Spatiotemporal Transformation of Social Media Geostreams: A Case Study of Twitter for Flu Risk Analysis

Myung-Hwa Hwang[1,2], Shaowen Wang[1,2,3], Guofeng Cao[1,4],
Anand Padmanabhan[1,2,3], Zhenhua Zhang[1]
Cyberinfrastructure and Geospatial Information Laboratory (CIGI)[1]
Department of Geography and Geographic Information Science[2]
National Center for Supercomputing Applications (NCSA)[3]
University of Illinois at Urbana-Champaign (UIUC)
Urbana, IL 61801
Department of Geosciences[4]
Texas Tech University
Lubbock, TX 79409
{mhwang4,shaowen,guofeng,apadmana,zzhcn}@illinois.edu

ABSTRACT

Georeferenced social media data streams (social media geostreams) are providing promising opportunities to gain new insights into spatiotemporal aspects of human interactions on cyber space and their relation with real-world activities. In particular, such opportunities are motivating public health researchers to improve the surveillance of disease epidemics by means of spatiotemporal analysis of social media geostreams. One essential requirement in achieving such geostream-based disease surveillance is to establish scalable data infrastructures capable of real-time transformation of massive geostreams into spatiotemporally organized data to which analytical methods are readily applicable. To fulfill this requirement, this study develops a data pipeline solution where multiple computational components are integrated to collect, process, and aggregate social media geostreams in near real time. As a test case, this solution focuses on one well-known social media geostream, the Twitter data stream, and one type of disease epidemics, the flu. The pipeline solution facilitates multiscale spatiotemporal analysis of flu risks by collecting geotagged tweets from the Twitter Streaming API, identifying flu-related tweets through keyword match, aggregating tweets at multiple spatial granularities in near real time, and storing tweets and the aggregate statistics in a distributed NoSQL database. Although developed for the surveillance of flu epidemics, the pipeline would serve as a general framework for building scalable data infrastructures that can support real-time spatiotemporal analysis of social media geostreams in the application domains beyond disease mapping and public health.

Categories and Subject Descriptors

H.3.1 [**Database Applications**]: Spatial databases and GIS; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Management, Design

IWGS '13:, November 05 - 08 2013, Orlando, FL, USA
Copyright 2013 ACM 978-1-4503-2532-5/13/11…$15.00.
http://dx.doi.org/10.1145/2534303.2534310

Keywords

Social media geostreams, Spatiotemporal analysis, Data pipeline, Disease surveillance

1. INTRODUCTION

Social media are a collection of online tools designed for social networking, blogging, microblogging, and sharing of opinions, messages, and multimedia [26]. By inviting users both to consume and produce contents, social media have gained popularity in recent years as an important space for personal communications and social interactions. A distinctive aspect of social media as a data source is that they produce real-time, massive data streams of users' conversations and the associated location information. An example of such media is the Twitter[1] that allows the user to post messages of 140 characters or tweets. In 2012, this service alone generated more than 340 million tweets every day and more than 3,900 tweets every second[2]. Even if only 0.8% of these tweets had location information as Ghosh and Guha [11] reported of data from the Twitter Search API, Twitter produces over 2.7 million geotagged tweets per day. This rate of data collection indicates the volume of tweet data stream is large and growing rapidly. The continuous, unprecedented growth of georeferenced social media data streams or social media geostreams (SMG) presents promising avenues for examining spatiotemporal dynamics in various human behaviors as well as for advancing theories and methods for spatiotemporal analytics of geostreams.

Disease surveillance has taken a pioneering step towards harnessing SMG, together with spatiotemporal analytics [31]. Traditional approaches to disease surveillance have relied on officially reported statistics based on clinical data of patient visits and diagnosis [21]. These statistics, however, have been released with time lags, providing limited value to the surveillance efforts. Recently, data of indirect signals of the epidemics have been used for surveillance [17, 28]. Examples include call volumes to telephone triage advice lines, over-the-counter drug sales, web queries, and social media data [10]. In comparison to other types of data, social media data streams have a few advantages for disease surveillance: they are available immediately after a message is posted, accessible easily through application programming interfaces (APIs), obtainable for free to a certain

[1] https://twitter.com/

[2] https://blog.twitter.com/2012/twitter-turns-six

extent, and less prone to privacy issues due to volunteered data contribution [27].

The potential of SMG in disease surveillance may be best revealed when such data are examined with spatiotemporal analytics. The occurrence and spread of diseases are a dynamic process that is significantly affected by human mobility. This is because individuals' movements would strongly influence where and when they were exposed to the potential causal factors of diseases, including other infected individuals [16]. Unlike temporal trend analyses which aim to detect whether an abnormal change in disease risk is occurring or not on a global spatial scale, spatiotemporal analysis methods are designed to capture how such abnormal behaviors vary geographically and how these variations change over time (e.g., periodicity) and interact with other spatial processes such as the movements of infected patients or susceptible population [5]. These capabilities of spatiotemporal analytics lend themselves well to the surveillance of dynamic disease processes and geotagged data streams from social media.

The objective of this study is to provide a methodological framework for facilitating the use of SMG for spatiotemporal surveillance of disease epidemics. In particular, it focuses on answering the following question: "*what data infrastructure is needed to enable the application of spatiotemporal analytics for SMG in support of real-time surveillance of disease epidemics?*" The motivation for this question lies in multiple challenges with transforming SMG in near real time into spatiotemporal data entities to which spatiotemporal analytics are readily applicable. These challenges can be classified into two types: those related to the nature of SMG and exploratory spatiotemporal analysis.

Two features of SMG present challenges for spatiotemporal disease surveillance: (a) the continuous streaming of massive spatial data and (b) the textual, unstructured representation of data contents. The conversations in social media are generated incessantly, thus requiring continuous collection and processing of the streamed data. Since the volume of data streams tends to be large, this processing requires scalable computing infrastructures; however, conventional tools for spatial data handling often lack such scalability by their design [33]. In addition, social media data are recorded in a form of raw text. Since disease surveillance focuses on the outbreaks and spread of disease, the relevance of individual conversations in social media to a particular disease needs to be determined through various methods for natural language processing or text mining [25]. The unstructured nature of social media data also necessitates the retrieval and organization of spatial, temporal, and topical attributes associated with individual conversations.

The challenges also arise due to the features of spatiotemporal analytical methods. Two of such features include (c) the need for interactive data exploration and (d) sensitivity to spatial and temporal scales. Disease surveillance involves continuous monitoring of health indicators and their spatial behaviors (e.g., spread). While interactive exploration of spatiotemporal dynamics in such indicators facilitates the detection of significant patterns and anomalies [28], it is challenging to provide quality user experience in such interaction without the support of backend infrastructures for efficient data processing and query [16]. And the spatiotemporal processes underlying disease outbreaks and spread are complex and may be better understood through scale-specific risk indicators. Established methods for spatiotemporal analysis are usually designed to produce risk estimates specific to a spatiotemporal scale. To accommodate this sensitivity of spatiotemporal analytical methods to scale, SMG-based systems for disease surveillance need to provide capabilities to transform and organize social media data in support of multiscale analysis.

