GeoJournal (2013) 78:319–338
DOI 10.1007/s10708-011-9438-2

Harvesting ambient geospatial information from social media feeds

Anthony Stefanidis · Andrew Crooks · Jacek Radzikowski

Published online: 4 December 2011

Abstract Social media generated from many individuals is playing a greater role in our daily lives and provides a unique opportunity to gain valuable insight on information flow and social networking within a society. Through data collection and analysis of its content, it supports a greater mapping and understanding of the evolving human landscape. The information disseminated through such media represents a deviation from volunteered geography, in the sense that it is not geographic information per se. Nevertheless, the message often has geographic footprints, for example, in the form of locations from where the tweets originate, or references in their content to geographic entities. We argue that such data conveys ambient geospatial information, capturing for example, people's references to locations that represent momentary social hotspots. In this paper we address a framework to harvest such ambient geospatial information, and resulting hybrid capabilities to analyze it to support situational awareness as it relates to human activities. We argue that this emergence of ambient geospatial analysis represents a second step in the evolution of geospatial data availability, following on the heels of volunteered geographical information.

Keywords Social media · Social network analysis · Volunteered geographic information · Ambient intelligence

A. Stefanidis (✉) · J. Radzikowski
Center for Geospatial Intelligence and Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22030, USA
e-mail: astefani@gmu.edu

A. Crooks
Department of Computational Social Science, George Mason University, Fairfax, VA 22030, USA

Introduction

The recent civil unrest events of the Arab Spring, spreading across North Africa and the Middle East in the first months of 2011, confirmed the unprecedented power of social media to communicate information within these societies, and from them to the outside world. This comes only 20 months after the early, experimental use of *twitter*, *facebook,* and *YouTube* in June 2009 to provide real-time accounts of the situation in the streets of Teheran by disseminating images, video, and news (Newsweek 2009), social media were again at the forefront of information transmission. The information disseminated through such media represents a deviation from Goodchild's (2007a) notion of *volunteered geography*, in the sense that it is not geographic information per se. Unlike *Wikimapia* or *OpenStreetMap*, social media feeds do not aim to empower citizens to create a patchwork of geographic information: geography is not their message. Nevertheless, the message has geographic

footprints, for example, in the form of locations from where the tweets originate, or references in their content to geographic entities (e.g. the numerous references to Tahrir Square during the Egyptian revolution). Accordingly, we argue that such data conveys *ambient geospatial information*, capturing for example, people's references to locations that represent momentary social hotspots. Harvesting this ambient geospatial information provides a unique opportunity to gain valuable insight on information flow and social networking within a society, and may even support a greater mapping and understanding of the human landscape and its evolution over time. In this paper we address a framework to harvest such ambient geospatial information, and resulting hybrid capabilities to analyze it.

This paper addresses the emergence of new analysis techniques, and resulting hybrid capabilities that take advantage of ambient geographical information (AGI) to support situational awareness as it relates to human activities. We argue that this emergence of *ambient geospatial analysis* represents a second step in the evolution of geospatial data availability, following on the heels of volunteered geographical information (VGI).

The paper is organized as follows. In "Tracing the rise of ambient geospatial information" we trace the rise of ambient geospatial information following the evolution of Web 2.0 technologies and the emergence of social media. In "System architecture for harvesting information from social media feeds" we present a general framework/architecture for collecting ambient information from social media. "Case studies: turning ambient geospatial data into knowledge" presents case studies and novel hybrid types of geospatial analysis that can be performed using ambient geospatial information in a non-technical way along with utilizing social network analysis techniques. In "Discussion and outlook" we offer our outlook assessment.

Tracing the rise of ambient geospatial information

Much of what is now possible with respect to social media feeds relates to the growth and evolution of Web 2.0 technologies. In this section we present the defining characteristics of Web 2.0 and its relation to geospatial information gathering and dissemination. The term Web 2.0 can be traced back to O'Reilly Media in 2004, who used it to define web applications that facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the World Wide Web. Utilizing technologies of social networking, social booking marking, blogging, Wikis and RSS/XML feeds (Graham 2007). Web 2.0 can be defined by six often overlapping concepts: (1) individual production and user-generated content, (2) harnessing the power of the crowd (e.g. crowdsourcing, see Howe 2006), (3) data on a massive scale, (4) participation-enabling architectures, (5) ubiquitous networking, and finally, (6) openness and transparency (see O'Reilly 2005; Anderson 2007; Batty et al. 2010; for further discussions). Examples of such Web 2.0 applications include *MySpace*, *facebook*, *flickr*, *YouTube*, and *Wikipedia*. The growth of Web 2.0 technologies relies heavily on our ability to communicate and share data and information through simple, freely available tools, in contrast to static websites and data repositories of the past. The aim of Web 2.0 tools are that they can be learnt quickly and effectively without immersion in professional activities (see Hudson-Smith et al. 2009a) such as advanced computer programming skills. Some are describing this change in as the cult of the amateur with respect to information gathering and content sharing (see Keen 2007).

With relation to spatial data, Web 2.0 has led to a renaissance of geographic information (Hudson-Smith and Crooks 2009). This renaissance was fueled by the immense popularity of tools like Google Maps which made place matter, and Google Maps Application Programming Interface (API) which allows practically anyone to create mashups (see Haklay et al. 2008; for more information and discussion), and also by the growth in use of the geobrowsers (e.g. Google Earth, NASA's World Wind). This renaissance put renewed focus on early work to exploit geographical information present in web pages to support various web queries (e.g. Buyukkokten et al. 1999; Gravano et al. 2003).

Considering the particularities of geospatial content as it relates to the above-mentioned six defining themes of Web 2.0, let us consider individual production and user-generated content, which also results in massive amounts of data. In the past, the production and collection of geospatial data (either primary or secondary) was often the first and most vital task of any geographical information system (GIS) project, with data capture costs often accounting for up to 85% of the cost of a GIS (Longley et al. 2010). This has been tipped on its head through crowdsourcing and VGI.

Representative examples include the post-earthquake mapping of Haiti in 2010 (e.g. Norheim-Hagtun and Meier 2010; Zook et al. 2010), the Christmas Bird Count (National Audubon Society 2011) via dedicated services (such as *OpenStreetMap* or *Google Map Maker*). *OpenStreetMap* and *Google Map Maker* also are perfect examples of participation-enabling architectures that support contributions by both domain experts and amateurs alike. While data collection is still an important aspect, some would argue that harnessing the power of the crowd reduces the burden of data collection. Authors have already started to assess the quality of VGI like *OpenStreetMap*, by comparing it to established authoritative mapping organizations such as the United Kingdoms Ordnance Survey (Haklay 2010). One could consider VGI '*good enough*' for its purpose, especially in situations where it presents the only reasonable means to collect information in a timely manner. In addition, some would argue that *OpenStreetMap* like *Wikipedia*, is a process of evolving a good product, not a complete product in itself because there is no end goal in sight as to what constitutes the best map (or the best entry in the case of Wikipedia, see Hudson-Smith et al. 2009b). Moreover, the internet is becoming more portable and there has been a considerable rise in location-aware devices like smartphones, GPS-enabled cameras, and tablets for data generation. As such data collection is dependent on increased access to the Internet, digital divide issues (the 'haves' and 'have not') still remain important, leading for example, to variations in developed versus less-developed countries access to information therefore contribution (see: Longley et al. 2006; Buys et al. 2009). The transition from desktop to a shared and distributed paradigm for data access and contribution allows greater openness and transparency, from top-down government-led efforts such as http://www.whitehouse.gov/open to more bottom-up initiatives such as http://geocommons.com/, both being key components of the emerging Geospatial Web (see Elwood 2010).

All these developments have supported over the past few years the emergence and growth of *volunteered geography*, with citizens as sensors, actively collecting and contributing geospatial information (Goodchild 2007b) utilizing Web 2.0 technologies and advances and reduction in terms of cost of data collection mechanisms (such as GPS enabled devices). Just as Web 2.0 has changed how we interact and share information on the Web, so too has web mapping evolved during the last decade, from viewing static data (such as MapQuest) to more dynamic sites with user generated content. This can be correlated with the emergence of relevant sites having well-defined APIs (such as Google's My Maps API). Coinciding with sites having APIs are technologies allowing for distributed GIS data collection, from smartphones with GPS to sites with digitization features such as *Google Map Maker*. Such sites allow people to collect and disseminate geospatial information while bypassing traditional GIS software. Furthermore, through APIs users can create bespoke applications to serve such data, e.g. through web mashups, which have seen substantial increases since 2005, as Google allowed users to access its Google Maps API. However, the analysis capabilities of such tools are often limited. One could consider this to be a legacy of GIS education, in the sense that people often consider GIS as just maps and map displaying, and not the underlying techniques to build such maps. But it also revolves around the purpose of many map mashups: to display data and not to manipulate it. Another barrier to carrying out spatial analysis is of course access to dedicated geographical information software (such as ArcGIS or MapInfo), which was traditionally limited to experts rather than the public at large. This latter concern is however changing through the development of opensource geographical information-related software (such as QGIS, and R), enabling people to manipulate and analyze data, just as OpenOffice allows people to use word and data packages.

