7	6	Yahoo	Mitt+Romney""	2012-01-10	http://www.biography.com/peo	Mitt Romney Biography - Facts, Birthday, Life		33.7542	-118.2019	90802
8	7	Yahoo	Mitt+Romney""	2012-01-10	http://www.huffingtonpost.com	Mitt Romney Plan Raises Taxes On Poor Far		33.7542	-118.2019	90802
9	8	Yahoo	Mitt+Romney""	2012-01-10	http://www.politico.com/news/s	GOP elites: Hard to stop Mitt Romney now - J		33.7542	-118.2019	90802
10	9	Yahoo	Mitt+Romney""	2012-01-10	http://elections.nytimes.com/2(	Mitt Romney - Election 2012 - NYTimes.com		47.5839	-122.2995	98144
11	10	Yahoo	Mitt+Romney""	2012-01-10	http://www.csmonitor.com/tag	Topic: Mitt Romney - CSMonitor.com		33.7542	-118.2019	90802
12	11	Yahoo	Mitt+Romney""	2012-01-10	http://www.cbsnews.com/830	Mitt Romney: I like being able to fire people" fo		33.7542	-118.2019	90802
13	12	Yahoo	Mitt+Romney""	2012-01-10	http://abcnews.go.com/blogs/p	Mitt Romneyâ€™s Moment - ABC News		33.7866	-118.2987	91521
14	13	Yahoo	Mitt+Romney""	2012-01-10	http://www.boston.com/Bostor	Mitt Romney takes Iowa by 8 votes over Rick		42.3477	-71.0384	02210

Mitt-Romney-Jan10-2012-Yahoo

b) Tweets search results (by Twitter API with keyword "Mitt Romney".

	text	from_user_name	to_user	location	from_user	created_at
1086	Romney Making a Big Push Days Before Iowa Caucus: Republican Presidentia	FOX 42 News (KPTM)		Omaha, NE	fox42news	Mon, 02 Jan 2012 02:48:08
1087	@LarrySabato Will Newt or Santorum bring up Bain Capital against Romney sir	Kevin Olson	LarrySabato	Iowa	baseballPOL	Mon, 02 Jan 2012 02:34:09
1088	21 voicemails from Mitt Romney..#stop	Erin Swartzendruber		looowwwaaa!	erin_swartze	Mon, 02 Jan 2012 02:27:27
1089	A closed-down Blockbuster on Ingersoll. That's Mitt Romney's campaign HQ. B	Marc Hogan		Des Moines	DesNoise	Mon, 02 Jan 2012 01:53:31
1090	RT @HawkeyeJosh: Mitt Romney just makes my skin crawl. #phony #corporate	lehi		Iowa	lehimesa	Mon, 02 Jan 2012 00:46:34
1091	Mitt Romney just makes my skin crawl. #phony #corporateshill #ihatemittromne	Josh		Iowa	HawkeyeJosh	Mon, 02 Jan 2012 00:34:21
1092	Five-plus Occupy protesters at Mitt Romney's headquarters on Ingersoll. They a	Nicole R. Paseka		Iowa	npaseka	Mon, 02 Jan 2012 00:13:30
1093	Mitt Romney thinks "it could be worse" is same as "let them eat cake". Trying to	LindaMcSchuler		USA, Iowa	LindaMcSchuler	Mon, 02 Jan 2012 00:12:06
1094	About five protestors at Mitt Romney headquarters #occupy #ocupycaucus http	Nicole R. Paseka		Iowa	npaseka	Mon, 02 Jan 2012 00:12:04

Figure 3. Two types of communication channels (media) collected by SWARMS.

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

Web Search Engines and Mapping the Web Pages

Most web search engines rely on web crawlers (or web robots) to collect and index web page content into a centralized database. For example, Google collects millions of web

page indexes in its search engine databases daily by deploying thousands of web crawlers from their servers. Web crawlers are dynamic network programs designed for collecting and duplicating targeted website contents (remotely) into web index databases. Each web crawler can switch its targeted websites by examining the hyperlinks found in the original web pages, often HyperText Markup Language (HTML) documents. Therefore, 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. For example, to determine the importance of web pages, Google developed its famous PageRank method, "*a global ranking of all web pages, regardless of their content, based solely on their location in the Web graph structure*" (Page et al. 1999, 15). PageRank relies on the external referred pages (other pages linked to the targeted web page) to calculate the ranks. For example, the SDSU web page will be more important if it was referred by two important web pages (the California State University System web page and the CNN.com/university web page). Another website will be ranked lower than the SDSU web page if it were referred by two less important websites (such as not-important.com/chats.htm and mypersonal-blogs.com/note.htm). The referring structure of web pages will determine their ranking numbers in the Google search engine.