This study addresses the aforementioned challenges by developing a data pipeline solution in which multiple computational components are integrated to collect, process, and aggregate SMG for the monitoring of flu risk indicators. This solution, termed the FluMapper data pipeline in this paper, addresses the challenge (a) by combining a bespoke social media crawler with a scalable, high-performance, and open source NoSQL database for continuous, sliding-window-based collection and processing of SMG. The challenges (b), (c), and (d) are then addressed by filtering social media data through keyword match and organizing the resultant data in a form of multiscale hierarchical spatial grids in which individual conversations are aggregated by varying sizes of unit cell in the grids. A record in the grids contains summary statistics of the conversations and episodic movements of social media users. Here, each cell in the grid serves as a measurement unit for observing the presence and locational changes of social media activities, and discontinuity in the observed locations of such activities renders the collected movement data episodic rather than continuous [2]. The data pipeline solution of the study is discussed in the context of harnessing streaming tweets data for early detection of flu epidemics in the conterminous United States (US).

In what follows, Section 2 provides a background of disease surveillance and its use of SMG. Section 3 motivates this study by highlighting the needs for combining real-time processing of SMG with spatiotemporal data frameworks for disease epidemics surveillance. Section 4 explains the data pipeline solution of the study, namely the FluMapper data pipeline. Section 5 illustrates the capacity of the solution and a case study of analyzing the collected data for flu risk analysis. Section 6 concludes the study.

2. BACKGROUND

Social media geostreams (SMG) have been increasingly used for disease surveillance. This section discusses the concept and practices of disease surveillance and how it benefits and is challenged from SMG.

2.1 Disease surveillance

The Centers for Disease Control and Prevention (CDC) defined disease surveillance as "the ongoing systematic collection, analysis, and interpretation of outcome-specific data for use in planning, implementation, and evaluation of public health practice" [6]. The crucial purpose of disease surveillance is to enable early identification of infectious diseases that may put a large number of people into health dangers ranging from simple coughs to fatal conditions [21]. The importance of early detection of disease epidemics has been increasingly recognized: Today, disease can spread across the world as quickly as a few hours due to advances in transportation systems; disease infections have occasionally been used for bioterrorism; and the process of disease infection and recovery incurs substantial costs across local, national, and world economies. Even for the case of influenza-like illness (ILI), for example, a CDC report showed that the total estimate of flu-related economic burden was about $87 billion when a 2003 population was used as the baseline [23].

Conventional approaches to disease surveillance are to monitor temporal changes in the pre-diagnostic health indicators and statistics derived from clinically relevant data sources. Such data sources include daily visits to physician offices and hospital emergency departments, hospital admissions, and laboratory-test requests [12]. Although these clinical data may capture direct signals of disease outbreaks by collating information of patient-reported symptoms, they often become available with some time lags that may have substantial effects on the prevention and responses to disease spread. To address this drawback, health researchers and practitioners have started to use data containing indirect signals of disease infections. Such data are collected through systematic reporting networks or gleaned from informal reports like Internet news and discussions in social media [17].

Multiple non-clinical data-reporting systems exist that enable the identification of emerging diseases. Examples include pharmacy chains systems of over-the-counter drug sales, databases of citizen calls to emergency medical services and 911, data systems for telephone triage hotlines, and records of school absenteeism and nurses [21]. Unlike clinical data sources, these systems provide real-time data about people's health-seeking behaviors that may be associated with changes in their health status. Research in public health found that outbreaks of certain diseases could be detected earlier with such indirect signals than clinical data [22]. Different types of data are also found to increase true positives or decrease false positives of certain diseases (e.g., triage calls being more sensitive to acute illness; [15]). The data-reporting systems of indirect health signals, however, have critical limitations: they are difficult to build beyond regional scales without investing substantial resources, and the data access tends to be restrictive.

Alternative to formal data-reporting systems, indirect signals of disease outbreaks can be captured through informal data sources whose original purpose has little to do with health indicators per se. Such data sources include Internet news [17], online documents, queries submitted to journal services and web search engines [12], and user-contributed contents available at social media [27]. Provided a set of target diseases to detect, the data from these sources undergo multiple steps of data processing to determine the relevance of data reports to certain syndromic symptoms. Temporal variations in the resulting statistics or predictions from analytical models using the statistics are monitored to identify abnormal increases in disease onsets. Despite the need for intensive data processing, the informal data are useful for early detection of disease epidemics especially at global scales. This is because they become available immediately after news or reports are released online, and Internet users are distributed worldwide and employ the online resources to understand and represent the disease symptoms of their interest.

2.2 Social-media-based disease surveillance

Arguably, social media geostreams (SMG) are the most popular informal data that have been investigated in the latest literature of disease surveillance. For instance, Quincey and Kostkova [25] highlighted the potential of Twitter data in informing epidemics intelligence by examining collocations of flu-related keywords in tweets. Collier and Doan [8] took a step further to develop an automated text mining system in which tweets were classified in real time to six categories of disease syndromes so as to serve as risk indicators for various classes of illness. Gomide et al. [13] constructed a computational model for analyzing sentiment, temporal trend, and spatiotemporal clustering of dengue-related tweets in Brazil. Chen et al. [7] and Achrekar et al. [1] developed research infrastructures to facilitate flu trend prediction based on Facebook and Twitter data. Similarly, Lampos et al. [18] and Signorini et al. [27] constructed SMG-based information systems to estimate flu rates and scores by means of temporal regressions and text mining algorithms. Bifet et al. [3] developed data infrastructures in which the contents of tweets were analyzed and annotated in terms of their relevance to disease syndromes.

The rising interest in SMG-based disease surveillance is strongly related to the relative advantages of SMG over traditional forms of data. Five such advantages are as follows. First, the geostream data not only include rich information about various social themes but also manifest the dynamic process of how humans interact through networks of communications over time as well as over cyber and real-world spaces [29]. This aspect of social interactions is difficult to capture at a large scale through surveys, interviews, or other reports on health-seeking behaviors. Second, most traditional data, such as data from surveys, interviews, and clinical reports, produce retrospective information due to the cost involved in data collection. In contrast, social media provide up-to-date information of social phenomena through real-time data collection, thus facilitating early detection of disease epidemics or growing public concerns about them [27]. Third, SMG contain individual-level observations with spatial and temporal references. From an epidemiological standpoint, this detailedness of the data allows for exploratory analysis of spatiotemporal dynamics of disease processes across varying ranges of scales [16]. Such multiscale explorations have been often disallowed with conventional data because they have been collected only at one type of predefined spatial unit for a limited time period. Fourth, social media data are accessible for free to a certain extent. For example, the Twitter provides public data streams from which 1% of the entire tweets are obtainable. Usually, social media services provide easy-to-use APIs to facilitate users' data access. Finally, since users contribute contents voluntarily to social media, the resultant data are less restrictive regarding data usage, sharing, and publication, in comparison to survey data.

