Modeling Disease Transmission on a Location-Based Social Network

Kris Samala [*] ksamala@umd.edu
Carl Kingsford [*†] carlk@cs.umd.edu

May 18, 2012

Abstract

Location-based social networking services (LBSNs) provide an alternative form of gathering massive amounts of mobility data applicable to various fields of research that require human activity surveillance. This paper examines the relationship between locations extracted from observed patterns of LBSN usage and the effectiveness of the resulting network in tracing geographical spread of an infectious disease. Our results show that location connections formed through LBSNs demonstrate a reasonable approach in modeling trends in human mobility.

1 Introduction

The increasing prevalence of smart phones, tablets, and other mobile web devices has spurred a surplus of applications that track real-time geospatial user activity. Location-based social networks (LBSN), such as Gowalla [1] and Foursquare [2], allow users to share their whereabouts with friends and connect with new people of similar interests. These virtual communities yield large-scale datasets revealing a high resolution view of human mobility. In this paper, we study the potential use of the underlying structure of LBSNs for modeling the spread of an infectious disease within the United States. Disease dynamics may be generalized to encompass a wide variety of symptoms or tailored to a specific set of parameters for a particular contagion. We fit our model to simulate the transmission of the influenza virus.

Influenza ranks among the ten leading causes of death in the United States [28]. About three to five million severe cases are reported annually that lead to roughly 250,000 to 500,000 deaths worldwide [29]. The economic impact of a seasonal influenza epidemic costs an estimated total of $87.1 billion in the US every year attributed to medical costs, loss of earnings due to illness, and total deaths [19]. Vaccination is the most effective preventive measure against individual infection and epidemics [22]. However, the rapid evolution of its viral structure requires the vaccine's viral components to be updated regularly based on the recommendations of the World Health Organization (WHO) Global Influenza Surveillance Network. The selections are evaluated to match currently circulating strains, as well as emerging variants [12, 21]. However, due to manufacturing and regulation constraints, updated vaccines could take five months to a year to become available for mass distribution [1]. If a new strain emerges after the selection process ends, the vaccine may only provide partial immunity increasing the risk for a widespread pandemic.

Effectively predicting the spread of disease is crucial in minimizing seasonal epidemics and preventing a global pandemic. Several models have been developed to simulate the transmission of infections in a population. The susceptible-infected-recovered (SIR) model, introduced by Kermack and McKendrick in 1927, provides a generic mathematical approach to modeling epidemics. SIR assumes a population that is initially equally susceptible to a disease. Once an individual becomes ill, the disease spreads from the infected to the susceptible. Each infected person eventually retires from the infected stage either through recovery from illness or death [25]. The model has since been extended to take more complex forms. A contact network model provides a more realistic view of the dynamically changing interactions within a

[*]Department of Computer Science, University of Maryland College Park
[†]Center for Bioinformatics and Computational Biology, University of Maryland College Park
[1] gowalla.com
[2] foursquare.com

heterogenous population [27, 36]. Instead of assuming that the transmission probability is constant for all person-to-person contacts, the dynamic contact network allows asymmetrical transmission patterns that effectively represent higher risk subpopulations as well as variances in daily interactions.

Mathematical models tracing geographical spread of disease present a spatial view of epidemic progression [31, 32]. This feature is particularly important in identifying vulnerable regions in the case of an outbreak wherein necessary response strategies, such as school closings, travel restrictions, and quarantine, must be enacted. Airline travel is widely studied as the greatest facilitator of inter-regional influenza transmission [24, 4, 14]. The sharp increase in travel volumes and speed of inter-continental transport have had a significant influence on the spatio-temporal pattern of influenza transmission [4]. Airline flight itineraries are popular sources of mobility data. However, models based on this data set alone have several limitations. For example, air transportation generally covers only long-distance excursions. As a result, shorter range movement from other modes of transportation would be excluded [14, 24]. Mass gatherings, such as sporting events, or local social venues wherein a diverse group of people come into contact with each other are also not taken into consideration.

Alternative methods of representing human contact and mobility include telephone records, standard mail and courier services, financial transactions, and migration patterns [30]. Each has its strengths but often fail to capture some essential element that characterizes human behavior. The structure of offline social networks has shown great potential in replacing traditional contact network models for SIR disease simulations. A study on a flu outbreak at Harvard College showed that individuals at the center of a social network tend to acquire infections earlier and are more likely to spread the infection to their friends [8]. These nodes act as early detection sensors of an emerging outbreak and should be prime targets for vaccination. Collecting data for a similar contact network and observing the flow of infection on a global scale is a daunting, if not impossible, task. Thus, we propose an alternative method for simulating geographical disease propagation through user-provided activities with local and temporal detail.