However, the emergence of social media participation and information sharing is bringing forward a different model of geospatial information contribution, while its still true that users actively contribute geographic data. There is also a model where the users' intent was not to directly contribute geospatial data (e.g. a map), but rather to contribute information (e.g. a geotagged picture from a family vacation, or text describing a planned event) that happens to have an associated geospatial component and a corresponding footprint. Harvesting and analyzing such ambient information represents a substantial challenge needing new skillsets as it resides at the intersection of disciplines like geography, computational social sciences, linguistics, and computer science. Nevertheless, it can provide us with unparalleled insight on a broad variety of cultural, societal, and human factors, particularly as they relate to human and social dynamics, for example:

- mapping the manner in which ideas and information propagate in a society, information that can be used to identify appropriate strategies for information dissemination during a crisis situation;
- mapping people's opinions and reaction on specific topics and current events, thus improving our ability to collect precise cultural, political, economic and health data, and to do so at near real-time rates; and
- identifying emerging socio-cultural hotspots.

In the next section we present a general architecture for harvesting ambient geospatial information from social media feeds which could be perceived as a merging of crowd-sourcing, VGI, and social media sourcing.

System architecture for harvesting information from social media feeds

In this section we present a brief outline of a general architecture of a system for harvesting information from multiple social media feeds (not just for creating mashups but for detailed analysis). In the context of this paper the objective is not to introduce a novel architecture per se, but rather to highlight its components, and implicitly its operations in order to better communicate how such data can be harvested. Interested readers are referred to the rather extensive web mining literature (see: Bowman et al. 1995; Cooley et al. 1997; Kosala and Blockeel 2000) or more recent references focusing on Social Web in particular (e.g. Russell 2011a, b) for an overview of relevant solutions. A representative architecture of such a system is shown in Fig. 1 and presented in more detail below. It entails three components: extracting data from the data providers (various social media servers) via APIs; parsing, integrating, and storing these data it in a resident database; and then analyzing these data to extract information of interest.[1]

[1] While we present a general architecture upon which we based our own system to collect such information, we should note that there also exists a number of comparable tools such as 140kit (http://140kit.com/), or twapperkeeper (http://twapperkeeper.com/), but these are limited in their scalability with respect to large datasets. Sites such as ushahidi (http://www.ushahidi.com/) also provide a means to collect and disseminate information over the web. However, there are very few tools that allow one to add context to content, or to support detailed analysis.

Original social media feeds can be retrieved from the source data provider through queries. This entails submitting a query in the form of an http request and receiving in response data in XML format (e.g. *Atom* or *RSS*). The query parameters may be for example, based on location (e.g. specifying an area of interest to which the feed is related), time (e.g. specifying a period of interest), content (e.g. specifying keywords), or even by user handle/ID. In response to these queries, and depending on the characteristics of the information provided by the service, we can receive from the server just *metadata* or *metadata and actual data*. A representative example of the first case is *flickr*, where the query result contains exclusively metadata information (e.g. author, time, and geolocation when available), and information on how to access the actual image itself. A representative example of the second is *twitter*, where the data received in response to a query are actual tweets and associated metadata (e.g. user information, time of tweet publication, geolocation when available, and information on whether this particular tweet is in response to or retweet of an earlier message).

Once this information is harvested from the social media server it can be parsed to become part of a *local database*, mirroring the content of the server provider for the specific entries that were returned by our query. Data parsed from diverse sources are integrated by SMI, a *Social Media Ingestor*, capturing information that is common across diverse sources (e.g. time of submission, user name, originating location, keywords), as well as service-specific information (e.g. content, links to actual files). This allows us to establish an integrated multi-source local database that can be used to perform analysis of the harvested data that is not supported by the provider database interface (e.g. statistics on user activities) for various projects.

The storage and management system for the local database is built around the notion of an information *entry*. In our approach, an *entry* is a local database record (and accompanying attributes) that corresponds to a single piece of information received from a social media provider in response to our above-mentioned queries. Thus, an entry is a local record corresponding to one social media entry. Its attributes comprise generalized, source-independent properties (e.g. authorship, publication time, location) as well as source-specific data (e.g. links to the actual media file) as shown in Fig. 2.

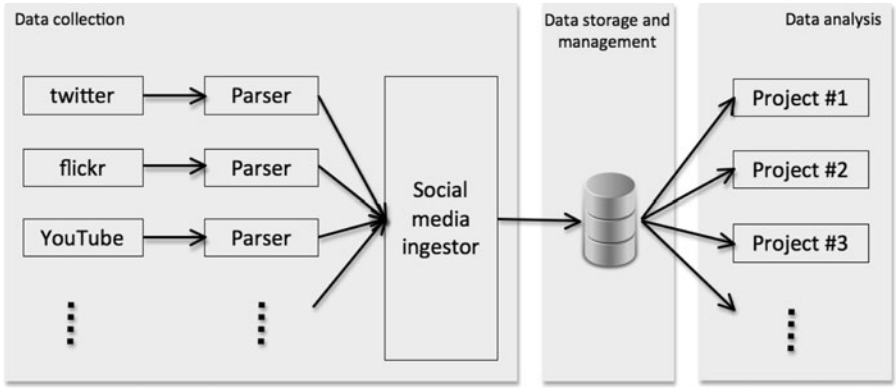


Fig. 1 General architecture of a system to harvest information from social media sites

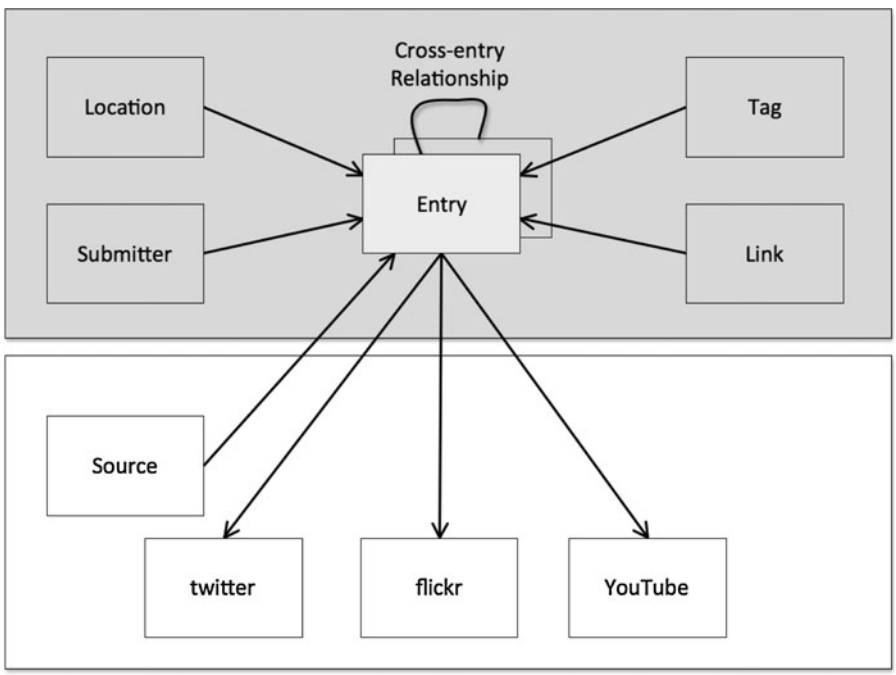


Fig. 2 Entry in the resident database. *The top gray* shaded area indicates source-independent attributes, while *the lower white* area is source-specific content

Cross-entry links are of particular importance, as they reveal semantic relations between entries. Within a particular service such links may become available from the original user community (e.g. retweets or responses in *twitter*). Additional links can be detected through analysis of the local database entries (e.g. linking entries that refer to the same real-life event, or ones that have comparable content, or using tags). By performing this analysis in our local database we can identify cross-medium links (e.g. linking *twitter* entries with *YouTube* videos and *flickr* images), which allows us to span the boundaries of source services. Relations between entries are valuable because they reveal relations between the submitters of these entries, allowing us to identify the structure of the underlying social network. The most important relations that form the network are forwarding entries written by other members of the social site, replying to messages, or mentioning of other entries. These are not only very strong indicators of the relations between the submitters, but also indicate information pathways, allowing us to recognize the links through which information is disseminated within different groups, and to identify original sources of the information, and social hubs which disseminate it within their networks.

The framework presented here has been implemented using PostgresSQL as a database. In our

prototype we issue http queries to the source social media sites, using their own query API. In our queries we specify certain content parameters (e.g. area of interest). As a response we receive XML files which are parsed to extract content of interest, and subsequently through SMI are inserted in the PostgresSQL database. For practical purposes such queries are non-continuous, but are rather issued periodically, depending on the source traffic and regulations (e.g. issued every 5 min for high traffic feeds, or less frequently for lower traffic). While the information harvested from social media in this manner is not explicitly geospatial, it does include implicit geospatial content, thus rendering it suitable for novel types of geospatial analysis as we present in the following section.

Case studies: turning ambient geospatial data into knowledge

The analytical spectrum of geospatial information harvested through the above-described processes ranges from the social to the geographical space. Feeds can be analyzed to identify geospatial information references within a dataset (for example, using entries in a gazetteer), or the locations of individuals who contribute information, and their social network links. Below we demonstrate two representative analytical operations:

- geospatial hotspot emergence, by monitoring variations over time of references to gazetteer entries ("Hotspot emergence"), and
- tracing information dissemination routes in an area of interest, and through it identifying and mapping local social networks in an area ("Tracing information dissemination and social networks").

While we use primarily *twitter* data to demonstrate these capabilities (the massive amounts and rapid updates of contributions make it more interesting for analysis), we can use the same techniques with any other social media feeds from among the ones collected by our prototype system to identify social networks, the locations of their members, and temporal variations of this information.