Despite the promising features, SMG also have notable characteristics that complicate their use for spatiotemporal surveillance of disease epidemics. Three such characteristics are discussed here, with the first attributable to the nature of data collection and the rest attributable to the syntax and semantics of the collected data [25]. First, data from SMG are massive and evolve continuously. Because of real-time data collection, the accumulative data easily grow beyond the capabilities of any mainstream geographic information systems (GIS). And location information included in SMG continues to change as the user moves around. While such movement data have the potential to shed insights into human-disease interactions, there are few data management and analytics solutions that can efficiently retrieve the movement data from massive geostreams and handle their developments over time [14, 16]. Second, social media data are in unstructured textual forms. Intensive work for data modeling and processing are thus required to transform messages in social media data into structured spatial data entities (e.g., point patterns and trajectories) to which spatiotemporal analytical methods are applicable [20]. Finally, social media data are conversations that need to be interpreted. Raw texts of conversations do not bear any meaning unless the researcher specifies the topic of interest [18]. Even when this topic is specified, the relevance of each conversation to the topic needs to be computed through techniques of natural language processing or text mining. This need for data interpretation, together with the reliance on sampled data (e.g., only 1% samples of tweets being accessible through Twitter's public data streams), raises substantial concerns about data quality and uncertainty.

The distinctive characteristics of SMG, either advantageous or challenging, provide numerous research opportunities for improving the science of spatiotemporal surveillance of disease epidemics [29]. Geosocial communications and networks that are observable in real time through social media would allow for advances in theories and practices of how people alert and should be alerted of health dangers and preventive measures. The richness of contextual information contained in SMG (e.g., who discussed what where, when, and how) is encouraging researchers to invent new methodological approaches to disease surveillance. The complicated aspects of SMG also motivate researchers in GIScience and other disciplines to develop new computational theories and methods for the modeling, processing, analysis, visualization, and syntheses of SMG, which would in turn advance the practice of spatiotemporal surveillance of disease outbreaks. This study contributes to these data-driven advances by providing a data infrastructure framework for near-real-time transformation of SGM in support of spatiotemporal analytics.

3. NEED FOR A SPATIOTEMPORAL DATA FRAMEWORK FOR ANALYTICS-ORIENTED PROCESSING OF SMG

The previous section highlighted that one key objective of disease surveillance is early detection of and fast response to disease epidemics and that SMG could facilitate the achievement of this goal through real-time provision of spatiotemporal information about health-related discussions on cyber space. The present approaches to disease surveillance, however, lack the capabilities to harness the spatiotemporal aspects of the massive geostreams from social media. This section discusses the need for filling this gap for enhanced disease surveillance.

3.1 Need for a data infrastructure for real-time geostream processing

The use of real-time data streams, such as SMG, is a relatively new approach in disease surveillance. This approach differs from conventional methods for disease surveillance in that it seeks to identify significant signals of disease epidemics through the processing and analysis of high-volume real-time data with low latency. The advantages of real-time surveillance have already been discussed in depth in the literature. For example, Wagner et al. [31] developed a real-time surveillance system in which free-text emergency department chief complaints were collected and analyzed in real time. Prospective evaluations of this system demonstrated that real-time surveillance could allow for earlier detection of disease epidemics within a few hours of the first identification of the relevant complaints [30]. Disease surveillance using real-time data streams would bring similar benefits in early detection of and fast response to disease outbreaks. Since social media streams are collecting data across the world, surveillance of these streams would bring additional benefits in enabling real-time disease surveillance in varying ranges of spatial scale [17].

The advantages of real-time geostreams, however, are yet to be achieved due to the challenge posed by the large volume of the data streams. Mainstream GIS and database technologies are often limited in providing scalable solutions for processing, storing, and managing flows of massive unstructured data [29, 33]. Furthermore, with those conventional technologies it is challenging to apply interactive queries and analytical operations even to obtain simple summary statistics of large data collections. New forms of data infrastructures are thus needed to handle and massage real-time SMG in efficient manners.

3.2 Need for an analytics-oriented spatiotemporal data framework

In disease surveillance, spatiotemporal aspects of disease outbreaks are of critical concerns to public health practitioners. Temporally, the practitioners have sought to determine the timing and duration of abnormal disease processes and the associations between these behaviors and other factors such as holidays, social events, and seasonal changes in temperature and precipitation [21]. Since the practitioners try to reduce the risks and impacts of disease epidemics by identifying and monitoring susceptible populations, they also have strived to understand the spatial aspects of disease processes. For example, the practitioners have made extensive use of health GIS to single out and characterize spatial clusters of the cohorts who are potentially exposed to risk factors, and to understand and visualize health-related and environmental factors that had affected the process and intensity of such exposures [9, 32].

Despite the usefulness of the temporal-only and spatial-only approaches, it has been frequently discussed that the spatial and temporal aspects of disease outbreaks should be considered simultaneously to better understand the interaction between dynamic disease processes and human mobility [16]. For instance, temporal changes in spatial disease patterns can be better explored through interactive animated maps of disease risk surfaces at varying ranges of spatiotemporal scale [5]. The reliability of disease risk assessments can be improved by introducing realistic assumptions of spatiotemporal heterogeneity in human exposures to risk factors and the effects of latency periods on disease onsets. Fine-grained data on human agents' space-time movements would also provide opportunities to shed lights on the spatial associations among disease exposure, transmission, and outbreaks [10]. Spatiotemporal patterns can be gleaned from the flows of such human movements, with some indicating that specific forms of spatial interactions may be correlated to aberrations in certain disease processes and that those correlations may be more apparent at a particular range of spatiotemporal scale than others.

The need for spatiotemporal disease surveillance has long been recognized, but methods and technologies are under development for its enablement with SMG [13]. For example, recent research in visual analytics has developed exploratory analytical methods for identifying spatiotemporal trends or anomalies in the thematic topics of social media for improved understandings of social media users and increased awareness of social events [28]. One challenge for extending these developments for spatiotemporal disease surveillance lies in the lack of data processing and organization framework that can facilitate real-time estimation and analysis of disease risk indicators from high-volume geostream. Spatiotemporal analytical methods are often computation-intensive and require input data of a specific structure, especially when those methods aim to support multiscale analysis. Conventional GIS lack the capabilities for data organization in support of those spatiotemporal analytics methods, let alone the abilities to deal with massive geostreams. In addition, current technologies for geostream processing and management still focus on enabling spatial queries and operations with geostreams and are still in their youth as to analytics support. New frameworks for spatiotemporal data organization thus need to be formulated to cater to the specific requirements for real-time spatiotemporal analysis of massive geostreams.

4. THE FLUMAPPER DATA PIPELINE

In response to the need for a spatiotemporal data framework for real-time processing of SMG, the study develops a solution named as FluMapper Data Pipeline. This solution facilitates near-real-time spatiotemporal surveillance of flu epidemics by processing tweet messages continuously, computing flows of Twitter users in a sliding window scheme, and aggregating the flow data in a hierarchical multiscale fashion. The resulting data of tweet events and tweeters' flows can be used for exploratory spatiotemporal analysis of flu risks and population mobility.

4.1 Goal and requirements

The overall goal of the FluMapper data pipeline is to enable near real-time processing of the tweet geostream so that the processing results can be easily used for multiscale exploratory space-time analysis of flu risks and the related population movements. In particular, this pipeline is assumed to serve as a data source for spatiotemporal analytical methods for flu risk estimation (e.g., Kernel density estimation) and movement pattern characterization (e.g., flow mapping). The pipeline transforms tweets data in near real time to points of events, and trajectories and flows of tweeters. It produces aggregate data for the most recent days of a fixed length at a constant interval for continuous monitoring of changes in flu risks and population mobility.