Currently, the Google search engine combines PageRank with other content-based analysis methods to make the keyword webpage search in Google more accurate and more effective. However, one major limitation of Google search engine is the restriction of its application programming interfaces (APIs). Current Google search APIs can only be used to retrieve up to 64 hits from Google search engine each time. Therefore, the SWARMS prototype requires the use of Yahoo and Bing search engines APIs because they provide up to 1000 hits 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 out the geolocation of IP addresses associated with each web page. Most geolocation operations are performed by sending requests to a WHOIS database server. 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), such as American Registry for Internet Number (ARIN) or Asia-Pacific Network Information Centre (APNIC). 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".

Researchers can then convert the street address of the registrant into latitude and longitude using geocoding services provided by GIS software or Google Map APIs. For example, the street address, "*5500 Campanile Drive, San Diego*" can be converted to "*32.773131,-117.0766*" (decimal degree coordinates). When an IP address or a Domain Name 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. A proxy server acts as a connector between users and the actual websites. 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 percent successful conversion rate for all geolocation procedures.

Twitter APIs with Geolocation-based Search

Social media (such as Twitter, Flickr, and Facebook) are powerful communication platforms for idea exchange, breaking news, personal networking, political opinions, and collective actions. By using smart phones, 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 impacts of these tools were so vividly demonstrated in the most recent anti-government protests, including the Arab Spring, Occupy Wall Street, and the London Riots. 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 (Golder and Macy 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). 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).

Our research team developed a Twitter spatial search tool (a python program) to retrieve tweets by using keywords and by defining searchable spatial range. The search results were saved into Excel files containing detailed tweets information, including user names, user ID, tweet text, created_time, and spatial locations. Duplicated tweets were removed using the unique tweet_id that came with each tweet. The spatial locations of tweets were tagged by Twitter API automatically (by using enabled GPS in mobile devices or by user-defined home towns). We performed searches using candidates' full name since searching with only first name or last name often returned un-related tweets. For example, searching with "Rick" or "Paul" returned many tweets referring to some other person named Rick or Paul instead of the two candidates. Figure 4 illustrates an example of our Twitter spatial search configuration (within the radius of 84 miles from the city center of Miami with the keyword search "Mitt Romney" and within the radius of 45 miles from the city center of Concord).

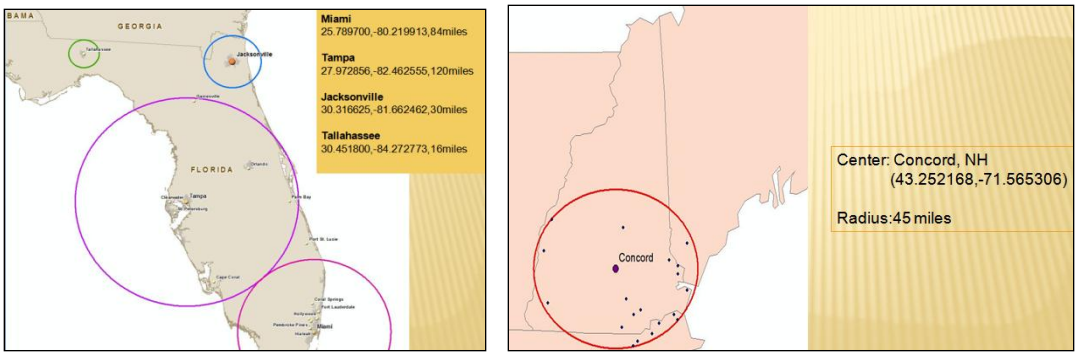

Figure 4. The pre-defined spatial ranges for tweet search in Florida and New Hampshire.

Our python spatial search tool can retrieve all tweets within a pre-defined spatial range (Figure 4). During the state primaries, we selected major cities with a population over 100,000 in the target states and set up the spatial range to cover major metropolitan areas without overlapping each other. Thus, the spatial range for each search center may vary based on the size of metropolitan area. However, there are two major limitations of our method: 1). the Twitter search tool can only retrieve tweets back to 7 days by using the Twitter search API; 2). if the search result exceeds 1,500 tweets, the Twitter search tool will only capture the latest 1,500 tweets. Figure 2 illustrates the pre-defined spatial search areas during the Florida election (four cities over 100,000 population) and the New Hampshire election (one city over 100,000 population).