Hotspot emergence

There has been a growing interest in using social media to track emerging trends, interest over specific events and daily activity patterns. Many of the applications focus on predicting trends such as forecasting box-office revenues for movies using social media (see Asur and Huberman 2010). Only recently have people started to take an interest in the geographical aspects of such trends. Sample demonstrations, such as the recent work of the Centre for Advanced Spatial Analysis at University College London and their new city landscapes or Tweetgeography[2] which data mines *twitter* data on a number of cities from around the world (e.g. London, New York and Paris) and maps the density of tweets to specific areas. While others have identified hotspots of activity such as around train stations or coffee shops with *Foursquare* data (Wall Street Journal 2011). This is a deviation from traditional social network analysis of such data (for example, who knows whom, who reads whom etc.), and shows the growing interest in location with respect to social media (something to which we will return in "Discussion and outlook").

Using an in-house implementation of the system presented in "System architecture for harvesting information from social media feeds" we collected tweets that included specific references to locations selected from a Gazetteer. In Fig. 3 we show for example, an analysis performed by our team on *twitter* feeds originating from Cairo (based on *twitter's* API location filtering) over the period 2/25/11–3/09/11, to identify ones that were labeled with the hashtag[3] '*Tahrir Square*'. The data are shown on the upper right-hand side of the figure: a chart shows the number of tweets per hour within a 10 km radius from Tahrir square (vertical axis) over the period 2/25/11–3/09/11 (horizontal axis). Tweets are grouped hourly, and we can identify a peak (marked by a circle) corresponding to the period 07:00–08:00 local time (UTC + 2) of 3/4/11. This is actually the morning of the day when the new prime minister of Egypt eventually addressed the people at the square. Such peaks can be identified using variants of tf-idf scores (Salton and McGill 1983) to recognize abnormally high references to

[2] Readers are referred to http://www.casa.ucl.ac.uk/tom/ and http://urbantick.blogspot.com/ for further information.

[3] Hashtags represent a bottom up, user-generated convention for adding content (in a sense, metadata) about a specific topic, by identifying keywords to describe content. Thus it allows easy searching of tweets and trends. Sites such as http://hashtags.org/ monitor such trends from tweets and provide relevant statistics, but only over short periods of times.

Fig. 3 *Twitter* data with a *Tahrir Square* label, 2/25/11–3/09/11, overlaid on a map of the square. Hashtag statistics over time show how off the spike (*circled*) is from normal

specific terms (Naaman et al. 2011). The data shown correspond to approximately 38,000 tweets with a *Tahrir Square* hashtag in that 2 week period, selected from among a total of 684,000 tweets from 40,000 persons with a Cairo label over that period.

The example presented above demonstrates one manner in which geographical hotspots emerge through such analysis: as footprints of landmark events. Such landmark events dominate social media reports before (when they are planned), during, or shortly after their occurrence, thus resulting in statistically significant deviations in the number of references to that specific location (manifested as spikes in the corresponding frequency table). This allows us to label such a location as a hotspot. In this specific situation we have advance notice of this event, as people started discussing it as soon as it became known. Depending on the nature of the event these spikes may be concurrent with (as we will see in Fig. 4 below), or follow the actual event. Thus this analysis can support timely resource allocation as needed to better monitor emerging hotspots.

While in the earlier example we focused on a single location and variations on references to it over time, the analysis can also be performed as a comparative study of multiple locations, as we show in Fig. 4. Here we show the relative traffic amount in social media sites with references to four different locations in Libya on 3/6/11 between 18:00 and 19:00 UTC time (Libya local time is UTC+2) at the height of the local civil unrest. We focus on 4 cities that were major theaters of conflict over that period, with major offensives by the Government attacking the rebels. Figure 4 is actually a snapshot from a video we created showing the variations in social media traffic for these 4 locations over a period of two weeks. For the specific instance depicted in Fig. 4, we show how references to Tripoli lead, followed by Benghazi, Tubruq, and Al-Zawiyah.[4] While this instantaneous analysis shows how Tripoli leads at that moment, by comparing the data to earlier traffic patterns we can identify that Al-Zawiyah shows a threefold increase on references compared to its past records, thus identifying it as an emerging hotspot from among the ones compared. As

[4] The tweets were gained by searching using the *twitter* API within a 30 km radius of the given city or using their *twitter* profile location.

Fig. 4 Social media traffic volume with references to four different locations in Libya. From *left* to *right* we have the cities of Al-Zawiyah, Tripoli, Benghazi and Tubruq. The data reflect social media traffic on the 3/6/11 between the hours of 18:00 and 19:00 UTC

a reference we should mention here that the period between 3/5/11 and 3/8/11 was the peak of the government offensive on Al-Zawiyah, with intense battles raging in the center of the town, and numerous casualties (see the Guardian 2011).

In an effort to assess *twitter*'s role as a news dissemination mechanism, Kwak et al. (2010) performed a comparison of *twitter* topics to Google Trends and CNN Headlines and their preliminary results appear to confirm the role of *twitter* as a news breaking mechanism. In the same broader research direction, Becker et al. (2011) present an approach to distinguish messages that convey real world event information from regular twitter traffic, while Sankaranarayanan et al. (2009) presented an approach to extract news content from noisy twitter feeds guided by a small sample of manually identified seeders which are in essence news-oriented tweeters, and Weng and Lee (2011) presented an approach based on clustering of wavelet-based signals. The breaking of new events (or first story detection) in *twitter* streams has been addressed by Petrovic et al. (2010) through the use of algorithms based on locality-sensitive hashing, and this can lead to the detection of news sources in *twitter* groups. Furthermore, the micro-blogging nature of *twitter* makes it particularly suitable for reporting real-time events, transforming humans to sensors. For example, Sakaki et al. (2010), presented a system to detect earthquakes in Japan through the aggregation of tweets from various users serving as social sensors. They demonstrated how this information can be used successfully for earthquake detection and reporting through the use of Kalman and particle filters.

While in the above presented figures we have focused on tweets, one also can harvest images and video files from sites like *flickr* and *YouTube* (via their APIs). Relating this to Tahrir Square, Fig. 5 shows *flickr* pictures geotagged in the square, while Fig. 6 shows geotagged *YouTube* movies. Through comparable analysis we can detect spikes in these datasets, reflecting increased interest due to various events. Considering the additional advantage of 2-D and 3-D visualization offered by these platforms, the combination of this information with microblogging provides enhanced situational awareness for the areas of interest. By applying the above-presented analysis to a number of areas of interest, identified e.g. from a gazetteer, we can monitor the emergence of hotspots among these multimedia data collections.

These data collections offer a unique analysis potential, as they comprise media and annotation content. While analyzing the media portion of the collection has been traditionally addressed in the image analysis community, as a query-and-match problem (finding buildings in a database that look like the one in an image; see fro example, Zhang and Kosecka 2006; Schindler et al. 2007). Taking advantage of the annotation content of these datasets provides new analysis opportunities, ranging from

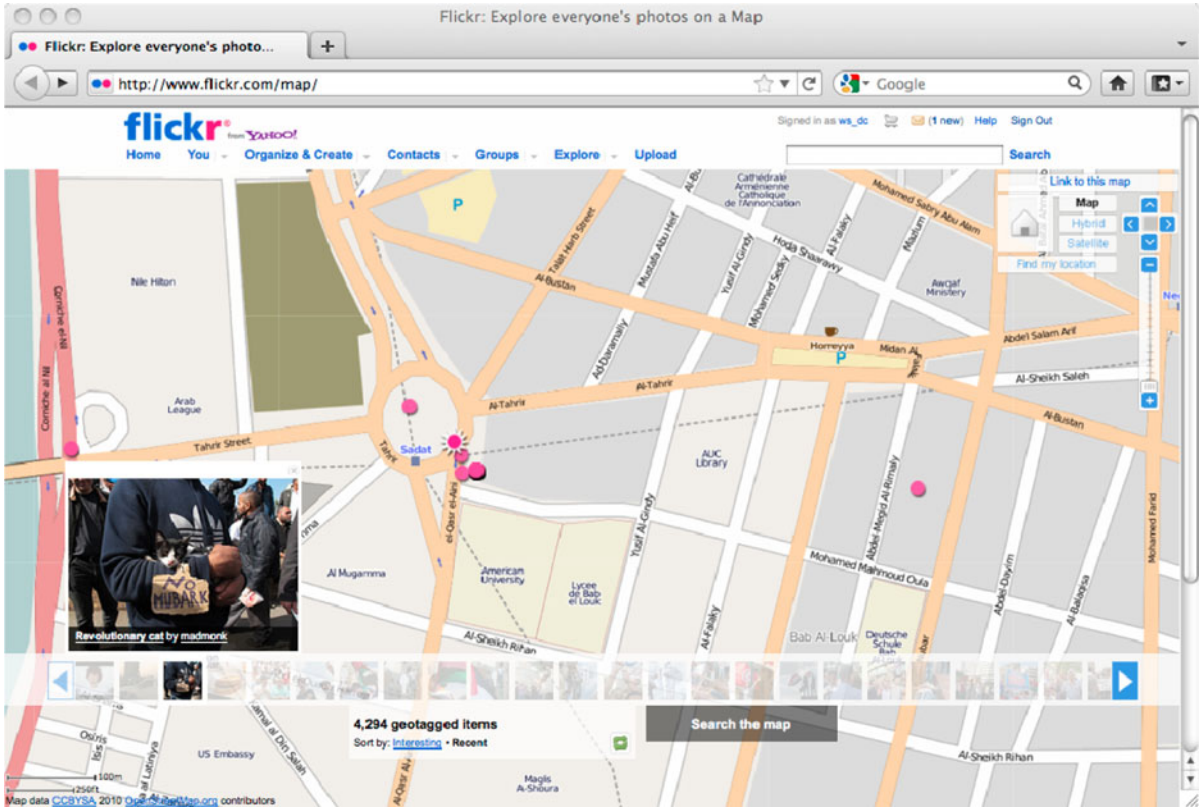


Fig. 5 *Circles* represent a selection of geotagged images for Tahrir Square, Cairo from *flickr*

improved geolocation solutions (Firedland et al. 2011) to extracting place semantics (Rattenbury and Naaman 2009), generating representative tags for different locations around the world (Kennedy et al. 2007), and even identifying events and corresponding social media documents (Becker et al. 2010). This potentially leads to improved understanding of such annotated multimedia user-contributed collections.