Online social media boasts a rich collection of observable activities voluntarily and openly shared by its users. However, the application of online behavior to realistic events, as in person-to-person contact, is not directly obvious. Due to the global scope and accessibility of web applications, online usage appears to have the ability to transcend physical and geographical barriers. In tracing social contact, such intuition implies that online social networking sites do not effectively represent actual physical contact between its users. Nevertheless, research exploring comparisons between online and offline social networks have found that physical proximity, travel patterns, national borders, and common language all exhibit significant influence over which connections are formed in online social networks [9, 35].

Recent studies have employed online social media to track influenza-related content in Twitter tweets [6, 11, 13, 26, 34], Facebook status updates [13], blog posts [15], and Google search queries [18]. Through mining and analysis of vast amounts of online traffic, these portals serve as real-time surveillance systems that are able to monitor and predict flu level trends and the possibility of an emerging epidemic. Communication patterns between YouTube users have been used to substitute human interactions in simulating the 2009 H1N1 epidemic [30]. The results show that simulations on the YouTube-based network were able to predict the first day of incidence in each country better than estimates from air traffic data. However, projections for the total number of cases are less accurate and underestimates the actual number of incidences. Merging the advantages of utilizing user-generated content and social contact networks derived from online social media, we look at the unique attributes of LBSNs and the potential application in disease simulation.

Advancements in web-based mobile technology have boosted the popularity of "apps" that encourage users to share their daily experiences with staggering detail. Facebook, Foursquare, and Gowalla are only a few of the numerous applications that exploit the ease of capturing precise time and coordinates in recording user activity. LBSNs have recently been garnering attention in fields of research related to human mobility and activity patterns [7, 2, 20], friendship prediction [9], link prediction [33], and recommendation systems [3]. These studies reveal promising results in tracing user mobility and online behavior. Our work focuses on extracting a location-based contact network from large volumes of fine-grained data gathered from LBSNs to model real-time physical social events, such as the transmission of disease.

Results from our simulations approximate the geotemporal spread of influenza during the 2009 H1N1 epidemic. We observe that transition probabilities between cities in our Gowalla network model play an important role in estimating the variation of disease incidence across different regions. Our model can be applied in tracing human movement while observing the frequency and timing of user activities that

Table 1: Sample Data

City	State	Longitude	Latitude	Time
San Jose	CA	-121.89	37.32	2010-06-21 13:21:19
Philadelphia	PA	-75.02	39.94	2010-06-21 13:59:58
Minneapolis	MN	-93.24	44.85	2010-06-21 14:09:13
New York	NY	-73.61	40.73	2010-06-21 14:49:06
New York	NY	-73.98	40.72	2010-06-21 16:21:48
San Francisco	CA	-122.40	37.78	2010-06-21 16:27:12
New York	NY	-73.61	40.73	2010-06-21 19:26:03

contribute to infection spread. Through a close examination of LBSN use, our model captures the overall travel patterns of its users at both local and long-distance levels.

2 Methods

2.1 Gowalla Check-in Network

Gowalla is a mobile social networking service where users submit location information to connect with friends, log daily routines, and share exciting new places they are visiting. A *check-in* is created when a user posts information about his current location. A *spot* is a specific place or venue, such as a restaurant, a park, or an airport. Check-ins are annotated with the spot name and location, user name, date and time of check-in, as well as optional user comments and photos.

We obtain our data set of 9 million check-ins from Berjani and Strufe [3], originally crawled using the Gowalla API. The data consists of check-ins from spots located in the US from February 2009 to October 2010. It includes field values for user ID, location ID, timestamp, and the latitude and longitude coordinates of the place of check-in.

The data is transformed to follow a set of conventions useful in developing our model. To standardize the collection of cities, a list of the most populated US cities with the corresponding geographical coordinates was obtained from the US Census Bureau [5]. The average distance of a spot's coordinates to its listed city's coordinates is 9.07 miles with a standard deviation of 26.48 miles. If a spot is registered to a city that is not on the census list, the entry is updated with the closest listed city with a mean distance of 26.70 miles and standard deviation of 24.03 miles. Gowalla logs a timestamp for a check-in based on the local time on the device used. In order to obtain a chronological sequence of check-ins, we update all timestamps to be in Eastern Standard Time.

Under the assumption that each check-in reflects a user's location at the specified timestamp, we observe a discrepancy regarding the check-in history of certain users. For example, Table 1 shows the activity of a single user on June 21, 2010 with check-ins between distant cities that are only within a few hours apart. The fastest plausible mode of transportation is by commercial airplane, which travels approximately 600mph on average. However, a small subset of users have apparently had the ability to register check-ins to multiple distant places faster than the expected physical travel time. Another Gowalla data set [7], gathered during the same time period, exhibits similar issues. The problem may be due to flaws in Gowalla's logging implementation, special user privileges, or perhaps a glitch during the crawling process. Nevertheless, this pattern occurs for only 0.3% of the total users so we eliminate these users' activities from our sample.