Our research utilizes the spatial-defined query function provided by Twitter Application Programming Interface (API) to compare the popularity of four U.S. Republican Presidential Candidates (Mitt Romney, Rick Santorum, Newt Gingrich, and Ron Paul) during the 2012 Primary Election. The full name of each candidate was used as "keywords" to search for tweets during the election period within the major city boundaries in the targeted states. Then we compared the percentages of tweets collected by each candidate (as the popularity index) to the actual election results. Our preliminary correlation analysis indicates that there is a strong correlation between the popularity of each candidate in Twitter (tweets created on one day before the election and tweets created on the actual election day) and the final election results.

Visualizing 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. But 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 associated with the actual geolocation of SDSU server's IP address, which is registered as "5500 Campanile Drive, San Diego, California". There will be many exceptions and inconsistence between the geolocations of IP addresses and the actual spatial content of the web pages. But the geographic probability of their spatial locations related to the spatial contents will be

much higher in the center of IP geolocations. 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 comparing 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 out that higher density areas of web page IP geolocations are associated with major U.S. cities with bigger population, such as New York, Los Angeles, and Houston. This indicates that the density (or geographic probability) of web pages may be closely related to the size of city populations. Bigger cities will have a higher density of web pages (or higher probability of hosting web pages) associated with selected keywords.

Our next step is to calculate the differences between two different keyword maps, such as "Mitt Romney" versus "Newt Gingrich". 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)

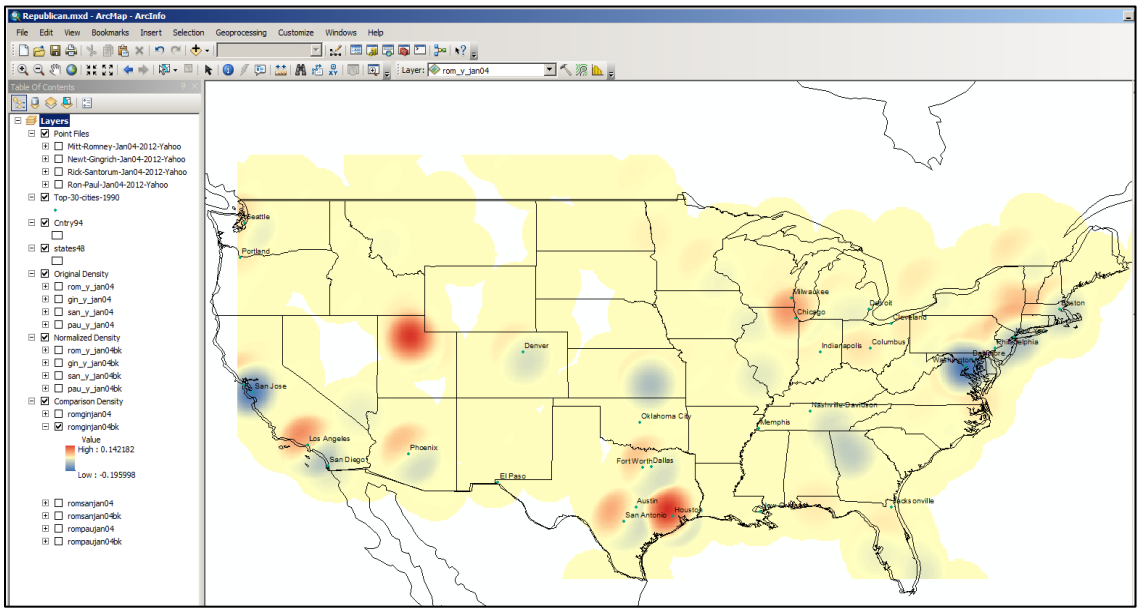

Figure 5. Differential value map of web page information landscape between "Mitt Romney" and "Newt Gingrich" keywords from Yahoo Engine on January 04, 2012.

Figure 5 illustrates the Web Page information landscape with the differential value between "Mitt Romney" and "Newt Gingrich". The red color areas have relative higher probability of hosting "Mitt Romney related web pages comparing to the probability of hosting "Newt Gingrich" web pages based on their web server IP addresses. The blue

color areas have relative higher probability of hosting "New Gingrich" web pages comparing to the probability of hosting "Mitt Romney" web pages.