Tracing information dissemination and social networks

While in the previous section we addressed references to physical locations in order to detect the emergence of geospatial hotspots, the analysis of social media feeds provides us with a unique capability to understand the human landscape in unprecedented temporal resolution and spatial detail. Of particular importance to us is the tracing of the manner in which information is disseminated among various groups, and the formation of social networks.

Coincident with the growth of VGI and crisis mapping (see Meier 2009; Biewald and Janah 2010; Parry 2011) we have seen the growth of data mining techniques to explore events. For example, there has been a number of studies using data mining techniques to trawl through traditional media such as news articles (Brownstein et al. 2008) or Internet search engines (e.g. Polgreen et al. 2008), and blogs (Corley et al. 2010) to explore disease outbreaks. Approaches also carry out geovisual analytics to support crisis management (see MacEachren et al. 2011 and SensePlace2[5]) often through mashups of geotagged information.[6] With the advent of micro-blogging the focus has also moved to using *twitter* messages to forecast influenza rates (Culotta 2010) or swine flu pandemics (Ritterman et al. 2009). However, one can use such

[5] http://www.geovista.psu.edu/SensePlace2/.

[6] Examples include: monitoring swine flu (http://compepi.cs.uiowa.edu/~alessio/twitter-monitor-swine-flu/) or twitter traffic and the Oscars (http://www.neoformix.com/2009/OscarTwitterMap.html).

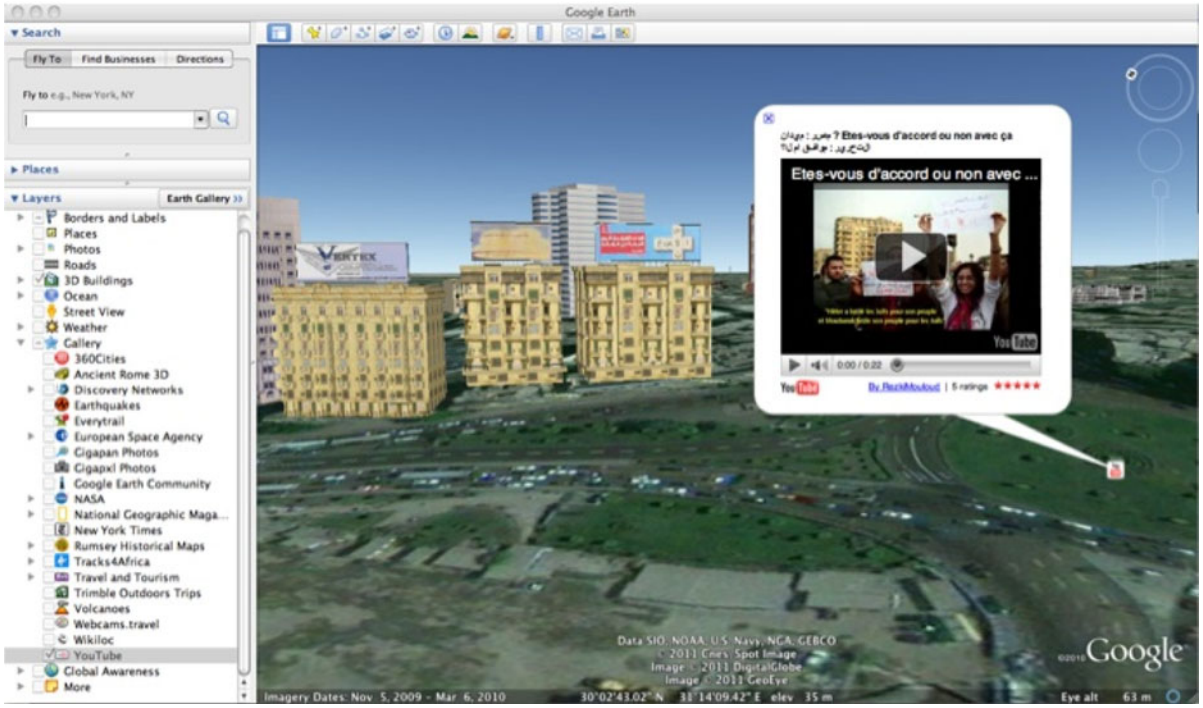


Fig. 6 Geotagged *YouTube* movies for Tahrir Square, embedded within Google Earth

information to also look at other trends and, more importantly for the scope of this paper, to gain information via social network analysis[7] about the social network structure: who is connected to whom, either directly or via common links, and how persons are clustered in groups sharing common interests. Critical in this analysis is the identification of information dissemination routes, by recognizing major nodes disseminating specific types of information, and their followers. Moreover, if one has locational information pertaining to these data one can geolocate (i.e. geotag) this information and thus map it out over an area of interest. This can be particularly important in crisis situations, supporting management and response, extending it beyond. Using our prototype system we started collecting *twitter* data relating to the devastating Sendai (Tohoku) earthquake in Japan (3/11/11), and we present here how this information can be analyzed to collect valuable social network and human landscape information.

Clusters of users manifest themselves through retweets and direct references. We collected *twitter* activity patterns within a 30 km radius around the center of Tokyo Metropolitan area, and present information dissemination patterns before and after a major event in Fig. 7. On the top we show the information dissemination pattern at UTC 13:35 of 3/11/11. We can see some major disseminators of information identified through their *twitter* names (e.g. NHK_PR, asahi_tokyo, etc.). The lines within the graph link users, and show retweets: every-time someone retweets (i.e. rebroadcasts) another user's post, this person is added to the original user's cluster. This is the typical cycle of information dissemination seen within social networks. At the bottom of the same figure we show the activity pattern in the same network 15 minutes later, at UTC 13:50. What has changed, is that in the meantime NHK_PR had issued three significant tweets related to the earthquake and

[7] Basically stated social network analysis (SNA) allows us to explore how different parts of a social system (e.g. people, organizations) are linked together. Moreover, it allows one to define the systems' structure and evolution over time (e.g. kinship or role-based networks). SNA is a quantitative methodology using mathematical graphs to represent people or organizations, where each person is a node, and nodes are connected to others via links (edges). Such links can be directed or undirected (e.g. friendship networks don't have to be reciprocal).

subsequent tsunami,[8] starting with a call for the general public to remain calm, and proceeding with a recommendation to evacuate, and then the ominous tsunami warning itself. In the bottom of Fig. 7 we see what is the equivalent to an information explosion within this network: a massive number of users rebroadcasts NHK_PR's tweets, forming a large cluster around it (represented by a star-like pattern). These are users who belong to NHK_PR's cluster and in that sense share the common characteristic of getting their information from the same source, and thus can be considered to be similarly-inclined in that sense.

By aggregating this type of activities over a longer period (3/11/11, from 05:00 to 19:00) we see in Fig. 8 the structure of this network. As is common in many complex networks, this network too is highly-skewed (Barabási and Albert 1999), in the sense that the majority of nodes have a low degree of connectivity (blue star-like shapes in the graph) while there are a small number of nodes which have a high degree (these can be considered as hubs of information dispersal and to some extent key actors in the social media sphere (see Asur and Huberman 2010)). For example, in this specific figure we can see that *NHK_PR* which is a national news organization, while *asahi_tokyo* and *TokyoMX* tweet mostly about local information in general, such as schools and metro services. These nodes, together with few other easily recognizable ones in Fig. 8 represent high value users within this community, as their actions reach large numbers of users and trigger cascades of activities. While we identified these users by visual inspection of the graph in Fig. 8, there exist a variety of models to describe the potential cascade effects of different members' activities within networks, considering a variety of parameters (e.g. network topology, topic), and thus identify influencers within the network (see Kimura et al. 2010; Watts and Dodds 2007; Dave et al. 2010). It should be noted here that the degree of influence that a user has within a given network has been proven to vary with the particular topic under examination (see Tang et al. 2009). In this example case we are focusing on the dissemination of a particular type of information (news and emergencies) within this network. If, for example, we were considering other information we may see different nodes emerging as influential leaders within the same network.