We find that the number of check-ins from Austin, TX is more than double the number for the second most frequently checked-in city (New York, NY). We attribute this to the fact that Gowalla's headquarters are located in Austin and possibly have a strong influence over the use of the service within its home city. To adjust for this bias, we look at the three US cities that have the closest population size to Austin — San Francisco, CA, Columbus, OH, and Indianapolis, IN. We define $city_n$ to be the number of users whose majority of check-ins are in *city* and compute the average ratio between the number of users from Austin and the number of users from other cities:

$$\bar{r} \quad = \quad \frac{1}{|A|} \sum_{city \in A} \frac{city_n}{Austin_n}, \quad A = \{San\ Francisco,\ Columbus,\ Indianapolis\} \tag{1}$$

We use $\bar{r}$ in the next section to scale down the impact of Austin users to be roughly proportional to users in similarly populated cities in constructing our location-based contact network.

2.2 Weighted City-to-City Contact Network

We formulate our contact network by observing the path of cities visited by each user. Let C be the set of nodes for each US city and let E be the set of weighted directed edges representing transfer connections between cities. For every pair of city nodes c_i and c_j in C, let $e_{i,j} = (c_i, c_j)$ be the directed edge between these nodes. Let U be the set of all users, then for $u \in U$, $L_u = [c_s, \ldots, c_t]$, is list of check-ins such that c_s, $\ldots$, c_t are the cities visited by user u in chronological order. For every $e_{i,j} \in E$, we compute its weight,

$$w(e_{i,j}) \quad = \quad \epsilon + (\#\text{ of consecutive occurrences of } c_i \text{ and } c_j \text{ in } L_u, \quad u \in U) \tag{2}$$

Let U_a be the set of users whose majority of check-ins are located in Austin. Then for $u_a \in U_a$, the weight for each consecutive c_i, c_j in L_{u_a} is scaled by $\bar{r}$ from equation 1 when added to $w(e_{i,j})$.

The result is a strongly connected graph with 501 nodes and 251001 weighted edges. A transition probability matrix, T, is constructed such that $T[i, j]$ denotes the probability of moving from city c_i to city c_j.

$$T[i,j] \quad = \quad \frac{w(e_{i,j})}{w(e_{i,\cdot})}, \quad \text{where } w(e_{i,\cdot}) \text{ is the sum of the edge weights leaving city } i \tag{3}$$

2.3 Contact Probability Distribution

The rate of disease propagation from a single person is directly correlated with the number of people an infected individual interacts with. The number of contacts fluctuates from person to person every day while the exact definition of what constitutes sufficient physical contact to transmit an infection varies for different diseases. There are currently limited sources of social contact data applicable for infectious disease transimission. One of these include an investigation of how social contact patterns influence the temporal spread of influenza. The study conducted a survey to determine the daily number of physical and conversational contacts for each participant [16]. They observed that lower contact numbers during school holidays coincides with decreased influenza incidence rates and that the timing of school sessions promoted the spread of the 2009 H1N1 epidemic. We use the survey data from this study to determine how many contacts a sick individual will attempt to infect in our simulation.

2.4 Model Dynamics

Our disease simulation follows an extended SIR model featuring an additional compartment, latent, which differentiates a subgroup of the population who have been exposed to infection but are not yet able to transmit the disease [24]. Figure 1 shows the transitions between and within each disease stage. When a susceptible individual contracts an infection, he transitions from S to a latent phase L which could last up to 3 days. We define $L_{i,j}$ as the i^{th} stage out of j total days of latency, where $i = 1, \ldots, j$ and $j = 1, \ldots, 3$. After the ${L_{j,j}}^{th}$ stage, the individual transitions to one of $I_{1,k}$, where $k = 1, \ldots, 10$, as the first stage of infection of up to k total days. As he reaches the final day of infection at $I_{k,k}$, he finally moves to R and recovers.

Seasonal factors affecting an influenza epidemic causes a periodic peak and decline in the number of incidences each year. Cases are usually higher during the colder months of the winter when people are more likely to stay indoors and decrease as the temperatures become warmer during the summer. To account for the seasonal variability in the rate of infection, we apply a sinusoidal function to the transmission probability as determined by Edlund et al [17].