The color patterns in the differential value map illustrated some interesting "signals" or "geospatial fingerprints" about the two keywords ("Mitt Romney" and "Newt Gingrich"). Salt Lake City in Utah has very high probability of hosting "Mitt Romney" web pages due to his previous political connection to the City and his religious preferences. "Romney" as former Massachusetts governor, may also cause the red color zones near the central New England areas. On the other hand, the state of Georgia shows the higher probability of hosting web pages related to "Newt Gingrich" due to his political connection (as former representative from Georgia). However, there are some patterns which are difficult to explain, such as the blue areas in San Francisco and the red zones in Houston.

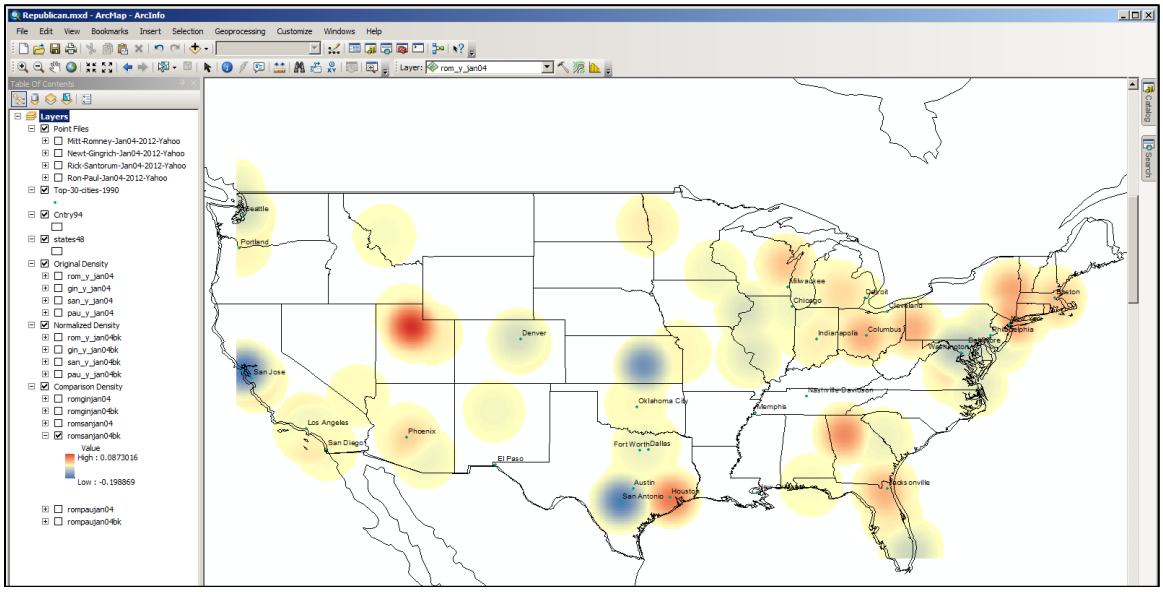

Figure 6. Differential value map of web page information landscape between "Mitt Romney" and "Rick Santorum" keywords from Yahoo Engine on January 04, 2012.

Figure 6 illustrates the Web Page information landscape with the differential value between "Mitt Romney" and "Rick Santorum". The red color areas have relative higher probability of hosting "Mitt Romney related web pages comparing to the probability of hosting "Rick Santorum" web pages based on their web server IP addresses. The blue color areas have relative higher probability of hosting "Rick Santorum" web pages comparing to the probability of hosting "Mitt Romney" web pages.

Similar to the previous map, we noticed that the Salt Lake City and New England regions still have higher probability of hosting "Mitt Romney" web pages (red zones). On the other hand, San Antonio and the State of Kansas has a higher probability of hosting "Rick Santorum" web pages.