To achieve the overarching goal, five requirements were identified for the FluMapper data pipeline. The first requirement was real-time collection of tweets data for earlier detection of abnormal changes in the spatiotemporal patterns of flu risks and tweeters' mobility. The second requirement was the identification of tweet messages related to ILI for the estimation of flu risks and flu-related flows of Twitter users. The third requirement was the conversion of unstructured textual tweets into spatial data entities to which analytical methods for risk estimation and flow mapping can be applied. For risk estimation, tweets should be stored as spatial events with temporal references. For flow mapping, tweets were converted to a cube of origin-destination (OD) matrices where origins and destinations were the unit cells of the spatial grids that were hierarchically partitioned in a quadratic fashion (Figure 1). This conversion was introduced to facilitate rapid cartographic visualization of flows across multiple spatial scales. The fourth requirement was the use of a scalable data storage system due to the massiveness of the Twitter data stream and the need for historical monitoring of flu risks and tweeters' movements. The last requirement was to provide service-based access to the preprocessed data of tweet messages; it was expected that spatiotemporal analytical methods for big spatial data would be implemented as distributed web services since they tend to be both data- and computation-intensive, and parallel algorithms and distributed high-performance computing resources, such as grids and clouds, would enable the application of such analytical methods to large spatiotemporal data.

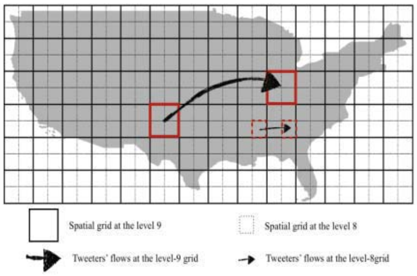

Spatial grid at the level 9
Spatial grid at the level 8
Tweeters' flows at the level-9 grid
Tweeters' flows at the level-8grid

Figure 1. Spatial grids used for aggregation of flow statistics

4.2 Architecture

Figure 2 shows the architecture of the FluMapper data pipeline. It consists of 1) four software components for data collection and processing, 2) a corpus of text files caching the raw tweets from Twitter, and 3) a set of distributed databases for managing spatial data entities of tweets and tweeters' flows. In particular, the four software components include the *crawler*, *tweet importer*, *cube generator*, and *data filter service*.

The *crawler* collected tweet messages from Twitter and cached the raw tweets as text files so as to enable historical data analysis. In particular, the *crawler* accessed Twitter's Streaming API and maintained a long-lived connection to it. Although Twitter also provides the Search API, it was not used in this study due to several limitations: 1) in most cases, it is slower than the Streaming API in the delivery of the latest tweets; 2) the Search API returns only recent tweets, i.e., up to 6~9 days old ones, necessitating special tools for longitudinal data collection; 3) the API also returns only a maximum of 1,500 tweets per search request; 4) the search algorithm focuses on relevance more than completeness, causing some omissions of tweets even though they meet search conditions; and 4) the quota for daily search is limited, thus complicating its use for real-time continuous surveillance of flu epidemics. In contrast, the Streaming API provides access to the end points of Twitter's public data streams. Once a client application is connected to the API, a feed of tweets is delivered continuously without any limits on the data-polling rate. Provided these pros and cons of Twitter APIs and the requirement of real-time data collection, the study used the Streaming API as the main data source for the FluMapper data pipeline. When the *crawler* accessed the Streaming API, it passed a spatial filter of North America, i.e., a bounding box ranging from (-167.276413, 5.499550) to (-52.233040, 83.162102). No other filters such as keyword-based ones were passed to the API so that a sample could be collected from the entire population of raw geotagged tweets. The resulting sample of raw tweets could then be used as a proxy for population at risk, an essential component for disease risk estimation [32]. ILI-related tweets, a proxy for disease cases, could also be identified from the raw tweets through keyword match or other text-mining techniques.

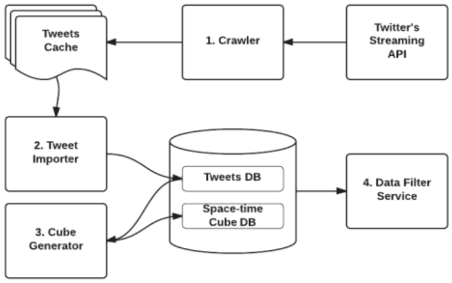


Figure 2. The Architecture of the FluMapper Data Pipeline

The second software component, the *tweet importer*, read raw tweets from the cache files and converted them into spatial events from which flu risks could be estimated. The *importer* first parsed individual tweets and filtered them out if they were retweets, had no information of their spatial footprints, and were posted outside of the conterminous US, which was the study region of the FluMapper data pipeline. The *importer* then determined the relevance of each tweet message to ILI by using keyword match. The keywords used for the match corresponded to 45 terms that referred to influenza-related names, symptoms, and medications, and were suggested in prior research to be appropriate for identifying influenza-related queries and conversations in web search and social media [11, 17, 26]. The terms included “flu”, "tamiflu", "h1n1", "swine", "influenza", "grippe", "sick", "illness", "coldness", "cold", "fever", "cough", and "headache." If a tweet message included one of these keywords, it was marked as flu-related. In addition, when the message was tagged with the coordinates of an exact geographic location rather than a place name, these coordinates were used to represent the messages as a spatial event. If only a place name was used for geographic referencing of the tweet message, the centroid of the bounding box of the place’s boundary was retrieved to transform the message into a spatial event. On average, more than 90 percent of the tweets in our database were tagged with exact geographic locations (Figure 5). In addition to geographic coordinates, a cell where each tweet was posted was identified in the finest level of the hierarchical spatial grids, and the cell’s ID was added as an attribute of the tweet in order to facilitate later computations of flow statistics of Twitter users. As the result of all these procedures, a tweet was converted into a record of a spatial event with five attributes (Figure 3). The final records of tweets were regularly loaded into *Tweets DB*, which supported scalable data storage through the use of MongoDB, a distributed, high-performance, NoSQL database management system. In support of near real-time surveillance of flu epidemics, the records in this DB were incrementally updated so that the DB retained only the data of the current sliding-time window.

While the transformation of tweets into spatial events allowed for flu risk estimation, it was not enough for understanding tweeters’ mobility. To collect data about the mobility of Twitter population, the FluMapper data pipeline traced the trajectories of Twitter users and kept track of their movements between two cells in the spatial grid at the finest level of the cube hierarchy. When a tweet message (***m***) is considered a location or a *presense* of a Twitter user *i* in a space-time setting and has an attribute *a*, it can be represented as follows:

$$\mathbf{m}_{i,t} = \{x_{i,t}, y_{i,t}, a_{i,t}\}$$

where t, x, y, and a represent a time point, x coordinate, y coordinate, and the attribute value of the tweet message, respectively.

The trajectory (***L***) of the Twitter user *i,* $\mathbf{L}_{i,}$ then becomes a chain of the locations of the tweet messages posted by the user *i*:

$$\mathbf{L}_i = \{\mathbf{m}_{i0}, \mathbf{m}_{i1}, \ldots, \mathbf{m}_{iT}\},$$

which means the user *i* tweeting at the location (x_{io}, y_{io}) in the beginning of the data collection ($t=0$), and moving to and tweeting at location (x_{i1}, y_{i1}) at time $t=1$. At the latest time point *T*, the user tweeted at (x_{iT}, y_{iT}). For each tweeter, the *tweet importer* collected all location data constituting his or her trajectory, and the relevance of a tweet message to ILI was added as an attribute of each location in the trajectory (see the *User trajectory* table in Figure 3).