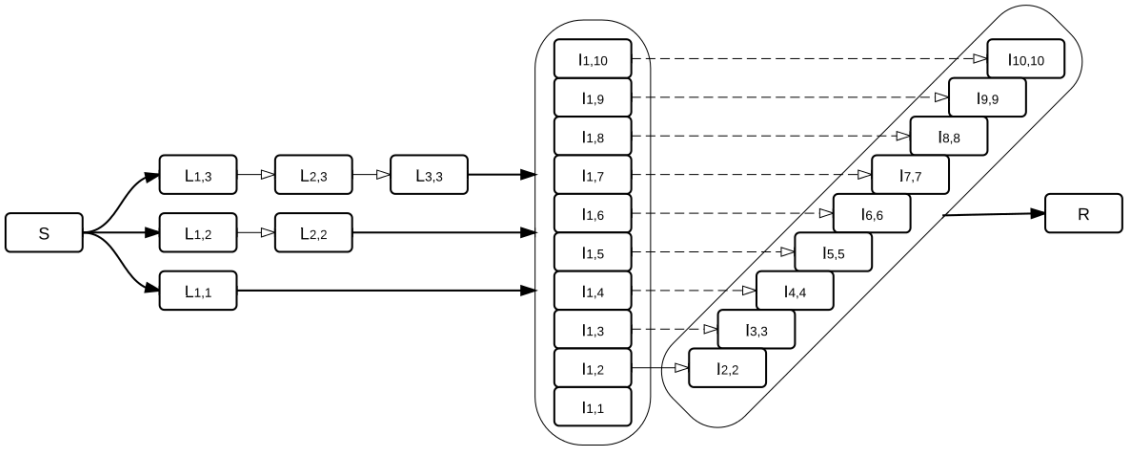

Figure 1: *SLIR Disease Model.*

Table 2: Simulation Parameters

Parameter	Definition
init	initial distribution of infected individuals for all cities
pop	population distribution for all cities
steps	number of time steps to run simulation
city	list of cities in network
contacts	contact probability distribution
p	disease transmission probability
T	city transition probability matrix
max_lat	maximum latency period
max_inf	maximum infectious period

Table 3: General Procedures

Procedure	Definition
GenAgent()	generates a new agent
Seasonal(t)	sinusoidal function of time t
Infect(x)	returns True with probability x
NextCity(T, c_i)	returns a city c_j with probability T[i,j]

2.4.1 Implementation

The simulation is executed with parameters and procedures listed in Tables 2 and 3. A distribution of influenza incidences per state is selected from Google Flu Trends data [23] to seed the simulation. The number of infected individuals in a state is randomly distributed among its cities for all states to obtain *init*. The overall population distribution among all cities in *pop* is acquired from the US Census Bureau [5]. Both *init* and *pop* are arrays of integers whose indices correspond to *cities*. *Simulation* initializes the set of agents for each disease compartment, S, L, I, R, and determines the length of infectious period for each infected agent. A call to *Spread* updates all sets of agents at every time step.

```
SIMULATION(init, pop, steps)
 1  S, L, I, R ← ∅
 2  for i ← 0 to len(city)
 3      do
 4          for n ← 1 to (pop[i] − init[i])
```

```
5            do agent ← GenAgent()
6               agent.city ← city[i]
7               agent.pd ← 0
8               S.add(agent)
9         for n ← 1 to init[i]
10           do agent ← new agent
11              agent.city ← city[i]
12              agent.pd ← random(1, max_inf)
13              I.add(agent)
14 for t ← 1 to steps
15    do SPREAD(p, contacts, T, max_lat, max_inf, t)
```

Spread simulates the transmission of disease from agents in I to agents in S. An infected agent comes into contact with c individuals, randomly chosen from *contacts*. The agent attempts to infect c susceptibles in its current city c_i with season-influenced transmission probability $Seasonal(p, t)$ then updates its location according to the city contact network. The remaining latency and infectious period for each agent in L and I, respectively, are decremented to reflect the changing state of infection as time passes. Once the remaining period reaches 0, the agents transition toward the next phase (L → I or I → R), and the infectious period for each previously latent individual is set.

```
SPREAD(p, contacts, T, max_lat, max_inf, t)
 1  for agent in I
 2     do c ← random(contacts)
 3        for k ← 1 to c
 4           do if Infect(Seasonal(p, t))
 5                 then infected_agent ← S.getAgentFromSameCityAs(agent)
 6                      infected_agent.pd ← random(1, max_lat)
 7                      S.remove(infected_agent)
 8                      L.add(infected_agent)
 9        agent.city ← NextCity(T, agent.city)
10        agent.pd ← agent.pd − 1
11        if agent.pd ← 0
12          then I.remove(agent)
13               R.add(agent)
14 for agent in L
15    do agent.pd ← agent.pd − 1
16       if agent.pd ← 0
17         then agent.pd ← random(1, max_inf)
18              L.remove(agent)
19              I.add(agent)
```

The procedure implements each set of agents, S, L, I, R, as vectors with the number of agents in each city. The vector for L is further subdivided such that $L = \{L_{i,j}\}$ for $i = 1, \ldots, j$ and $j = 1, \ldots, max\_lat$. Similarly, $I = \{I_{k,l}\}$ for $k = 1, \ldots, l$ and $l = 1, \ldots, max\_inf$. The counts on all vectors are iterated over to simulate the actions for each agent and updated as events unfold. Our simulations run for 365 time steps with varying p values and seed *init* with random pre-flu season estimates from Google Flu Trends. A simulation outputs a $steps \times len(city)$ matrix, M, that contains the number of infected agents at each time step for every city on our network.

