Predicting Disease Transmission from Geo-Tagged Micro-Blog Data

Adam Sadilek
Department of Computer Science
University of Rochester
Rochester, NY 14627
sadilek@cs.rochester.edu

Henry Kautz
Department of Computer Science
University of Rochester
Rochester, NY 14627
kautz@cs.rochester.edu

Vincent Silenzio
School of Medicine and Dentistry
University of Rochester
Rochester, NY 14627
v.m.silenzio@rochester.edu

Abstract

Researchers have begun to mine social network data in order to predict a variety of social, economic, and health related phenomena. While previous work has focused on predicting aggregate properties, such as the prevalence of seasonal influenza in a given country, we consider the task of fine-grained prediction of the health of specific people from noisy and incomplete data. We construct a probabilistic model that can predict if and when an individual will fall ill with high precision and good recall on the basis of his social ties and co-locations with other people, as revealed by their Twitter posts. Our model is highly scalable and can be used to predict general dynamic properties of individuals in large real-world social networks. These results provide a foundation for research on fundamental questions of public health, including the identification of non-cooperative disease carriers ("Typhoid Marys"), adaptive vaccination policies, and our understanding of the emergence of global epidemics from day-to-day interpersonal interactions.

Introduction

Recent work has demonstrated that micro-blogging data can be used to predict a variety of phenomena, including movie box-office revenues (Asur and Huberman 2010), elections (Tumasjan et al. 2010), and flu epidemics (Lampos, De Bie, and Cristianini 2010). Most research to date has focused on predicting aggregate properties of the population from the activity of the bloggers. A different kind of problem one can pose, however, is to predict the behavior or state of *particular individuals* within the social network. For instance, one could try to predict whether a person will go to a movie or vote for a particular candidate based on micro-blog data. The individual's own data may or may not be accessible. At one extreme, the task is to predict his behavior or state by considering only data from other people. For example, Sadilek, Kautz, and Bigham (2012) show that a person's location can be predicted with a high degree of accuracy based on only the geo-tagged posts (a.k.a. tweets) of his friends on Twitter.

This paper explores fine-grained prediction of the *health* of individuals on the basis of such social network data—an important instance of the general problem of modeling dynamic properties of participants in large real-world social networks. We begin by building upon previous work on classification of health-related text messages (Culotta 2010; Paul and Dredze 2011a; Sadilek, Kautz, and Silenzio 2012), to learn a robust SVM classifier that infers the health state of a person based on the content of his tweets. We then learn a conditional random field (CRF) model that *predicts* an individual's health status, using features derived from the tweets and locations of other people. Performance of the CRF is significantly enhanced by including features that are not only based on the health status of friends, but are also based on the estimated encounters with already sick, symptomatic individuals in the dataset, including non-friends. Thus, the model is able to capture the role of *location* in the spread of an infectious disease, the impact of the *duration* of co-location on disease transmission, as well as the *delay* between a contagion event and the onset of the symptoms. Using the Viterbi algorithm to infer the most likely sequence of a subject's health states over time, we are able to predict the days a person is ill with 0.94 precision and 0.18 recall. These results far outperform alternative models.

This work is an important step towards the development of automated methods that identify disease vectors, trace the transmission between concrete individuals, and ultimately help us understand and predict the spread of infectious diseases with fine granularity. It provides a foundation for research on fundamental questions of public health, such as: How does an epidemic on a population scale emerge from low-level interactions between people in the course of their everyday lives? Can we identify a potentially non-cooperative individual who is a vector of a dangerous disease, *i.e.*, a "Typhoid Mary"? What is the interaction between friendship, location, and co-location in the spread of communicable diseases?

Our results also prove useful for deploying sickness prevention resources, and for applications that help an individual maintain his or her health. For example, a person predicted to be at high risk of the flu could be specifically encouraged to get the flu vaccine, and recommendations can be made about which places pose a high risk of getting infected. Finally, the kinds of models we explore are not limited to the health domain. The close relationship between the spread of disease and information in general is well known (Easley and Kleinberg 2010). For example, by changing the map-

New York City Dataset	
Unique users	632,611
Unique geo-active users	6,237
Tweets total	15,944,084
GPS-tagged tweets	4,405,961
GPS-tagged tweets by geo-active users	2,535,706
GPS-tagged tweets by geo-active users that show a symptom of an illness	2,047
Distinct visited locations	57,109
“Follows” relationships between geo-active users	102,739
“Friends” relationships between geo-active users	31,874

Table 1: Summary statistics of the data collected from NYC. Geo-active users are ones who geo-tag their tweets relatively frequently (more than 100 times per month). Note that following reciprocity is about 31%, which is consistent with previous findings (Kwak et al. 2010). The number of distinct visited locations is calculated as the number of cells (100 by 100 meters) of the NYC grid that have been visited by at least one geo-active individual.

ping from text to features, the same approach can be used to model and predict the transmission of political ideas, purchasing preferences, or practically any other behavioral phenomena.