According to [2], the trajectory data collected here are considered episodic movement data rather than continuous one, since tweeters post messages irregularly and the intermediate locations between two tweeting activities cannot be reconstructed easily through traditional interpolation methods. To reduce the side effects of such episodic movement data (e.g., missing data and spatiotemporal uncertainties of individuals’ specific moving paths), the study aggregated the collected trajectories data into flows of individual tweeters according to the number of *presences*[2]. Let’s suppose that the spatial grid at the finest level is $\boldsymbol{G_0}$ consisting of ***R*** rows and ***K*** columns. If $\boldsymbol{m}_{i,t}$ is within a cell at the *rth* row and *kth* column, i.e., *c(r,k)*, the cell id of $\boldsymbol{m}_{i,t}$ is equal to $(r-1)*\boldsymbol{K} + k$. Two consecutive locations in the trajectory $\boldsymbol{L}_i$ is referred to as a segment $\boldsymbol{S}_{i,t,t+1}$, and this segment can be represented as follows:

$$\mathbf{S}_{i,t,t+1}=\{\mathbf{m}_{i,t},\mathbf{m}_{i,t+1}\}=\{c(r,k),c(r',k')\}=\{\ (r-1)*\boldsymbol{K}+k\ ,\ (r'-1)*\boldsymbol{K}+k'\ \}$$

In the study, a flow was considered to occur if $r \neq r'$ or $k \neq k'$, i.e., if the tweeter *i* moved across two different cells in $\boldsymbol{G_0}$. While updating $\boldsymbol{L}_i$, the *tweet importer* inspected if the trajectory was extended across two different cells in $\boldsymbol{G_0}$, and if so, added a new record into the *Flow* table of the *Tweets DB* (Figure 3). As in the case of tweets, flows were retained only while their ending times were within the sliding time window.

The FluMapper data pipeline retrieved information of individual tweets, user trajectories, and tweeters’ flows by regular execution of the *tweet importer*. For efficient spatiotemporal exploration of flu risks and tweeters’ mobility across varying spatial scales, these individual-level data needed to be further aggregated along the hierarchical system of spatial grids. The *cube generator* in Figure 2 undertook this aggregation every three hours by querying tweets and flows data of the most recent seven days from the query initiation time and counting the numbers of raw tweets, flu-related tweets, tweeters, outflows, and inflows for each cell in the finest level of the spatial grids. The cube consisted of ten layers of spatial grids, and the finest level of the cube contained a spatial grid of 3072×7168 cells of about 30 arc-seconds (≈1 km) width. Once flow statistics were gathered at this level of the cube’s hierarchy, the statistics were recursively aggregated up to a higher level of spatial grid until the grid dimension was reduced to 6×14 cells (Figure 1). The results of the aggregation were stored in the *Space-time Cube DB* (Figure 4) as a new data collection for two days so that temporal dynamics in tweeters’ movements could be explored smoothly in a three-hour interval.

The primary goal of the FluMapper data pipeline was to provide real-time tweet geostream as inputs for advanced spatiotemporal data analyses such as inference-enabled kernel density estimation and one-to-all/all-to-all flow mapping. When these analytical methods are designed to handle massive geostreams like the tweets data collected by the pipeline, they often take the form of web services behind which distributed high-performance computing resources are utilized for parallel data processing, computation, and visualization. To facilitate the interface between these distributed analytical services and the pipeline’s databases, the *data filter service* was developed. At present, this element supports only temporal queries on the *Tweets DB*. In particular, the *filter service* returns the first tweets posted by each tweeter within a user-defined time period. The default period for the temporal query is a week from the query initiation time. The *data filter service* returns only the first tweets in order to eliminate repeated tweets that a single user posted about ILI. This data elimination was adopted under the assumption that the first flu-related tweet of each user would have stronger associations with the locations of flu outbreaks than later tweets. Reports from CDC follow a similar assumption by accounting only for the first visit of each patient within the period of a single episode of illness [1]. Although CDC considers a patient visit a new encounter of illness after more than six weeks has passed from the most recent encounter [19], the FluMapper data pipeline used a week as the default temporal filter because tweets data were found to show the highest level of correlation with CDC reports when a week was used to identify the first encounter of flu cases [1].

4.3 Databases

Two databases were used to store the data processed by the FluMapper data pipeline: the *Tweets* (Figure 3) and *Space-time cube DBs* (Figure 4). As discussed earlier, the *Tweets DB* contained three collections of tweets, user trajectories, and flows, respectively. Each record in the tweets collection had six attributes: tweet ID (*_id*), cell ID (*cell_id*), Twitter user ID (*user_id*), geographic location (*geo*), relevance to ILI (*isflu*), and timestamp (*time*). A user trajectory was a set of three lists containing the spatial (*geo*), temporal (*time*), and flu-relatedness (*flu*) of tweets posted by a tweeter. A record of a flow also consisted of six attributes: the ID of the flow (*_id*), the IDs of origin (*from_cell_id*) and destination (*to_cell_id*) cells, and the timestamps of the tweet messages at the origin (*from_time*) and destination (*to_time*) cells, and flu-relatedness (*isflu*) of the flow. If the tweet message either from the origin or the destination cell was about ILI, the flow was considered flu-related.

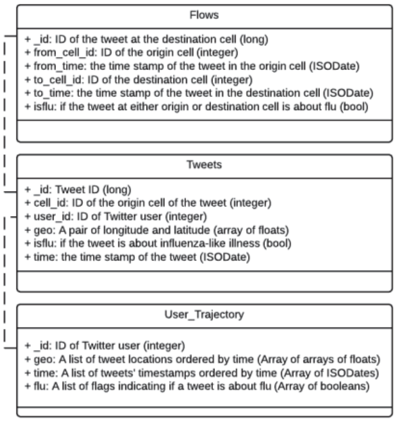

Flows
+ _id: ID of the tweet at the destination cell (long)
+ from_cell_id: ID of the origin cell (integer)
+ from_time: the time stamp of the tweet in the origin cell (ISODate)
+ to_cell_id: ID of the destination cell (integer)
+ to_time: the time stamp of the tweet in the destination cell (ISODate)
+ isflu: if the tweet at either origin or destination cell is about flu (bool)
Tweets
+ _id: Tweet ID (long)
+ cell_id: ID of the origin cell of the tweet (integer)
+ user_id: ID of Twitter user (integer)
+ geo: A pair of longitude and latitude (array of floats)
+ isflu: if the tweet is about influenza-like illness (bool)
+ time: the time stamp of the tweet (ISODate)
User_Trajectory
+ _id: ID of Twitter user (integer)
+ geo: A list of tweet locations ordered by time (Array of arrays of floats)
+ time: A list of tweets' timestamps ordered by time (Array of ISODates)
+ flu: A list of flags indicating if a tweet is about flu (Array of booleans)