While this analysis captures the manner in which information is disseminated within the network, and the different levels of influence of various nodes, it also supports another purpose. Studies support the driving role of homophily in social network activities at large (McPherson et al. 2001; Singla and Richardson 2008) and twitter in particular (Choudhury et al. 2010), with individuals preferring to associate and interact (and thus cluster) with users of similar background and interest. Accordingly, a valid argument can be made that clusters identified through this analysis comprise of individuals who share opinions and are likely to belong to similar classes in the human landscape of the area under investigation. Furthermore, these clusters formed in *twitter* tend to be spread over larger geographical areas, unlike *facebook* for example, where studies have shown that friendship relations tend to be geographically clustered and inversely proportional to distance (Backstrom et al. 2010). As a matter of fact Huberman et al. (2009) showed that *twitter* users interact with small subsets of their social connections, following instead interest- and topic-driven motives in their interaction patterns (Java et al. 2007). Nevertheless, when dealing with a newsworthy situation (extraordinary events) these networks tend to cluster locally. Reports indicate that some degree of correlation appears to exist between the physical location of tweeting individuals relative to the reported event and their network importance (locals gain importance in the network when reporting about a local event), as well as that local networks tend to become denser when addressing local events (Yardi and Boyd 2010).

In order to better understand the behavior of tweeters and the structure of their groups we analyzed the Tokyo data from 3/11/11. As we see in Fig. 9a there are a large number of tweeters who only tweet once during this time period, and a small number of tweeters who tweet a lot, approaching a power law

[8] The 3 most retweeted tweets in this specific dataset in chronological order were: Tweet 1 from NHK_PR: 2011-03-11 13:51:23, asking people to remain calm; the number of retweets in the period of interest was 303. Tweet 2 from NHK_PR: 2011-03-11 13:54:59 was a warning to switch off power to homes before evacuating; the number of retweets in the period of interest was 435. Tweet 3 from NHK_PR: 2011-03-11 14:05:45 was a tsunami warning; the number of retweets in the period of interest was 258.

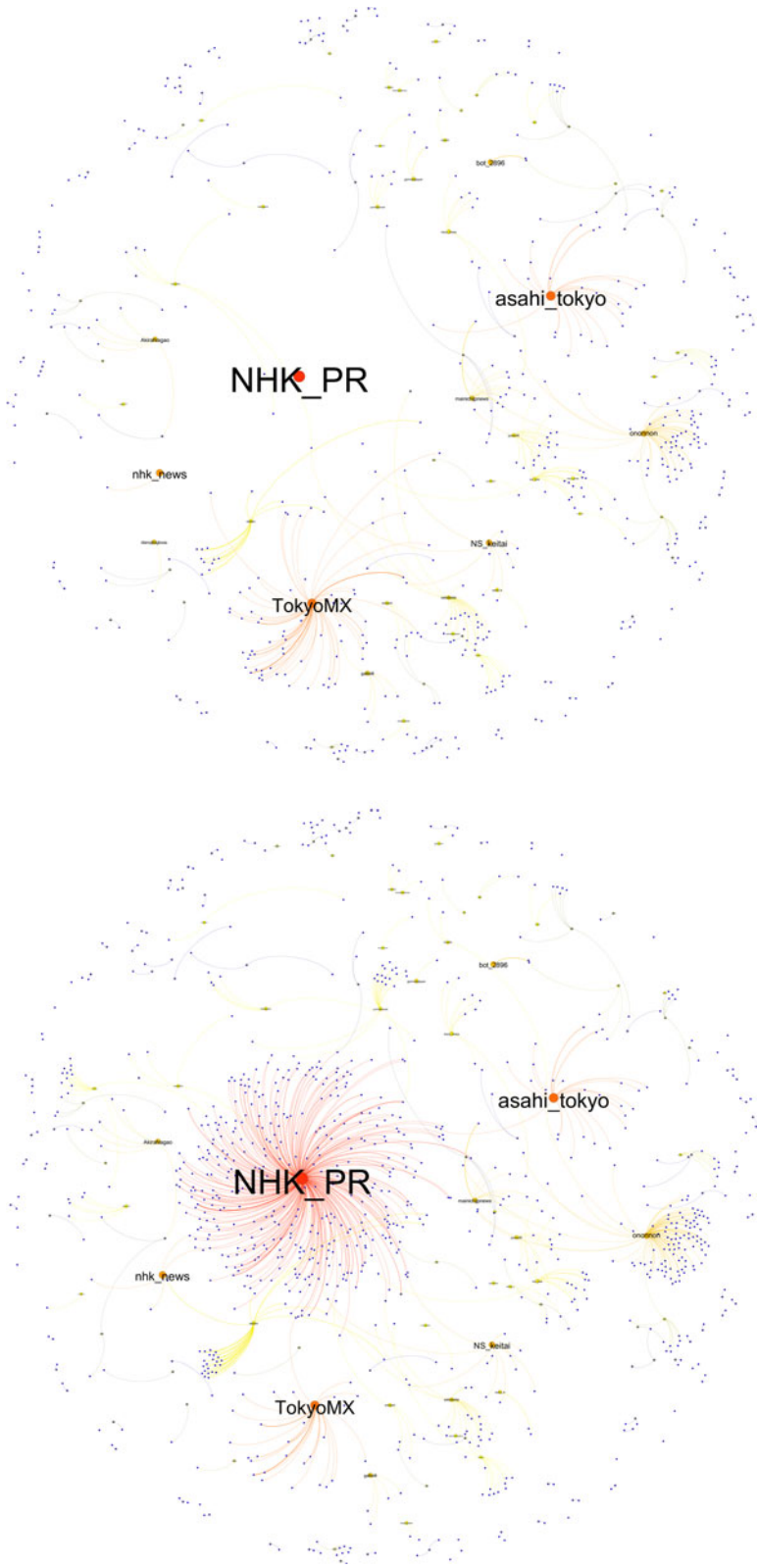


Fig. 7 Network cores and retweeting nodes before and after the earthquake. *Top* Retweets over a 15 min period at 13.35 (UTC time) *Bottom* Retweets over a 15 min period at starting at 13.50 (UTC time). The explosion in retweets is due to the three news releases tweeted by NHK_PR listed in footnote 8

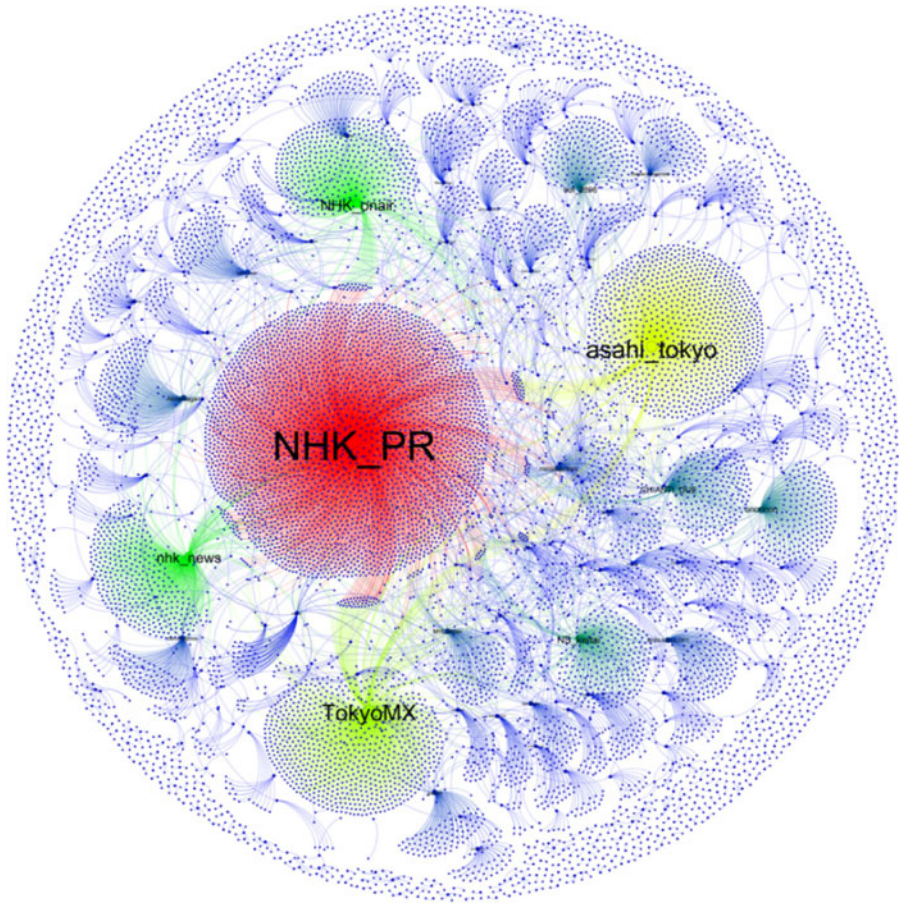


Fig. 8 Network cores and users (nodes) retweeting. The color of the edges matches the outgoing node (retweet), which is colored according to degree of importance (from *blue* for lowest, to *green*, *yellow* and *red* for highest). (Color figure online)

distribution with an exponent of 3.35 (readers interested in power laws are referred to Newman 2005). This behavior is confirming observed blogosphere characteristics (e.g. Shi et al. 2007) and comparable behavioral patterns observed in online forums (see Zhang et al. 2007) or file sharing sites (e.g. Adar and Huberman 2000). In comparison, Fig. 9b shows the retweeting behavior, which deviates from a power law distribution as it has a very heavy tail. Very few *twitter* users have their messages retweeted hundreds (or even thousands) of times, making them important disseminators of information within this network. The major hubs of Fig. 8 (e.g. NHK_PR) reside at the tail portion of the graph. Attempts to fit a power law distribution to these data fail (red line in Fig. 9b), whereas if we truncate the data and exclude the tail component on the right-hand side of the graph we could detect a power law distribution with an exponent of 1.84. Such power-law analysis is useful for modeling, in the sense that if we were attempting to predict the number of tweets we have something to validate it against (see Clauset et al. 2009, for further discussions on power law modeling).