The Data

Our experiments are based on data obtained from Twitter, a popular micro-blogging service where people post message updates at most 140 characters long. The forced brevity encourages frequent mobile updates, as we show below. Relationships between users on Twitter are not necessarily symmetric. One can follow (subscribe to receive messages from) a user without being followed back. When users do reciprocate following, we say they are *friends* on Twitter. There is anecdotal evidence that Twitter friendships have a substantial overlap with offline friendships (Gruzd, Wellman, and Takhteyev 2011). Twitter launched in 2006 and has been experiencing an explosive growth since then. As of June 2011, over 300 million accounts are registered on Twitter.

Using the Twitter Search API[1], we collected a sample of public tweets that originated from the New York City (NYC) metropolitan area. The collection period was one month long and started on May 18, 2010. Using a Python script, we periodically queried Twitter for all recent tweets within 100 kilometers of the NYC city center. Altogether, we have logged nearly 16 million tweets authored by more than 630 thousand unique users (see Table 1). To put these statistics in context, the entire NYC metropolitan area has an estimated population of 19 million people.[2] Since this work studies the effects of people’s location and co-location on disease transmission, we concentrate on accounts that posted more than 100 GPS-tagged tweets during the one-month data collection period. We refer to them as *geo-active users*, and our dataset contains 6,237 such individuals.

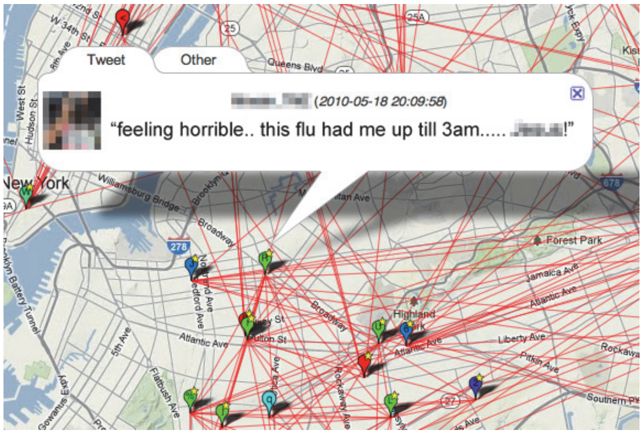


Figure 1: Visualization of a sample of friends in New York City. The red links between users represent friendships, and the colored pins show their current location on a map. We see the highlighted person X complaining about her health, and hinting about the specifics of her ailment. This work investigates to what extent can we predict the day-to-day health of individuals by considering their physical encounters and social interactions with people like X.

Methodology and Models

Given that five of your online friends have flu-like symptoms, and that you have recently met eight people, possibly strangers, who complained about having runny noses and headaches, how accurately can we predict that you will soon become ill as well? In the remainder of this paper, we propose and evaluate a model that provides answers to such questions across a large sample of people participating in online social media (see Fig. 1).

In this section, we first review our method for automatic detection of Twitter messages that suggest the author contracted an infectious disease[3] (Sadilek, Kautz, and Silenzio 2012). We then develop a CRF model that leverages the labeled tweets and makes accurate predictions about people’s health state.

Detecting Illness-Related Messages

In order to train and evaluate a predictive model of personal health, we first need to identify ill individuals, and estimate the time when they became contagious. We focus on self-reported symptoms and complaints that appear in the text of Twitter status updates. Our prior work has shown we can identify them with high precision as well as high recall, even though such messages are rare (Sadilek, Kautz, and Silenzio 2012). We achieve this by learning a linear support vector machine (SVM) binary classifier C_f while directly optimizing the area under the ROC curve (Joachims 2005). This SVM is robust even in the presence of strong class imbalance, where for every health-related message there are more than 1,000 unrelated ones. This is a necessary precondition

[1] http://search.twitter.com/api/

[2] http://www.census.gov/popest/metro/

[3] In this paper, such diseases include those with symptoms that overlap with, but are not necessarily limited to, influenza-like illness (http://en.wikipedia.org/wiki/Influenza-like_illness).