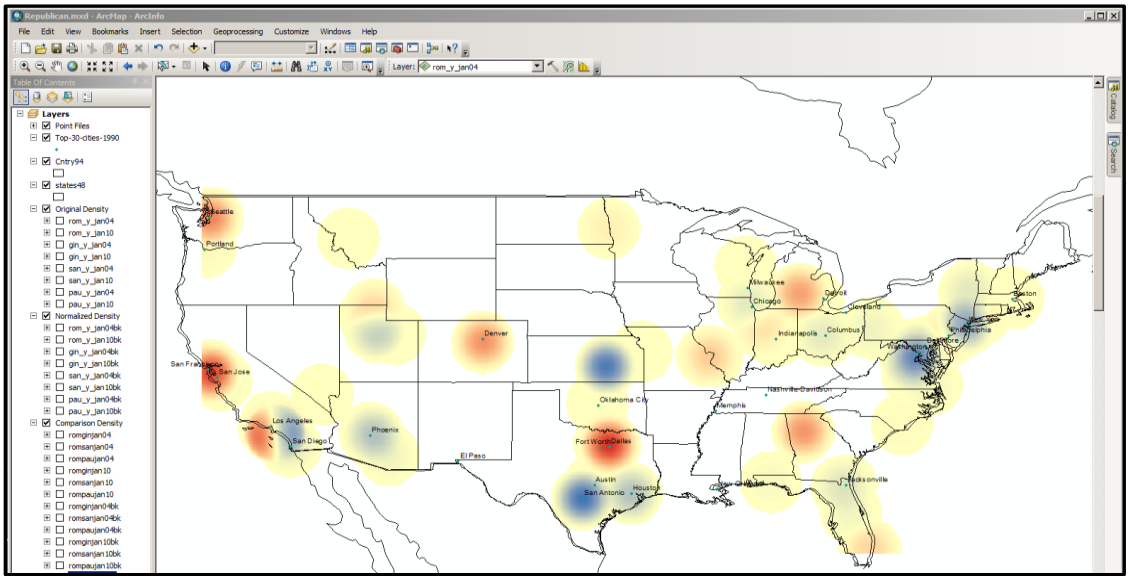

Figure 7. Differential value map of web page information landscape of "Mitt Romney" between January 04, 2012 and January 10, 2012 (one week later) from Yahoo Search Engine.

Figure 7 illustrate the temporal change of web page information landscapes within one week using the same keyword, "Mitt Romney". The red color areas indicated the increased probability from January 04, 2012 to January 10, 2012 (regarding the probability of hosting "Mitt Romney" in the area). The blue color areas indicated the decreased probability from January 04, 2012 to January 10, 2012. The visualization patterns may reveal some interesting findings as the following:

- The probability increased zones, such as Dallas, TX; Denver, CO; Atlanta, GA; West Coast, may reflect the winning of Primary election by Mitt Romney in Iowa caucus.
- The decreased areas (blue color), such as southern Texas and Phoenix may indicate the increased presence of Rick Santorum and Newt Gingrich (Mitt Romney's opponents).

Geolocation-Based Tweet Content Analysis

In addition to analyze the web page information landscape of "Mitt Romney". Our project started the pre-defined spatial search for tweets on Jan 03, 2012 (Iowa Caucus) and finished the last search task on April 04, 2012 (Wisconsin Primary) when Rick Santorum dropped out of the presidential race (Mitt Romney became the only candidate for the Republican Party). During the three months, we collected 81,898 tweets using the keywords of the four candidate names and generated 1,050 excel files. Each file contains tweets by using one candidate's full name search. Tweets were filtered by their created time stamps and assigned to three categories: *Tweets_two_days_before_election*, *Tweets_one_day_before_election*, and *Tweets_on_the_election_day*. Figure 8 illustrated the tweet search results in three states (Ohio, Georgia, and Oklahoma). The green color

bars indicated the percentage of tweets (created on one day before election) mentioned each candidate's name. The blue color bars indicated the percentage of tweets (created on the day of election) mentioned each candidate. The red color bars show the actual election results for each candidate (their voting/supporting percentages).