2.5 Comparing with Google Flu Trends

Google Flu Trends approximates the proportion of influenza-like illness (ILI) cases from aggregated flu-related queries on the web [23]. The study [18] revealed that the frequency of certain ILI-related search terms match the seasonal fluctuation of influenza reports indicating a good predictor for flu activity across various regions.

The Centers for Disease Control and Prevention (CDC) provides weekly surveillance reports with FluView based on observed ILI-related outpatient visits [10]. However, data collected for the surveillance system can be delayed by up to a few weeks. Google Flu Trends offers daily estimates of flu activity by monitoring the occurrences of ILI-related search terms. They tested the top 50 million queries in their database and developed a linear model that produced a mean correlation of 0.9 with CDC observed ILI reports.

We compare our results with Google Flu Trends' weekly US influenza estimates from September 2003 to May 2012. The counts of infected agents in M are summed every seven days to obtain a 52-week tally normalized by each state total. These are evaluated against every 52-week subset of the Flu Trends data to find a closest match. A score for each pairing is determined by computing the average Euclidean distance, $\bar{d}$, between corresponding column vectors for each state in the two matrices. The lowest distance matchings are evaluated on a state-by-state level for further analysis.

3 Results

Simulation results for the Gowalla network model fit closest with Google Flu Trends estimates for May 2009 to May 2010 with $\bar{d} = 0.22$. We find that our model approximates the timing and activity levels for each state in the continental US throughout the flu season, with the exception of Alaska, Vermont, West Virginia, and Wyoming, for which the Gowalla network did not include any check-in data. We developed two baseline comparison networks to evaluate the strength of our model. The first is a permuted version of the Gowalla network wherein a random permutation of the probability vector for a city's connections to all other cities is substituted in the transition probability matrix T for all cities. The other is a randomized contact network which draws a random value from a Gaussian distribution as the probability that an agent will move from one city to another. Simulations using the same initial parameters as the Gowalla network results in $\bar{d}$ values for the permuted and random networks of 0.24 and 0.83, respectively, when compared with Google Flu estimates during the same flu season.

Figure 2 exhibits weekly influenza activity heat maps for Google Flu data and the simulation results for the Gowalla model, permuted, and randomized contact networks. The x-axis lists the 52-week simulation period while the y-axis lists the US states observed. Values for each tile are obtained from the percentage of total incidences for a state at a given week normalized across all states for that week. The weekly heat map estimates a general trend of how the activity levels of each state influence the prevalence of flu on a national level. The Gowalla network model follows a similar pattern of varying flu levels for each state throughout the year as Google Flu. The permuted network shows little variation while the randomized network is mostly uniformly distributed.

Meanwhile, Figure 3, displays a geographical snap shot of flu activity across the US for the four different sets. The geographic heat maps show the relative distribution of influenza incidences across the region for the week of October 11, 2009. Google Flu and Gowalla network highlight different states with high activity but follow a general overall distribution of the various ranges of flu levels. As evident in the weekly heat maps, the permuted and randomized networks show a much limited distribution of values in comparison. Specifically, the randomized network exhibits a sparse distribution of incidences for only a few select states. The results entail that the Gowalla network model has an observable impact on the geo-temporal spread of disease. The actual distribution of which state induces increased prevalence on another state differs as a result of bias with the manner of Gowalla service use which may distort the representation of certain states.

The overall proportion of incidence during the 52-week period, shown in Figure 4, exhibits high correlation between the three simulated models, with Euclidean distances of less than 0.05 from each other and approximately 0.14 against Google Flu. The pattern may be due to the applied seasonal factor varying the transmission probabilities over time with a sine wave function. This indicates that the total incidences on the national level is unaffected by the city-to-city transition probabilities since individual contact and transmission rates are sampled from the same distribution.