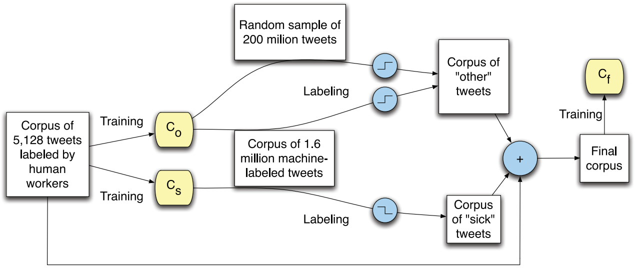

Random sample of 200 milion tweets
Corpus of "other" tweets
Labeling
Training
Corpus of 5,128 tweets labeled by human workers
C_o
Corpus of 1.6 million machine-labeled tweets
+
Final corpus
C_f
Training
C_s
Labeling
Corpus of "sick" tweets

Figure 2: A diagram of our cascade learning of our SVM classifier C_f that we use to detect tweets indicating an infectious sickness of the author. The _|‾ and ‾|_ symbols denote thresholding of the classification score, where we select the bottom 10% of the scores predicted by C_o (*i.e.*, tweets that are normal with high probability), and the top 10% of scores predicted by C_s (*i.e.*, likely "sick" tweets).

Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
sick	0.9579	sick of	0.4005
headache	0.5249	you	0.3662
flu	0.5051	lol	0.3017
fever	0.3879	love	0.1753
feel	0.3451	i feel your	0.1416
coughing	0.2917	so sick of	0.0887
being sick	0.1919	bieber fever	0.1026
better	0.1988	smoking	0.0980
being	0.1943	i'm sick of	0.0894
stomach	0.1703	pressure	0.0837
and my	0.1687	massage	0.0726
infection	0.1686	i love	0.0719
morning	0.1647	pregnant	0.0639

Table 2: Example positively and negatively weighted significant features of our SVM model C_f.

for further progress, as false negatives and false positives cannot be traded-off against each other in this domain—they both carry equal importance. In this work, we use C_f to distinguish between tweets indicating the author is afflicted by an infectious ailment (we call such tweets "sick"), and all other tweets (called "other" or "normal").

We need to obtain sufficient amount of labeled training data in order to learn C_f. We do this by first training two "helper" SVMs, C_s and C_o, on a dataset of 5,128 tweets, each labeled as either "sick" or "other" by multiple Amazon Mechanical Turk workers and carefully checked by the authors. C_s is highly penalized for inducing false positives (mistakenly labeling a normal tweet as "sick"), whereas C_o is heavily penalized for creating false negatives (labeling symptomatic tweets as normal). After training, we used C_s and C_o to label a set of 1.6 million tweets that are likely health-related, but contain some noise. We obtained both datasets from Paul and Dredze (2011a), and they are completely disjoint from our NYC data.

The intuition behind this cascading process, illustrated in Fig. 2, is to extract tweets that are with high confidence about sickness with C_s, and tweets that are almost certainly about other topics with C_o from the corpus of 1.6 million tweets. We further supplement the final corpus with messages from a sample of 200 million tweets (also disjoint from all other corpora considered here) that C_o classified as "other" with high probability. We apply thresholding on the classification score to reduce the noise in the cascade, as shown in Fig. 2.

As SVM features, we use all unigram, bigram, and trigram word tokens that appear in the training data. For example, a tweet *"I feel sick."* is represented by the following feature vector:

$$\left(i, feel, sick, i\ feel, feel\ sick, i\ feel\ sick\right).$$

Before tokenization, we convert all text to lower case, strip punctuation and special characters, and remove mentions of user names (the "@" tag) and re-tweets (analogous to email forwarding). However, we do keep hashtags (such as "#sick"), as those are often relevant to the author's health state, and are particularly useful for disambiguation of short or ill-formed messages. Table 2 lists examples of significant features found in the process of learning C_f.