Figure 3. The Tweets DB schema

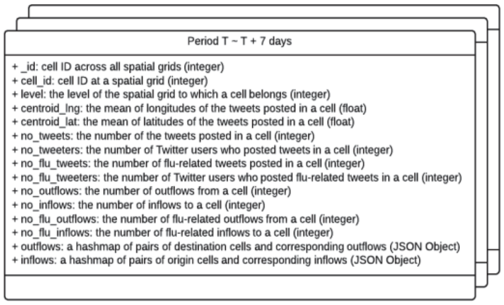

Period T ~ T + 7 days
+ _id: cell ID across all spatial grids (integer)
+ cell_id: cell ID at a spatial grid (integer)
+ level: the level of the spatial grid to which a cell belongs (integer)
+ centroid_lng: the mean of longitudes of the tweets posted in a cell (float)
+ centroid_lat: the mean of latitudes of the tweets posted in a cell (float)
+ no_tweets: the number of the tweets posted in a cell (integer)
+ no_tweeters: the number of Twitter users who posted tweets in a cell (integer)
+ no_flu_tweets: the number of flu-related tweets posted in a cell (integer)
+ no_flu_tweeters: the number of Twitter users who posted flu-related tweets in a cell (integer)
+ no_outflows: the number of outflows from a cell (integer)
+ no_inflows: the number of inflows to a cell (integer)
+ no_flu_outflows: the number of flu-related outflows from a cell (integer)
+ no_flu_inflows: the number of flu-related inflows to a cell (integer)
+ outflows: a hashmap of pairs of destination cells and corresponding outflows (JSON Object)
+ inflows: a hashmap of pairs of origin cells and corresponding inflows (JSON Object)

Figure 4. The Space-time cube DB schemaThe *Space-time cube DB* was composed of at most 16 data collections where each contained a cube of OD matrices based on the hierarchical spatial grids for a week period. A cell of an OD matrix corresponded to a record in a data collection. The record had 15 attributes: a unique ID of the current cell across multiple matrices (*_id*), a unique ID of the cell within an OD matrix (*cell_id*), the level of the OD matrix to which the cell belongs (*level*), the means of longitudes (*centroid_lng*) and latitudes (*centroid_lat*) of all tweets posted in the cell, the number of all (*no_tweets*) and flu-related tweets (*no_flu_tweets*) posted in the cell, the number of unique tweeters who posted raw (*no_tweeters*) or flu-related (*no_flu_tweeters*) tweets in the cell, the total numbers of all (*no_outflows*) and flu-related outflows (*no_flu_outflows*) at the cell, the total numbers of all (*no_inflows*) and flu-related (*no_flu_inflows*) inflows at the cell, all and flu-related outflows from the cell to each destination (*outflows*), and all and flu-related inflows from each origin to the cell (*inflows*).

4.4 Implementation

Multiple software tools were used to implement the FluMapper data pipeline. As discussed earlier, Twitter's Streaming API was employed for real-time collection of tweets data. The *crawler* was implemented as a Java-based daemon that interacted with the Streaming API through the Twitter4J[3] Version 2.2 library. The *tweet importer* and *cube generator* were developed as Python scripts using the PyMongo 2.4 library. The *data filter service* was also a Python-based application deployed as Common Gateway Interface. The *Tweets* and *Space-time Cube DBs* were managed in two separate MongoDB instances. MongoDB was chosen because it scales better than conventional RDBMS, is freely available, and is easier to use [3]. For improved scalability and rapid data query, the *Tweets* and *Space-time Cube DBs* were distributed across two separate computer nodes. Each database node has a single 2.40 GHz Intel(R) Xeon(R) CPU and 8 GB memory.

5. Illustration

This section illustrates the data collection and processing capacity of the FluMapper data pipeline and how the collected data can be used for exploratory analysis of flu risk indicators. The pipeline so far used a sliding window of 7-day length since our earlier experiments showed that a window size less than this length did not capture significant changes in tweeters' movements across varying spatial scales. Because of the use of the sliding window scheme, the illustrations here should be considered an average snapshot of the performance measures of the pipeline and a case study of tweet-based flu risks indicators.

5.1 Data collection and processing

The first measure of the data pipeline was how many tweets it collected and processed on a daily basis. Figure 5 shows daily volumes of tweet messages that were collected by the *crawler*, were imported into the *Tweets DB*, had geographic coordinates of the locations of tweet posting, and contained one or more keywords of ILI-related symptoms and names. For the period from May 23 to June 5, 2013, the *crawler* cached about 2.7 million tweets created in the North America area. About 25% of the collected raw tweets were filtered out because they were retweets, had inappropriate georeferences, or were posted outside the conterminous US. Consequently, an average of 2 million tweets was loaded into the *Tweets DB* per day, and because of the use of the 7-day sliding window, the tweets collection in the DB retained about 14 million records on average. As presented in Figure 5, about 92.6% of these records had information of exact geographic coordinates of tweet locations, and only 1.2% of them were found to be ILI-related. Despite little variations in the daily volumes of tweets, the statistics displayed in Figure 5 reveal that the FluMapper data pipeline had consistently dealt with large volumes of tweet geostreams, which would have been challenging to handle in conventional health GIS used for disease surveillance.

Starting its operation from February 16, 2013, the pipeline had collected a total of 2,807,234 user trajectories until June 5, 2013. The cumulative number of user trajectories had increased steadily, the average daily increment between May 29 and June 5, 2013 amounted to 18,927. Data of tweeters' movements or flows have been collected from these user trajectories. Figure 6 presents the temporal changes in the daily volumes of the flows. For the period

[3] http://twitter4j.org/en/index.html

from May 29 to June 5, 2013, the pipeline identified 0.25 million flows of Twitter users from a neighborhood of 1 km^2 to another. Not surprisingly, the Twitter population made more travels across neighborhoods during weekends (06/01-02) than weekdays. On average, tweeters posted messages containing ILI-related keywords during about 1% of these cross-neighborhood travels. The statistics of tweeters' trajectories and movements demonstrate the unique capability of the FluMapper data pipeline, i.e., the provision of large spatiotemporal data of the mobility of Twitter population. The small amount of flu-related flows may also indicate that SMG are specific enough to identify few flu cases when the flu season has actually ended.