4 Discussion

Our observations demonstrate that the Gowalla network model approximates the geographical propagation of disease as a consequence of human mobility patterns. The location-based contact network relies on the

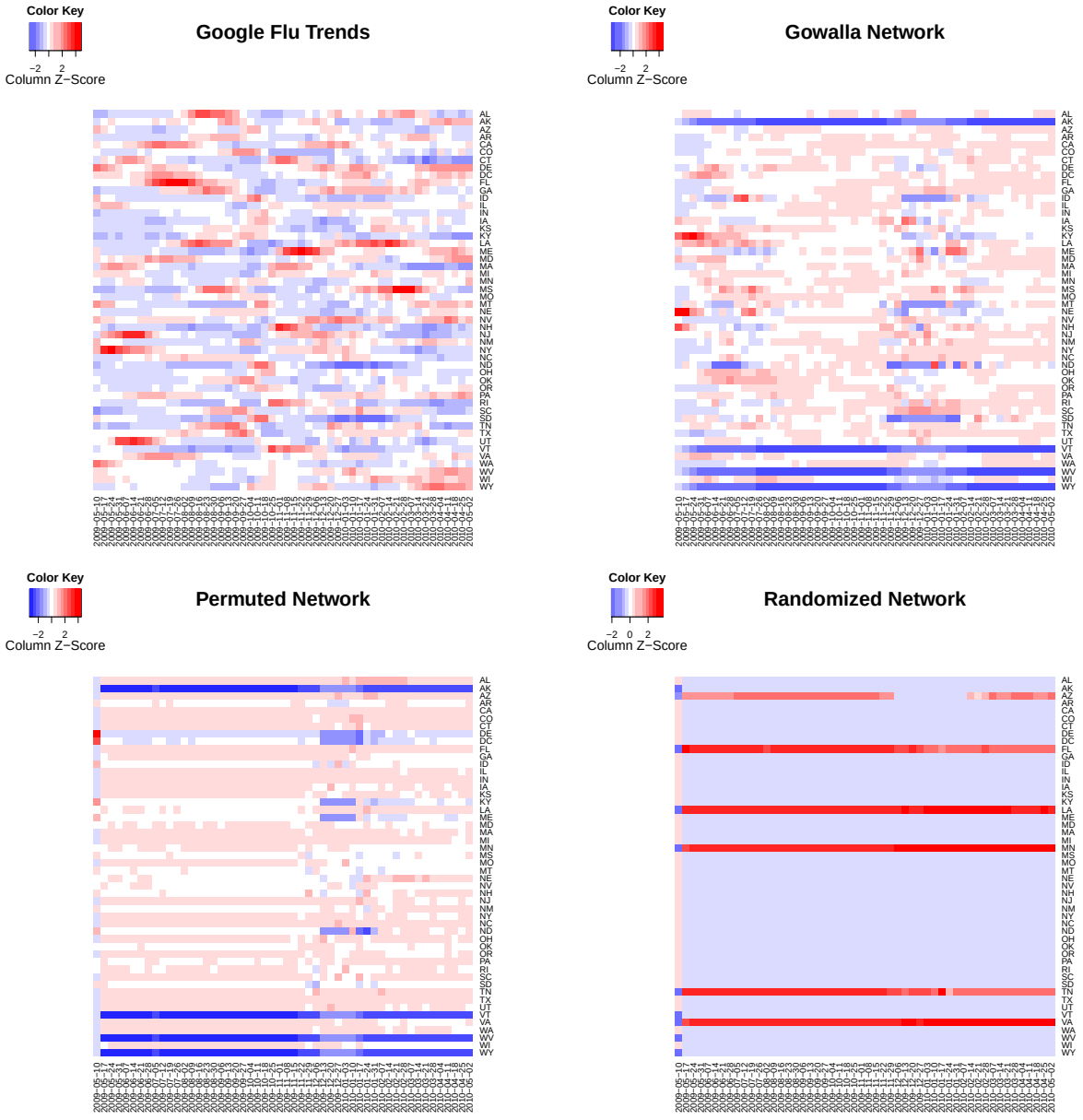

Figure 2: *Week-by-week Heat Maps.*

check-in activity of Gowalla users to detect movement from place to place. Infection spread is simulated over the network using established disease model parameters to determine spatio-temporal epidemic progression. Although the model uses a unique measure of movement through LBSNs, our results show that the transition probability distribution computed from our network is indicative of actual travel probabilities influenced by local and long-distance transportation methods.

It is worth noting that the Gowalla state-by-state distance comparisons against Google Flu estimates best match the 2009 H1N1 flu pandemic. This suggests that our model may be more appropriately used for high incidence disease outbreaks than seasonal epidemics. The disease parameters used may have played a role in the emergence of such performance, however, the random and permuted networks that incorporated the same parameters did not display the same behavior. If the model is updated to reflect better representation of regional activity through the use of larger and more up-to-date collections of check-in information, our