Evaluation of C_f on a held-out set shows 0.98 precision and 0.97 recall. Furthermore, the correlation between the prevalence of infectious diseases predicted by C_f and the predictions made by Google Flu Trends specifically for New York City is 0.73. The official Center for Disease Control and Prevention data for NYC is not available with sufficiently fine granularity, but previous work has shown that Google's predictions closely correspond to the official statistics for larger geographical areas (Ginsberg et al. 2008). Google Flu Trends may have greater specificity to "influenza-like illness", whereas our approach may be less specific, but more sensitive to detect other, related infectious processes exhibiting these nonspecific features in Twitter content.

Predicting the Spread of Disease

Human contact is the single most important factor in the transmission of infectious diseases (Clayton, Hills, and Pickles 1993). Since the contact is often indirect, such as via a doorknob, we focus on a more general notion of *co-location*. We consider two individuals co-located if they visit the same 100 by 100 meter cell within a time window (slack) of length T. For clarity, we show results for T 12 hours, but we obtained virtually identical prediction performance for $T \in \{1, \ldots, 24\}$ hours. We use the 100*m* threshold, as that is the typical lower bound on the accuracy of a GPS sensor in obstructed areas, such as Manhattan. Since we focus on geo-active individuals, we can calculate co-location with high accuracy. The results below are for a condition, where we consider a person ill up to four days after they write a "sick" tweet. As with the parameter T, it is important to note that the results are consistent over a wide range of duration of contagiousness (from 1 to 7 days). Few diseases with influenza-like symptoms are contagious for periods of time beyond these bounds.

Statistical analysis of the data shows that avoiding encounters with infected people generally decreases your chances of becoming ill, whereas a large amount of contact with them makes an onset of a disease almost certain

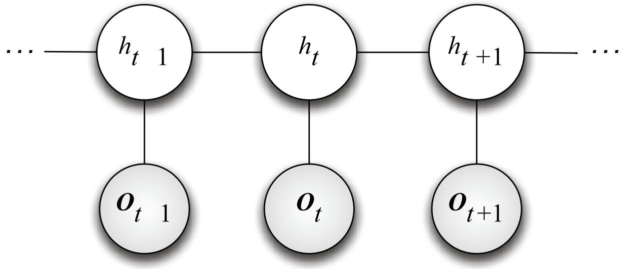


Figure 3: This conditional random field models the health of an individual over a number of days (h_t). The observations for each day (o_t) include day of week, history of sick friends in the near past, the intensity of recent co-location with sick individuals, and the number of such individuals encountered.

(Sadilek, Kautz, and Silenzio 2012). We find a definite exponential relationship between the intensity of co-location and the probability of getting ill. Similarly, by interpreting a Twitter friendship as a proxy for unobservable phenomena and interactions, we see that the likelihood of becoming ill increases as the number of infected friends grows. For example, having more than 5 sick friends increases one's likelihood of getting sick by a factor of 3, as compared to prior probability, and even more with respect to the probability given no sick friends. Additionally, we model the *joint* influence of co-location and social ties, and conclude that the latent impact of friendships is weaker (linear in the number of sick friends), but nonetheless important, as some observed patterns cannot be explained by co-location alone (Sadilek, Kautz, and Silenzio 2012).

Our goal now is to leverage the interplay of co-location and friendships to predict the health state of any individual on a given day. For this purpose, we learn a dynamic conditional random field (CRF), a discriminative undirected graphical model (Lafferty 2001). CRFs have been successfully applied in a wide range of domains from language understanding to robotics. They can systematically outperform alternative approaches, such as hidden Markov models, in domains where it is unrealistic to assume that observations are independent given the hidden state.

In our approach, each person X is captured by one dynamic CRF model with a linear chain structure shown in Fig. 3. Each time slice t contains one hidden binary random variable (X is either "healthy" or "sick" on day t), and a 25-element vector of observed discrete random variables o_t given by

$$o_t \quad \left(\text{weekday}, c_0, \ldots, c_7, u_0, \ldots, u_7, f_0, \ldots, f_7\right),$$

where c_n denotes the number of estimated encounters (co-locations) with sick individuals n days ago. For example, the value of c_1 indicates the number of co-location events a person had a day ago ($t-1$), and c_0 shows co-location count for the current day t. Analogously, u_n and f_n denote the number of unique sick individuals encountered, and the number of sick Twitter friends, respectively, n days ago. For all random variables in our model, we use a special missing value to represent unavailable data.

Before we turn to our experiments, we will discuss the limitations that apply to any indirect method of modeling public health.