While the preceding discussion in "Tracing information dissemination and social networks" shows how we can identify and analyze social networks arising from social media activities, like many other studies (as discussed above) it is still addressing the geographically amorphous social media space. However, as we mentioned earlier there is also accompanying locational information, and we now discuss linking the social media space to actual geographic space. In Fig. 10 we show the spatial footprint of tweet-retweet pairs in Tokyo, captured on a random instance (6/13/2011 at 8 am in this case). We can see the location of the original tweeter (marked as a sitting bird next to the 'source author' tag), and a link to the location of the person retweeting the original message (marked as the flying bird next to the 'retwitter' tag). Location information is either available directly by

original authors (having the geolocation information on while they tweet, as is the case in this specific example), or can be deduced from IP addresses using any of the IP geolocation solutions (see Eriksson et al. 2010). The accuracy of this geolocated information may range from building level all the way to broader neighborhood, or city level, depending in the manner in which it is acquired (see for example, Poese et al. 2011). Hecht et al. (2011) showed that, left to their own devices, the vast majority of users (64% of the geolocated tweets) prefer to provide their locational information at the city level, with state level coming in as second choice. This is consistent with the overall heuristic principles describing the manner in which people decide what type of locational information they are willing to disclose as presented by Ludford et al. (2007). Furthermore, work by Cheng et al. (2010) showed that even without any IP or geolocation information, users' location can often be estimated at the city level based purely on the content of their messages. Comparable work has been performed on predicting the location of *flickr* users through an analysis of their contributions' content (Popescu and Grefenstette 2010), and of *facebook* users through an analysis of their social network (Backstrom et al. 2010).

In the *twitter* feeds that we collected for this and comparable experiments we have approximately 16% of the feeds with detailed location information (coordinates), while another 45% of the tweets had locational information at coarser granularity (e.g. city level). There is a disparity of reported values regarding the rates of disclosure of geolocation information by twitter users. For example, shortly after twitter's introduction Java et al. (2007) reported that approximately 52% of *twitter* users (39 k out of 76 k) in their study had provided some location information in the corresponding entry of their profiles, while more recently Hecht et al. (2011) reported that two out of three users in their study provided some type of geolocation information. More recently Cheng et al. (2010) reported that 5% of users in their study listed locational information at the level of coordinates, with another 21% of users listing locational information at the city level. These variations in the reported percentages of geolocated tweets could be attributed to the fact that precise locational information is more often associated with mobile devices, and thus we have higher percentage of such information available in areas where latest technology is more easily and rapidly adopted. Regardless of the manner in which locational information was obtained, once it is available it can be used to identify the spatial footprint of clusters of social networks like the ones identified in Fig. 8.

In Fig. 11 we map the originating location of twitters in the Tokyo area, expressing this information as density of tweets per cell at two different instances: 9 pm on 3/12/11, and 2 am of 3/13/11 (just after the catastrophic earthquake). Information has been aggregated in cells of 2×2.5 km for visualization purposes, but can also be available at a finer resolution (city block level). We have collected approximately 200,000 tweets from this area over a two-day period, and the data presented in the figure show the relative density of tweets from this area (red and deeper yellow shows the highest number of tweets per cell, lighter shows lesser).

The examples we presented above demonstrate newfound capabilities offered by harvesting geospatial information from social media feeds. The collected data can be analyzed to identify clusters of users who share interests and opinions. Further analysis of these clusters of social networks can reveal valuable network information, for example, the main providers of information within them, the manner in which this information is disseminated to large groups of users, and other relevant characteristics. While in our approach we focused on particular geopolitical events (e.g. the Arab Spring events, and the Japan earthquake), the analysis can be performed at any instance and focusing on any topic. In this manner we can collect a variety of parameters describing the composition of crowds, ranging from cultural and political to health and economics. This information can be subsequently analyzed to identify similarities in citizen groups. By geolocating this information we are presented with unprecedented opportunities to harvest human geography data in real-time and at fine spatial resolutions. This information can be valuable to a wide range of operations, ranging from natural disaster response to product market analysis.

Discussion and outlook

The motivation for this paper came from the unprecedented developments in social media and the resulting effects on actively and passively contributed

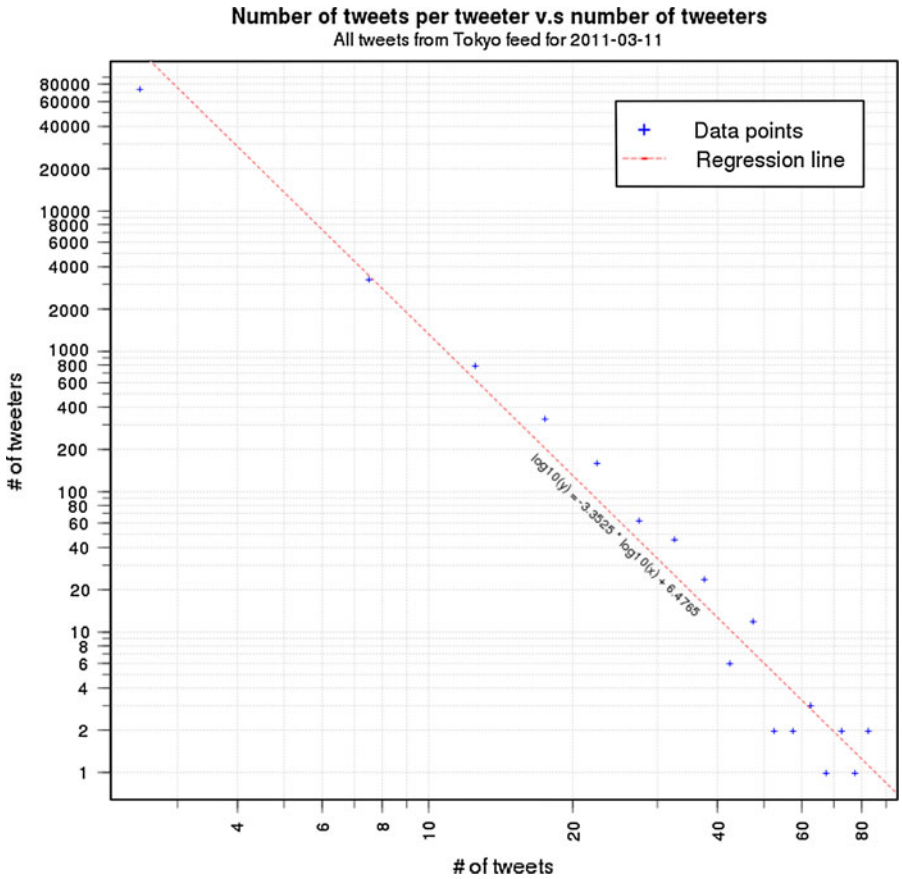


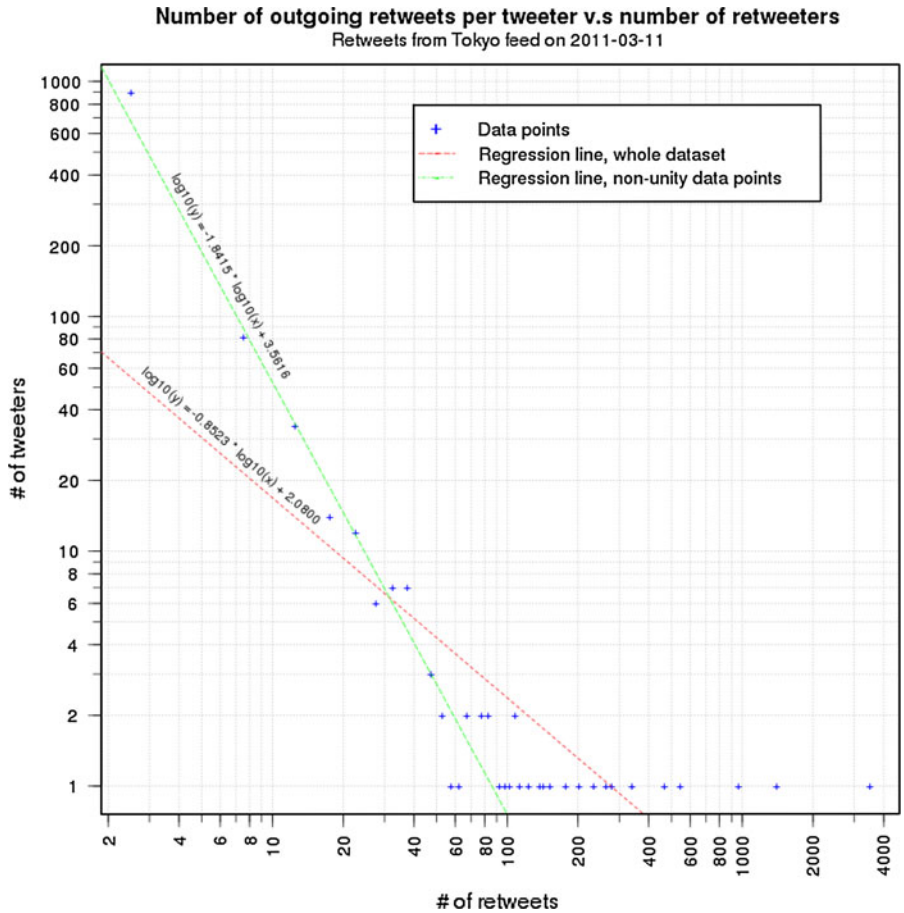
Fig. 9 **a** Log-log plots of tweets and tweeters of geolocated messages on the 3/11/11 for Tokyo: the number of tweets per tweeter (*horizontal axis*) versus the number of tweeters (*vertical axis*). The most active tweeters (e.g. news organizations) are clustered at the *bottom right* of the graph, whereas occasional users are at the *top left*. **b** Log-log plots of tweeters and retweets of geolocated messages on the 3/11/11 for Tokyo: number of outgoing retweets per originating tweeter (*horizontal axis*) versus the number of retweeters (*vertical axis*). Tweeters with large follower base are at the lower right-hand side part of this graph