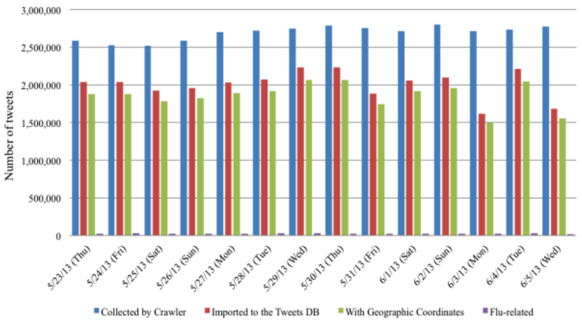


Figure 5. Daily volumes of tweets

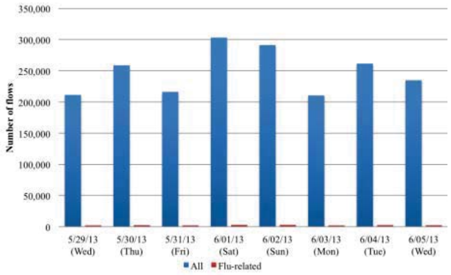


Figure 6. Daily volumes of flows

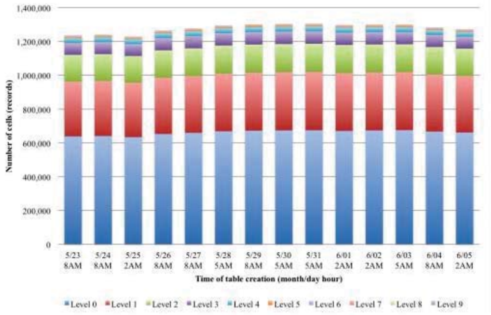


Figure 7. Sizes of a data collection in the Space-time cube DB

Figure 7 shows the number of data records containing aggregate statistics of tweet messages and tweeters' flows at each level of the hierarchical spatial grids. The average number of cells from which at least one tweet was posted during a week period was about 1.3 million across all levels of the grids. Because each dimension of a spatial grid reduced to half as the level increased, the cells at the level-0 grid occupied more than 50% of the entire documents in the cube. This pattern was consistently observed in the data tables of the *Space-time cube DB* due to the use of the 7-day sliding window scheme. On average, the *Space-time cube DB* retained 16 data collections across which the total number of records amounted to about 20.8 millions. These metrics of the cube DB indicates that multiscale spatiotemporal exploration of tweeters' mobility demands structured spatial data of a large size and that the design of the FluMapper data pipeline was efficient in addressing such demand.

The final measure we examined was the time spent on processing raw tweet messages in the *tweet importer*, collecting aggregate statistics of tweeters' flows in the *cube generator*, and retrieving the locations of unique tweeters through the *data filter service*. During the period from May 23 to June 5, 2013, the *cube generator* and *data filter service* were invoked every three hours concurrently. While these two components did not run, the *tweet importer* was called every seven minutes. Figure 10 shows the average processing times by the three components. The *tweet importer* spent about 1.5 ~ 8 minutes on processing 10,000 raw tweets and importing them into the *Tweets DB*. The average time for one execution of the *cube generator* was 2.3 hours with a standard deviation of 46 minutes. Since the cube generator produced data for about 1.3 million cells at a time, the execution time means it took about 1 minute to collect statistics for 10,000 grid cells. When the *data filter service* was used to obtain the first locations of Twitter users from a 7-day tweets dataset, the average run time of the service was 1.1 hours with a standard deviation of 20 minutes. Since the *data filter service* returned an average of 773,314 locations, its run time indicates that about 49 seconds was needed to identify the locations of 10,000 tweeters.

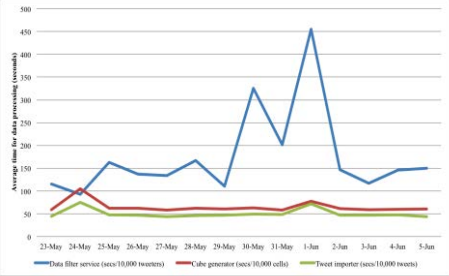


Figure 10. Average time for data processing

5.2 Analysis of flu risk indicators

Although efficient processing of massive SMG was the first concern in the design of the FluMapper data pipeline, the data content it provides was a crucial aspect as well. This section describes the distinctive characteristics of the data content with a focus on illustrating how the data can afford multiscale spatiotemporal exploration of flu risk indicators. A cross-sectional data set for the period of March 25 9AM to April 1 9AM, 2013 (GMT) was retrieved and used below for the illustration purpose.

Figure 8 provides a snapshot distribution of the Twitter users who posted at least one tweet in the study period. The total number of unique tweeters shown in this figure was about 701,450. Among these tweeters, 13.6% (95,233 users) posted at least one message including ILI-related keywords. The spatial distribution of the Twitter population qualitatively resembled the population density in US, showing higher densities of tweet messages in the eastern and coastal areas. Figure 9b shows the flu risk indicators that our application FluMapper (https://flumapper.org) produced by applying adaptive Kernel density estimation to the data shown in Figure 8. When compared with the indicators derived from the data of nine hours ago (Figure 9a), some local changes in the flu risk was noticeable in the areas of South Dakota, New Mexico, Mississippi, and Alabama. The changes in the vicinity of Alabama were especially notable since the CDC report of the same period indicated that the level of ILI-related activities was the highest in this state until about a week ago from April 1, 2013. Since then, no report has been available from CDC up to April 1, 2013. Yet, the data from the FluMapper data pipeline were collected continuously, thus enabling exploration of hourly changes in the flu risks through near-real-time applications of analytical methods for risk estimation.

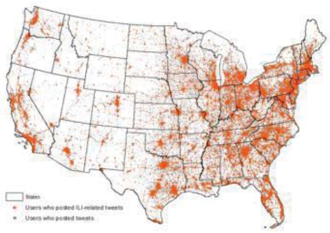
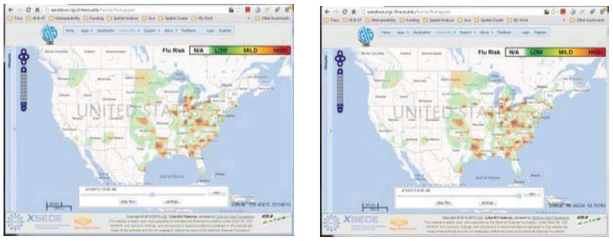


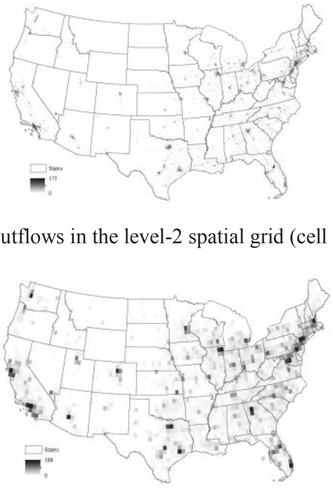
Figure 8. Spatial distribution of Twitter users (March 25 9AM ~ April 1 9AM, 2013)



(a) March 25 0AM ~ April 1 0AM (b) March 25 9AM ~ April 1 9AM

Figure 9. Temporal changes in flu risk indicators

Another type of content that the pipeline produced is the data of tweeters' movements or flows. As illustrated in Figure 10, the data from the pipeline facilitate spatial exploration of these flow statistics across varying spatial scales by providing the statistics measured at hierarchal spatial grids. The use of a sliding temporal window that moves forward every three hours also enabled temporal exploration of the changes in the movement patterns of Twitter users.



(a) Tweeters' outflows in the level-2 spatial grid (cell size: 4×4 km^2)

(b) Tweeters' outflows in the level-6 spatial grid (cell size: 64×64 km^2)

Figure 10. Spatial distributions of tweeters' outflows (March 25 9AM ~ April 1 9AM, 2013)

When combined with advanced techniques for spatiotemporal analysis and visualization such as flow mapping, the data of tweeters' movements may have the potential to shed lights on the associations between human mobility and the process of disease exposure, transmission, and outbreaks. For example, Figure 11 shows a flow map that was generated from the data of the study period. This figure suggests that soon after Alabama experienced the highest level of ILI activity between March 15 and 22, 2013, many tweeters traveled from Alabama to other states, particularly to those in the southeast and northeast regions. This kind of information may be of interest to public health practitioners who want to inform local health districts about spatial influences of neighboring areas that are subject to high levels of ILI activity. The FluMapper data pipeline enables near-real-time production of such information through spatiotemporal organization of geostream data from social media.