Limitations Our observations are limited by the prevalence of public tweets in which users talk about their health, and by our ability to identify them in the flood of other types of messages. Both these factors contribute to the fact that the number of infected individuals is systematically underestimated, but evaluation of C_f suggests that the latter effect is small. We can approximate the magnitude of this bias using the statistics presented earlier. We see that about 1 in 30 residents of NYC appears in our dataset. If we strictly focus on the geo-active individuals, the ratio is roughly 1:3,000. However, the results in this paper indicate, that by leveraging the latent effects of our observations, such a sampling ratio is sufficient to predict the health state of a large fraction of the users with high precision.

We note that currently used methods suffer from similar biasing effects. For example, infected people who do not visit a doctor, or do not respond to surveys are virtually invisible to the traditional methods. Similarly, efforts such as Google Flu Trends can only observe individuals who search the web for certain types of content when sick. A fully comprehensive coverage of a population will require a combination of diverse methods, and application of AI techniques—like the ones presented in this work—capable of *inferring* the missing information.

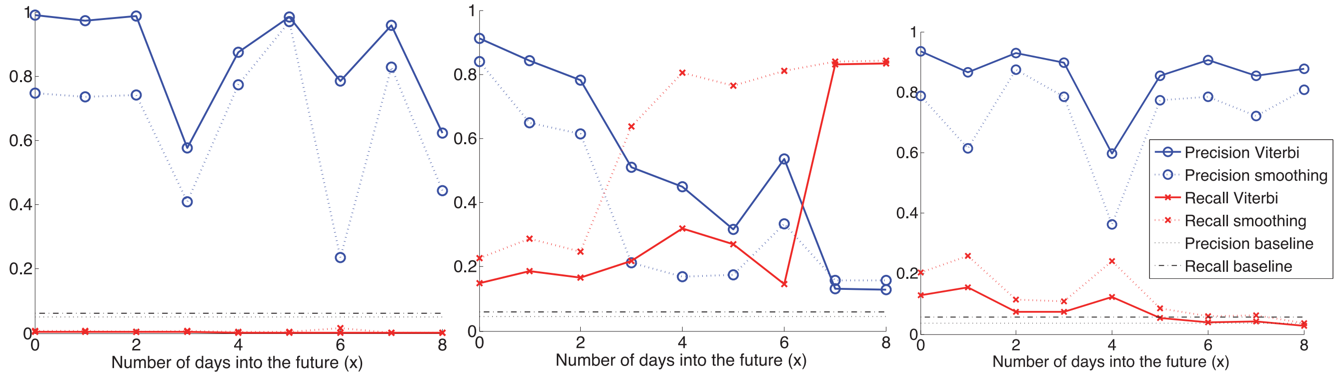
Experiments and Results

In this section, we evaluate our approach in a number of experimental conditions, compare the results of our CRF models with a baseline, and discuss insights gained. We perform 6237-fold cross-validation (the number of geo-active users), where in order to make predictions for a given user, we train and test the CRF while treating all other users as observed. We report results aggregated over all cross-validation runs.

While the structure of the CRF model remains constant across our experiments, we consider two types of inference: Viterbi decoding, and the forwards-backwards algorithm (smoothing). While the former finds the most likely *sequence* of hidden variables (health states) given observations, the latter infers each state by finding maximal marginal probabilities. The tree structure of our CRF allows for scalable, yet exact, learning and inference by applying dynamic programming (Sutton and McCallum 2006), while the rich temporal features capture longer-range dependencies in the data. L1 regularization is used to limit the number of parameters in our model. Maximum-likelihood parameter estimation is done via quasi-Newton method, and we are guaranteed to find a global optimum since the likelihood function is convex.

As a baseline, we consider a model that draws its predictions from a Bernoulli distribution with the "success" parameter p set to the prior probability of being sick learned from the training data.

Fig. 4 summarizes the performance of our models (Bernoulli baseline, and CRF with Viterbi and forwards-backwards inference, respectively) in terms of precision and



(a) Only Friends Observed

(b) Only Co-Location Observed

(c) Both Observed

Figure 4: Summary of results. Each plot shows the precision and recall of our three models for predictions made with hindsight ($x = 0$), and up to 8 days into the future ($x = 8$). We see that when leveraging the effect of social ties or co-locations individually (plots **(a)** and **(b)**, respectively) the CRF models perform inconsistently as we make predictions further into the future. By contrast, when considering friendships and co-location *jointly* **(c)**, the performance stabilizes and improves, achieving up to 0.94 precision and 0.18 recall (AUC of 0.85).

recall along two main dimensions. The first dimension is the type of features the CRF leverges: only information about sick friends (plus weekday) is observed in Fig. 4a; only co-location (plus weekday) is leveraged in Fig. 4b; and the full observation set o_t is available in Fig. 4c. The second dimension is the time for which we make predictions, shown on the horizontal axes (x) in Fig. 4. For x 0, the plots show the performance when inferring the most likely health state for the entire observation sequence (*i.e.*, up to the present day). For $x > 0$, we show the precision and recall when predicting x days into the future, where observations are *not* available. (As described in the previous section, variables corresponding to future observations are simply set to the special "missing" value.)