GEOGRAPHICAL HEAT MAPS

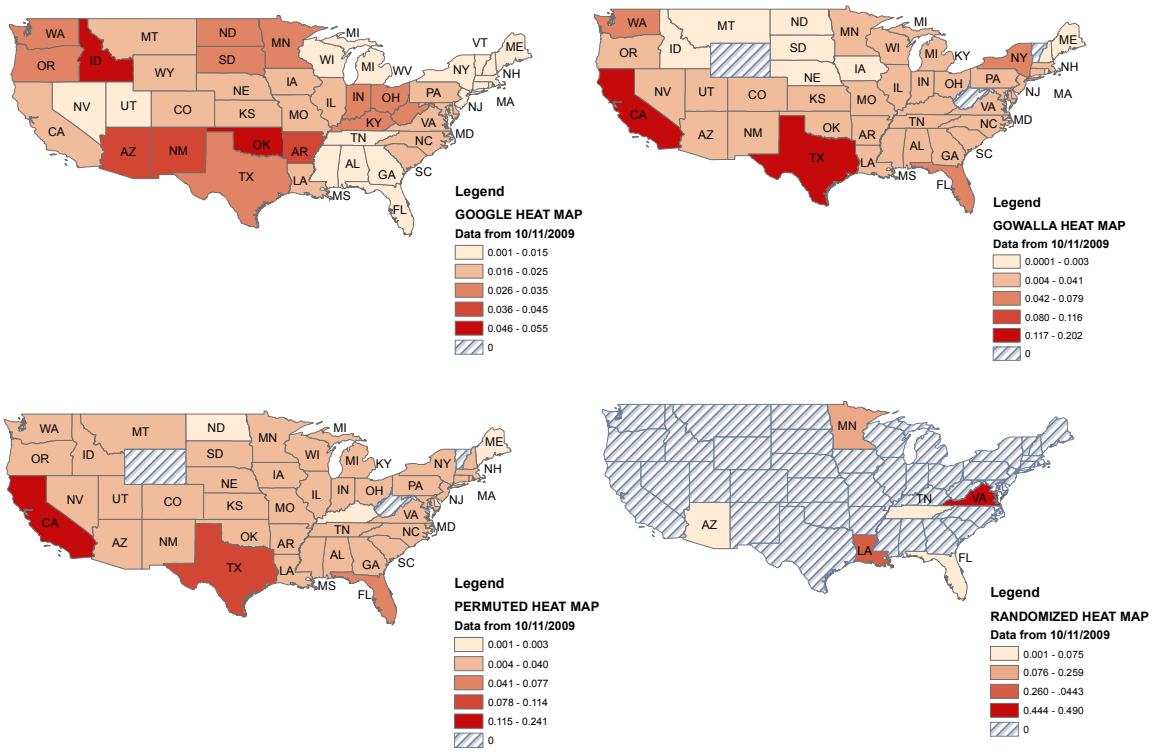

Figure 3: *Geographical Heat Maps.*

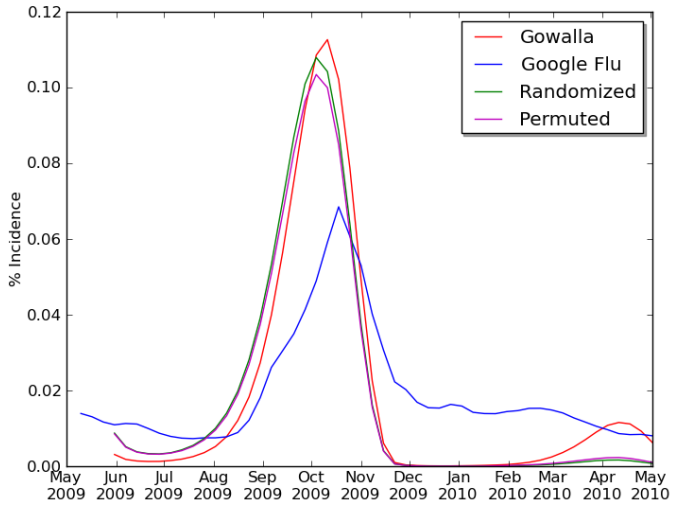

Figure 4: *Influenza Overall Estimates.*

simulations may be able to provide real-time metrics for user mobility and contact.

Location-based social networks offer intriguing insights into the mobility pattern of its users. It demonstrates behavioral trends through the timing and frequency of check-ins to favorite spots and visits to new places. The large-scale availability of such information provides an interesting alternative to tracing human activity. However, the power and shortcomings of the data lie on its dependence to self-reported user activity. On one hand, it allows people to conveniently record accurate details of day-to-day excursions. Instead of relying on recalling a busy day's worth of commute when filling out survey forms, one can simply press a button on a device that is always at hand. However, the types of individuals that these services cater to may be limited to a particular niche. Although the use of these applications has grown dramatically over the last few years, there remain large sections of the population that are excluded from LBSN data sets. Older generations may not be accustomed to using new technologies and younger children have yet to acquire sole access to these gadgets. Regardless, LBSN captures an essential element of human behavior that is lacking on conventional surveillance methods thanks to its unique ability of reporting immediate and precise information regarding a user's whereabouts.