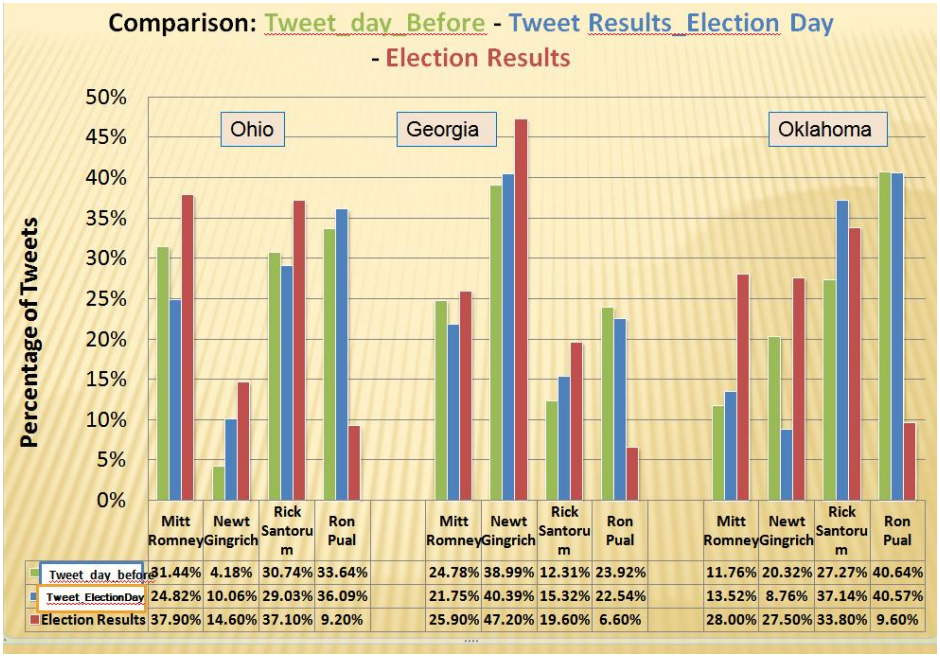

Figure 8. Tweet search results (two days) with four Republican Presidential Candidates in three states (Ohio, Georgia, and Oklahoma) comparing to the actual election results.

This research also use regional poll data and the actual election results from the RCP website (*www.realclearpolitics.com*) for comparison purposes during each primary and caucus. RCP is a well-known popular political website providing aggregated regional polls and political news. We used the RCP poll data in each state to compare to our tweet search results. Some states have over 400 tweets per day for each candidate. But a few states only have very few tweets for each candidate due to the smaller population. Therefore, we only selected 19 states where the major candidates have at least 20 tweets in the three-day timeframe for our correlation analysis. A total of 41,941 tweets from the three categories mentioned above were calculated into the percentage ratio between each candidate and to compare with RCP poll data and election results

We composed scripts in R, a free statistical computing environment, to calculate correlation coefficient values between the following five variables: *Tweets_two_days_before_election%*, *Tweets_one day_before_election%*, *Tweets_on_the_election_day%*, *Regional_poll% (from RCP),* and *Election_results% (from RCP).* When including all four candidates, correlation coefficient (R) between the **Tweets_one day_before_election%** and **Election_results%** is **0.56**. The coefficient value becomes 0.59 between Tweets_on_the_election_day% and Election_results%. We also noticed that Ron Paul is much more popular in tweets comparing to the poll or

election results. If we remove the Ron Paul from the correlation analysis, the correlation results improve to **0.75** (one day before%) and **0.86** (the election day%) respectively.

Summary and Future Research

This research demonstrated a research framework for tracking and analyzing the spatial content of social media (Twitter) and web pages, visualized the dynamic comparison of web information landscapes, and examined the correlation between the popularity of candidates on Twitter and the actual election results. Our next step is to perform further comprehensive sentiment analysis of tweets and web pages to identify the "pro" and "con" aspects for each candidate (supporters vs. opposers).

By tracking and analysing the contents of tweets and web pages, researchers might be able to reveal important social contexts of specific events (such as presidential elections or protests) and understand the temporal and spatial relationships among these short messages and human behaviours. In the fields of biology and physics, scientists can analyze massive amounts of data and information from scientific observations and measurements. Today 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.

Acknowledgements

This material 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 material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

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, Australia, 14 –18 April, 107-117.

Golder S.A., Macy M.W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333: 1878–1881.

Hägerstrand, T. (1967) Innovation Diffusion as a Spatial Process. The University of Chicago Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Baraba´si, A. L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science*, 323, 721-723.

Miller G. 2011. Social Scientists wade into the Tweet stream. *Science* 333: 1814–1815.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: Bringing order to the Web. http://ilpubs.stanford.edu:8090/422/ (last accessed 28 March 2011).

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.

Svantesson, D. J. B. 2005 Geo-identification: Now they know where you live. *Privacy Law & Policy Reporter* 11(6): 171-74.

Twitter. 2012. *Twitter turns six* [online]. Twitter Blog. Available from: http://blog.twitter.com/2012/03/twitter-turns-six.html [Accessed May 1 2012].

Ming-Hsiang Tsou, Professor, Department of Geography, San Diego State University, CA 92182-4493. Email <mtsou@mail.sdsu.edu >