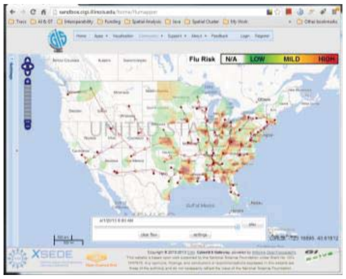


Figure 11. Outflows of Twitter users (March 25 9AM ~ April 1 9AM, 2013)

6. Conclusion

Motivated by the increasing interests in the use of social media data for the early detection of disease epidemics, this study investigated the problem of enabling spatiotemporal analysis of social media geostreams for the purpose of disease surveillance. To address the problem, the study developed a data pipeline solution for transforming social media geostreams into spatiotemporal data entities to which advanced techniques for spatiotemporal analysis (e.g., risk surface estimation and flow mapping) are easily applicable across varying ranges of spatial scales. The proposed solution differs from existing approaches to disease surveillance in that it supports near-real-time, multiscale spatiotemporal organization of geostream data for exploratory spatiotemporal analysis of disease-relevant conversations in social media. This focus of the solution on analytics also distinguishes it from current technologies for geostream processing, since much attention has not been paid in the latter to enabling multiscale spatiotemporal analysis of geostreams. Despite its focus on tweets, the solution is broadly applicable to other types of SMG.

While the illustrations of this study provided some measures of the efficiency and scalability of the FluMapper data pipeline, our future work will focus on making further enhancements of the data capacity and performance of the pipeline. We will expand the geographic coverage for data collection from the conterminous US to the North America continent. Such an expansion of data coverage will allow us to better understand the scalability challenge posed by the massive size of SMG. Emerging technologies for distributed data stream processing will be also investigated as potential solutions to scale up the data-processing capability of the pipeline in terms of both data capacity and processing speed.

7. ACKNOWLEDGMENTS

This work is partially funded by the National Science Foundation (NSF) under Grant Number OCI-1047916. Computational experiments used the Extreme Science and Engineering Discovery Environment (XSEDE) (resource allocation Award Number SES090019), which is supported by the National Science Foundation Grant Number OCI-1053575.

8. REFERENCES

[1] Achrekar, H. et al. 2011. Predicting flu trends using Twitter data. *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (2011), 702–7.

[2] Andrienko, N. et al. 2012. Visual analytics for understanding spatial situations from episodic movement data. *KI-Künstliche Intelligenz,* 26, 3 (2013), 241–251.

[3] Bifet, A. et al. 2011. Moa-TweetReader: Real-time analysis in Twitter streaming data. *Discovery Science* (2011), 46–60.

[4] Bonnet, L. et al. 2011. Reduce, you say: What NoSQL can do for data aggregation and BI in large repositories. *Database and Expert Systems Applications* (2011), 483–8.

[5] Castronovo, D.A. et al. 2009. Dynamic maps: a visual-analytic methodology for exploring spatio-temporal disease patterns. *Environmental Health*. 8, 61 (2009).

[6] Centers for Disease Control 1986. *Comprehensive plan for epidemiologic surveillance*. Centers for Disease Control.

[7] Chen, L. et al. 2010. Vision: Towards real time epidemic vigilance through online social networks: introducing SNEFT-social network enabled flu trends. *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* (2010), 4.

[8] Collier, N. and Doan, S. 2012. Syndromic classification of Twitter messages. *Electronic Healthcare*. P. Kostkova et al., eds. Springer. 186–95.

[9] Cromley, E.K. and McLafferty, S.L. 2012. *GIS and public health*. The Guilford Press.

[10] Fefferman, N.H. and Naumova, E.N. 2010. Innovation in observation: a vision for early outbreak detection. *Emerging Health Threats Journal*. 3, e6 (2010).

[11] Ghosh, D. and Guha, R. 2013. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*. 40, 2 (2013), 90–102.

[12] Ginsberg, J. et al. 2008. Detecting influenza epidemics using search engine query data. *Nature*. 457, 7232 (2008), 1012–4.

[13] Gomide, J. et al. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *ACM Web Science Conference (WebSci)* (2011), 1–8.

[14] Gonzalez, M.C. et al. 2008. Understanding individual human mobility patterns. *Nature*. 453, 7196 (2008), 779–82.

[15] Henry, J. V et al. 2004. Comparison of office visit and nurse advice hotline data for syndromic surveillance: Baltimore - Washington, DC, metropolitan area, 2002. *MMWR*. 53 Suppl, (2004), 112–6.

[16] Jacquez, G.M. et al. 2005. Design and implementation of a space-time intelligence system for disease surveillance. *Journal of Geographical Systems*. 7, (2005), 7–23.

[17] Keller, M. et al. 2009. Use of unstructured event-based reports for global infectious disease surveillance. *Emerging infectious diseases*. 15, 5 (2009), 689–95.

[18] Lampos, V. et al. 2010. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases*. J.L. Balcazar et al., eds. Springer. 599–602.

[19] Lazarus, R. et al. 2002. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerging infectious diseases*. 8, 8 (2002), 753-60.

[20] Li, L. et al. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*. 40, 2 (2013), 61–77.

[21] Lombardo, J.S. and Buckeridge, D.L. eds. 2007. *Disease surveillance: A public health informatics approach*. John Wiley & Sons, Inc.

[22] Magruder, S.F. et al. 2004. Progress in understanding and using over-the-counter pharmaceuticals for syndromic surveillance. *MMWR*. 24, 53 Suppl (2004), 117–22.

[23] Molinari, N.-A.M. et al. 2007. The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*. 25, 27 (2007), 5086–96.

[24] de Quincey, E. and Kostkova, P. 2010. Early warning and outbreak detection using social networking websites: The potential of Twitter. *Electronic Healthcare*. P. Kostkova et al., eds. Springer. 21–4.

[25] Padmanabhan, A. et al. 2013. FluMapper: An interactive CyberGIS environment for massive location-based social media data analysis. *Proceedings of the 2nd Conference of the Extreme Science and Engineering Discovery Environment (XSEDE): Gateway to Discovery* (2013), 1-2.

[26] Safko, L. 2010. *The social media bible: tactics, tools, and strategies for business success*. Wiley.

[27] Signorini, A. et al. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*. 6, 5 (2011), e19467.

[28] Thom, D. et al. 2012. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. *IEEE Pacific Visualization Symposium* (2012), 41-48.

[29] Tsou, M.-H. and Leitner, M. 2013. Visualization of social media: seeing a mirage or a message. *Cartography and Geographic Information Science*. 40, 2 (2013), 55–60.

[30] Tsui, F.-C. et al. 2003. Technical description of RODS: a real-time public health surveillance system. *Journal of the American Medical Informatics Association*. 10, 5 (2003), 399–408.

[31] Wagner, M.M. et al. 2004. Syndrome and outbreak detection using chief-complaint data - experience of the real-time outbreak and disease surveillance project. *MMWR*. 53, Suppl (2004), 28–31.

[32] Waller, L.A. and Gotway, C.A. 2004. *Applied spatial statistics for public health data*. Wiley-Interscience.

[33] Wang, S. et al. 2013. A CyberGIS Environment for Analysis of Location-Based Social Media Data. *Location-based Computing and Services*. A.K. Hassan and H. Amin, eds. CRC Press. In press.