geographic information. That provides us with unique opportunities to collect in real-time data in epic scale and geolocate this information for analysis. Unlike VGI where people are acting as sensors, in ambient geographical information (AGI) they are also the observations from which we can get a better understanding of various parameters of the human landscape. For example, people's tweets act as sensor measurements in the sense that the Japan data around earthquake shows concern, responses and also captures the way in which events become part of normal life. While the data from Egypt shows potential hot spot emergence. One could consider these as altering notions of how we explore the geographical and social systems. We can observe the collapse of a physical infrastructure as it is affecting its people (e.g. Japan), or the collapse of a social system while leaving the physical infrastructure intact (e.g. Cairo) or not (e.g. Libya). In a sense, such data streams harvested from human sensors have similarities to how one uses rain and stream gauges to monitor flooding in early warning systems.

Unlike VGI, AGI focuses upon passively contributed data and the paper has highlighted a number of applications whereby one can harvest this ambient geospatial data and turn this information into knowledge about what is happening around the world. However, one has to note that there is an issue of only getting a sample of the population when collecting AGI from social media feeds, namely individuals who are active in this arena. Nevertheless this sample is rapidly growing, as relevant technology adoption is becoming more ubiquitous.

This rise in social media and the ability for analysis raises several concerns with respect to the suitability of traditional mapping and GIS solutions to handle this type of information. We no longer map just buildings and infrastructure, but we can now map abstract concepts like the flow of information in a society, contextual information to place and linking both

Fig. 9 continued



quantitative and qualitative analysis in human geography. In a sense one could consider AGI to be addressing the fact that the human social system is a constantly evolving complex organism where people's roles and activities are adapting to changing conditions, and affect events in space and time. By moving beyond simple mashups of social media feeds to actual analysis of their content we gain valuable insight into this complex system.

What is to come next is difficult to predict. For example, consider that only 10 years ago the idea of location-based services and GPS embedded into mobile devices was still in its infancy. Advances in tools and software made geographical information gathering easier, resulting into growing trends in crowdsourcing geographical data rather than using authoritative sources (such as National Mapping agencies). More recently, the popularity of geographically tagged social media is facilitating the emergence of location as a commodity that can be used in organizing content, planning activities, and delivering services. We expect this trend to increase as mobile devices become more locationally aware. One could relate this to the growing usage of online activities and services (such as real-time information on social media sites as *Foursquare*, *facebook places*, *Google Latitude*, *twitter* and *Gowalla* and a host of new sites and services emerging with time). But also of more static sites (in the sense on can upload when wants) such as *flickr*, *YouTube* etc. provides means of viewing and in a sense forming an opinion of a place without actually visiting.

Harvesting ambient information brings forward novel challenges to the issue of privacy, as analysis can reveal information that the contributor did not explicitly communicate (see Friedland and Sommer 2010). But this is not a new trend; it has actually been happening for a while now. Google itself is basically a marketing tool using the information it collects to improve its customer service. Similarly, *twitter* makes

Fig. 10 Geolocating pairs of tweeters and retweeters

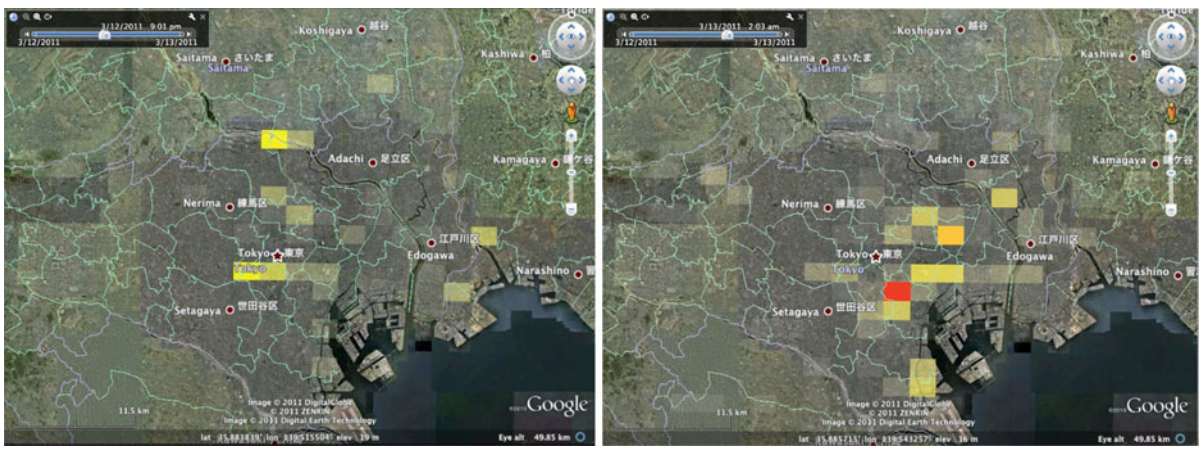


Fig. 11 Twitter traffic mapping: originating locations of tweets (density of tweets per 2 × 2.5 km lock) in the area around Tokyo on the evening after the Fukushima earthquake (3/13/11) (*left* 9.01 pm on the 3/12/2011; *right* at 2.03 am on the 3/13/2011)

money by licensing their tweet fire hose to search engines, while companies can pay for "promoted tweets"(see Financial Times 2010). And this trend has already spread to locational information. For example, TomTom (2011) has been using passively sensed data for helping police with placing speed cameras. This comes alongside iPhones storing locational data while the user is unaware (BBC 2011). However, people are making progress in highlighting the issue of privacy relinquishing when sharing locational information. Sites and apps such as such as *pleaserobme.com* or the *creepy,*[9] a geolocation aggregator have demonstrated the potential for aggregating social media to pinpoint user locations. Trying to protect people's identities in times of unrest is also a well-recognized concern, for

[9] http://ilektrojohn.github.com/creepy/.

example, the Stand by Task Force (2011) suggest ways of limiting expose and delay information for the recent unrest in North Africa.

But the power of harvesting AGI stems from gaining a deeper understanding of groups rather than looking at specific individuals. As the popularity of social media is growing exponentially we are presented with unique opportunities to identify and understand information dissemination mechanisms and patterns of activity in both the geographical and social dimensions, allowing us to optimize responses to specific events, while the identification of hotspot emergence helps us allocate resources to meet forthcoming needs.

References

Adar, E., & Huberman, B. A. (2000). Free riding on gnutella. *First Monday*, *5*(10–2). Available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/792/701.

Anderson, P. (2007). What is Web 2.0? Ideas, technologies and implications for education. *Horizon scanning report, JISC technology and standards watch*. Available at http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanning/hs0701.aspx.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (pp. 492–499). Toronto, Canada.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW'10* (pp. 61–70). Raleigh, NC.

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Batty, M., Hudson-Smith, A., Milton, R., & Crooks, A. T. (2010). Map mashups, Web 2.0 and the GIS revolution. *Annals of GIS, 16*(1), 1–13.

BBC. (2011). *iPhone tracks users' movements.* Available at http://www.bbc.co.uk/news/technology-13145562. Accessed on 27th, June, 2011.

Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings AAAI conference on weblogs and social media* (pp. 291–300). New York, NY.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Proceedings AAAI conference on weblogs and social media* (pp. 438–441). Barcelona, Spain.

Biewald, L., & Janah, L. (2010). TechCrunch: Crowdsourcing disaster relief. Available at http://techcrunch.com/2010/08/21/crowdsourcing-disaster-relief/. Accessed on January, 26th, 2011.

Bowman, C., Danzig, P., Hardy, D., Manber, U., & Schwartz, M. (1995). The harvest information discovery and access system. *Computer Networks and ISDN Systems, 28*(1–2), 119–125.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance Sans Frontiéres: Internet-based emerging infectious disease intelligence and the Healthmap project. *PLoS Medicine, 5*(7), 1019–1024.

Buys, P., Dasgupta, S., Thomas, T. S., & Wheeler, D. (2009). Determinants of a digital divide in sub-Saharan Africa: A spatial econometric analysis of cell phone coverage. *World Development, 37*(9), 1494–1505.

Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., & Shivakumar, N. (1999). Exploiting geographical location information of web pages. In *Proceedings ACM SIGMOD workshop on web and databases—WebDB*. Philadelphia, PA.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings ACM conference on information and knowledge management CIKM'10* (pp. 759–768). Toronto, Canada.

Choudhury, M. D., Sundaraman, H., John, A., Seligmann, D. D., & Kelliher, A. (2010). *Birds of a feather: Does user homophily impact information diffusion in social media?*' Available at http://arxiv.org/abs/1006.1702.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*(4), 661–703.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *International conference on tools with artificial intelligence—ICTAI'97* (pp. 558–567). Newport Beach, CA.