References

[1] World Health Organization: Global Alert and Response Briefing Notes. Pandemic influenza vaccine manufacturing process and timeline. `http://www.who.int/csr/disease/swineflu/notes/h1n1_vaccine_20090806/en/`. Accessed April 2012.

[2] A. Bawa-Cavia. Sensing the urban: Using location-based social network data in urban analysis, June 2011. The First Workshop on Pervasive Urban Applications (PURBA).

[3] B. Berjani and T. Strufe. A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, 2011.

[4] J. Brownstein, C. Wolfe, and K. Mandl. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Medicine*, 3:1826–1835, October 2006.

[5] US Census Bureau. 2010 Census Data. `http://2010.census.gov/2010census/`. Accessed April 2012.

[6] C. Chew and G. Eysenbach. Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS ONE*, 5, November 2010.

[7] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1082–1090, 2011.

[8] N. Christakins and J. Fowler. Social network sensors for early detection of contagious outbreaks. *PLoS ONE*, 5, September 2010.

[9] J. Cranshaw, E. Toch, Hong J., Kittur A., and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of Ubicomp '10*, pages 119–128, 2010.

[10] CDC Influenza Division. FluView: A weekly influenza surveillance report. `http://www.cdc.gov/flu/weekly/`. Accessed April 2012.

[11] Achrekar H. et al. Predicting flu trends using Twitter data. In *First International Workshop on Cyber-Physical Networking Systems*, pages 713–718, 2011.

[12] Barr et al. Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: Basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009–2010 Northern Hemisphere season. *Vaccine*, 28:1156–1167, 2010.

[13] Chen L. et al. Vision: Towards real time epidemic vigilance through online social networks. In *ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*, June 2010.

[14] Colizza V. et al. The role of airline transportation network in the prediction and predictability of global epidemics. In *Proceedings of the National Academy of Sciences of the United States*, volume 103, pages 2015–2020, February 2006.

[15] Corley C. et al. Monitoring influenza trends through mining social media. In *International Conference on Bioinformatics and Computational Biology*, July 2009.

[16] Eames K. et al. Measured dynamic social contact patterns explain the spread of H1N1v influenza. *PLoS Computational Biology*, 8, March 2012.

[17] Edlund S. et al. A spatio-temporal model for influenza. *Electronic Journal of Health Informatics*, 6, 2010.

[18] Ginsberg J. et al. Detecting influenza epidemics using search engine query data. *Nature*, 457, February 2009.

[19] Molinari et al. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25:5086–5096, March 2007.

[20] Noulas A. et al. An empirical study of geographic user activity patterns in Foursquare. In *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media*, July 2011.

[21] Russell et al. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26S:D31–D34, July 2008.

[22] Centers for Disease Control and Prevention: Advisory Committee on Immunization Practices. Prevention and control of influenza with vaccines. *Morbidity and Mortality Weekly Report*, pages 1–62, August 2010.

[23] Google.org. Flu Trends. `http://www.google.org/flutrends/`. Accessed April 2012.

[24] R.F. Grais and Ellis J.H. et al. Modelling the spread of annual influenza epidemics in the US: The potential role of air travel. *Health Care Managetent Science*, 7:127–134, 2004.

[25] W.O Kermack and A.G. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, volume 115, pages 700–721, August 1927.

[26] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *2nd International Workshop on Cognitive Information Processing*, 2010.

[27] L.A. Meyers, M.E.J. Newman, and B. Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240:400–418, 2006.

[28] S. Murphy, J. Xu, and K Kochanek. Deaths: Preliminary Data for 2010. *National Vital Statistics Report*, 60(4), January 2012.

[29] World Health Organization. Fact Sheets: Influenza. `http://www.who.int/mediacentre/factsheets/2003/fs211/en/`. Accessed April 2012.

[30] J. Osth, T. Niedomysl, and B. Malmberg. Using data from social networking sites to predict the spread of pandemic influenza. In *Internation Meeting on Emerging Diseases and Surveillance*, 2011.

[31] L. Rvachev and I. Longini. A mathematical model for the global spread of influenza. *Mathematical Biosciences*, 75:3–22, 1985.

[32] L. Sattenspiel and C. Simon. The spread and persistence of infectious diseases in structured populations. *Mathematical Biosciences*, 90:341–366, 1988.

[33] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054, 2011.

[34] A. Signorini, A.M. Segre, and P. Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic. *PLoS ONE*, 6, May 2011.

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

[36] E. Volz and L.A. Meyers. Susceptible-infected-recovered epidemics in dynamic contact networks. In *Proceedings of The Royal Society B*, volume 274, pages 2925–2933, September 2007.