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health, 7*(2), 596–615.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115–122). Washington, DC.

Dave, K., Bhatt, R., & Varma, V. (2010). Modeling action cascades in social networks. In *Proceedings of the fifth international AAAI conference on weblogs and social media* (pp. 121–128). Barcelona, Spain.

Elwood, S. (2010). Geographic information science: Emerging research on the societal implications of the geospatial web. *Progress in Human Geography, 34*(3), 349–357.

Eriksson, B., Barford, P., Sommers, J., & Nowak, R. (2010). A learning-based approach for IP geolocation. In: A. Krishnamurthy, B. Plattner (Eds.), *Passive and active measurement*, *lecture notes in computer science* (Vol. 6032, pp. 171–180). Berlin, Germany: Springer.

Financial Times. (2010). Coke sees 'phenomenal' result from twitter ads. *Financial Times* (June, 25th). Available at http://www.ft.com/cms/s/0/6726ef4e-805a-11df-8b9e-00144feabdc0.html#axzz1OoiRiNWO. Accessed on 27th, June, 2011.

Firedland, G., Choi, J., Lei, H., & Janin A. (2011). Multimodal location estimation on flickr videos. In *Proceedings of MultiMedia'11*. Scottsdale, AZ.

Friedland, G., & Sommer, R. (2010). Cybercasing the joint: On the privacy implications of geotagging. In *Proceedings of*

the fifth USENIX workshop on hot topics in security (HotSec 10). Washington, DC.

Goodchild, M. F. (2007a). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Goodchild, M. F. (2007b). Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, *2*(1), 24–32.

Graham, P. (2007). Web 2.0. Available at http://www.paulgraham.com/web20.html. Accessed on May 1st, 2008.

Gravano, L., Hatzivassiloglou, V., & R. Lichtenstein. (2003). Categorizing web queries according to geographical locality. In *Proceedings of the conference on information and knowledge management—CIKM* (pp. 325–333). New Orleans, LA.

Guardian. (2011). Assault on Zawiyah. *The Guardian* (March 8th). Available at http://www.guardian.co.uk/world/2011/mar/08/arab-and-middle-east-protests-libya. Accessed on 27th, June, 2011.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B, 37*(4), 682–703.

Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass, 2*(6), 2011–2039.

Hecht, B., Hong, L., Suh, B., & Chi, E. (2011). Tweets from Justin Bieber's heart: The dynamics of the 'location' field in user profiles. In *Proceedings of the ACM CHI conference on human factors in computing systems*. Vancouver, Canada.

Howe, J. (2006). The rise of crowdsourcing. *Wired 14.06*, 161–165. Available at http://www.wired.com/wired/archive/14.06/crowds.html. Accessed on September 25th, 2008.

Huberman, B. A., Romero, D. M., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday, 14*, 1–5.

Hudson-Smith, A., Batty, M., Crooks, A. T., & Milton, R. (2009a). Mapping tools for the masses: Web 2.0 and crowdsourcing. *Social Science Computer Review, 27*(4), 524–538.

Hudson-Smith, A., & Crooks, A. T. (2009). The renaissance of geographic information: Neogeography, gaming and second life. In H. Lin & M. Batty (Eds.), *Virtual geographic environments* (pp. 25–36). Beijing: Science Press.

Hudson-Smith, A., Crooks, A. T., Gibin, M., Milton, R., & Batty, M. (2009b). Neogeography and Web 2.0: Concepts, tools and applications. *Journal of Location Based Services, 3*(2), 118–145.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Joint 9th WEBKDD and 1st SNA-KDD workshop'07* (pp. 56–65). San Jose, CA.

Keen, A. (2007). *The cult of the amateur: How today's internet is killing our culture*. New York, NY: Currency.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., & Rattenbury, T. (2007). How flickr helps us make sense of the world: Context and content in community-contributed media collections. In T*he proceedings of multiMedia'07* (pp. 631–640). Augsburg, Germany.

Kimura, M., Saito, K., & Motoda, H. (2010). Extracting influential nodes on a social networks for information diffusion. *Data Mining and Knowledge Discovery, 20*(1), 70–97.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations, 2*(1), 1–15.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *WWW'10* (pp. 591–600). Raleigh, NC.

Longley, P. A., Ashby, D. I., Webber, R., & Li, C. (2006). Geodemographic classifications, the digital divide and understanding customer take-up of new technologies. *BT Technology Journal, 24*(3), 67–74.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2010). *Geographical information systems and science* (3rd ed.). New York, NY: Wiley.

Ludford, P. J., Priedhorsky, R., Reily, K., & Terveen, L. G. (2007). Capturing, sharing, and using local place information. In *Proceedings of CHI '07* (pp. 1235–1244). San Jose, CA.

MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J., & Mitra, P. (2011). Geotwitter analytics: Applications in crisis management. In *Proceedings of the 25th international cartographic conference*. Paris, France.

McPherson, M., Smith-Lovin, L., & Cook, J. C. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*, 415–444.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology, 62*(5), 902–918.

National Audubon Society. (2011). *Christmas bird count*. Available at http://birds.audubon.org/christmas-bird-count. Accessed on 26th April, 2011.

Newman, M. E. J. (2005). Power laws, pareto distributions and Zipf's Law. *Contemporary Physics 46*(5), 323–351.

Newsweek. (2009). A twitter timeline of the iran election. *Newsweek*. Available at http://www.newsweek.com/2009/06/25/a-twitter-timeline-of-the-iran-election.html. Accessed on April 27th, 2011.

Norheim-Hagtun, I., & Meier, P. (2010). Crowdsourcing for crisis mapping in Haiti. *Innovations: Technology Governance, 5*(4), 81–89.

O'Reilly, T. (2005). What is Web 2.0: Design patterns and business models for the next generation of software. Available at http://www.oreillynet.com/lpt/a/6228. Accessed on February 20th, 2009.

Parry, M. (2011). Academics join relief efforts around the world as crisis mappers. *The Chronicle of Higher Education* (March, 27th). Available at http://chronicle.com/article/Academics-Join-Relief-Efforts/126912/#. Accessed on 27th June, 2011.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. In: *Human language technologies—HLT'10* (pp. 181–189).

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? *Computer Communication Review, 4*(2), 53–56.

Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases, 47*(11), 1443–1448.

Popescu, A., & Grefenstette, G. (2010). Mining user home location and gender from flickr tags. In *Proceedings international conference on weblogs and social media—ICWSM'10* (pp. 307–310). Washington, DC.

Rattenbury, T., & Naaman, M. (2009). Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web, 3*(1), 1–30.

Ritterman, J., Osborne, M., & Klein, E. (2009). Using prediction markets and twitter to predict a swine flu pandemic. In F. M. Carrero, J. M. Gómez, B. Monsalve, E. Puertas & J. C. Cortizo (Eds.), *1st international workshop on mining social media* (pp. 9–18). Sevilla, Spain.

Russell, M. A. (2011a). *Mining the social web*. O'Reilly Media.

Russell, M. A. (2011b). *21 recipes for mining twitter*. O'Reilly Media.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW '10*, Raleigh, NC.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. NewYork: McGraw-Hill.

Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., & Sperling, J. (2009). TwitterStand: News in tweets. In *ACM GIS 2009* (pp. 42–51). Seattle, WA.

Schindler, G., Brown, M., & Szeliski, R. (2007). City-scale location recognition. In *IEEE conference on computer vision and pattern recognition (CVPR'07)* (pp. 1–7). Minneapolis, MN.

Shi, X., Tseng, B., & Adamic, L. A. (2007). Looking at the Blogosphere topology through different lenses. In *Proceedings of the international conference on weblogs and social media* (*ICWSM 2007*). Boulder, CO.

Singla, P., & Richardson, M. (2008). Yes, there is a correlation-from social networks to personal behavior on the web. In *WWW'08* (pp. 655–664). Beijing, PRC.

Standby Task Force. (2011). The security and ethics of live mapping in repressive regimes and hostile environments. Available at http://blog.standbytaskforce.com/?p=259. Accessed on 27th, June, 2011.

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 807–816). Paris, France.

TomTom. (2011). This is what we really do with your data. Available at http://www.tomtom.com/page/facts. Accessed on 27th June, 2011.

Wall Street Journal. (2011). Where the young and tech-savvy go? *Wall Street Journal* (May 19). Available at http://blogs.wsj.com/digits/2011/05/19/a-week-on-foursquare/. Accessed on 27th June, 2011.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion. *Journal of Consumer Research, 34*(4), 441–458.

Weng, J., & Lee, B.-S. (2011). Event detection in twitter. In *Proceedings of the AAAI conference on weblogs and social media* (*ICWSM-11*). Barcelona, Spain.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitters. In *Proceedings of the web search and data mining* (*WSDM'10*). New York, NY.

Yardi, S., & Boyd, D. (2010). Tweeting for the town square: Measuring geographic local networks. In *Proceedings of fourth international AAAI conference on weblogs and social media* (pp. 194–201). Washington, DC.

Zhang, J., Ackerman, M.S., & Adamic, L. (2007). Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (pp. 221–230). Banff, Canada.

Zhang, W., & Kosecka, J. (2006). Image based localization in Urban environments. In *International symposium on 3D data processing, visualization, and transmission* (pp. 33–40). Chapel Hill, NC.

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Medical and Health Policy, 2*(2), 7–33.