We see that the results of our CRFs significantly outperform the baseline model. When leveraging the effect of social ties or co-locations individually (Figs. 4a and 4b, respectively), the CRF models perform inconsistently as we make predictions further into the future. By contrast, when considering friendships and co-location *jointly*, the performance stabilizes and improves, achieving up to 0.94 precision and 0.18 recall (Fig. 4c).

In general, we see that Viterbi decoding results in better precision and worse recall, whereas forwards-backwards inference yields slightly worse precision, but improves recall. The relatively low recall indicates that about 80% of infections occur without any evidence in social media as reflected in our features. For example, there are a number of instances of users getting ill even though they had no recent encounters with sick individuals and all their friends have been healthy for a long time.

Clearly, there are complex events and interactions that take place "behind the scenes", which are not directly recorded in online social media. However, this work posits that these latent events often exhibit themselves in the activity of the sample of people we can observe. For instance, we have seen that having online social ties to infected people significantly increases one's chances of becoming ill in the near future (Sadilek, Kautz, and Silenzio 2012). However, we do not believe that the social ties *themselves* cause or even facilitate the spread of an infection. Instead, the Twitter friendships are proxies and indicators for a complex set of phenomena that may not be directly accessible. For example, friends often eat out together, meet in classes, share items, and travel together. While most of these events are never explicitly mentioned online, they are crucial from the disease transmission perspective. However, their likelihood is modulated by the structure of the social ties, allowing us to reason about contagion.

Related Work

Since the famous cholera study by John Snow (1855), much work has been done in capturing the mechanisms of epidemics. There is ample previous work in computational epidemiology on building relatively coarse-grained models of disease spread via differential equations and graph theory (Anderson and May 1979; Newman 2002), by harnessing simulated populations (Eubank et al. 2004), and by analysis of official statistics (Grenfell, Bjornstad, and Kappey 2001). Such models are typically developed for the purposes of assessing the impact a particular combination of an outbreak and a containment strategy would have on humanity or ecology (Chen, David, and Kempe 2010). However, the above works focus on simulated populations and hypothetical scenarios. By contrast, we address the problem of predicting the health of *real-world* populations composed of individuals embedded in a fine social structure. As a result, our work is a major step towards prediction of actual threats and the emergence of disease outbreaks.

In the context of social media, Krieck et al. (2011) explore augmenting the traditional notification channels about a disease outbreak with data extracted from Twitter. By manually examining a large number of tweets, they show that self-reported symptoms are the most reliable signal in detecting if a tweet is relevant to an outbreak or not. This is because people often do not know what their true problem is until diagnosed by an expert, but they can readily write about how they feel. Researchers have also concentrated on capturing the overall *trend* of a particular disease outbreak, typically influenza, by monitoring social media (Culotta 2010; Lampos, De Bie, and Cristianini 2010; Chunara, Andrews, and Brownstein 2012). Freifeld et al. (2010) use information actively submitted by cell phone users to model aggregate public health. However, scaling such systems poses considerable challenges.

Other researchers focus on a more detailed modeling of the *language* of the tweets and its relevance to public health in general (Paul and Dredze 2011a), and to influenza surveillance in particular (Collier, Son, and Nguyen 2011). Paul et al. develop a variant of topic models that captures the symptoms and possible treatments for ailments, such traumatic injuries and allergies, that people discuss on Twitter. In a follow-up work Paul and Dredze (2011b) begin to consider the geographical patterns in the prevalence of such ailments, and show a good agreement of their models with official statistics and Google Flu Trends.

Even the state of the art systems suffer from two major drawbacks. First, they produce only coarse, aggregate statistics, such as the expected number of people afflicted by flu in Texas. Furthermore, they often perform mere passive monitoring, and prediction is severely limited by the low resolution of the aggregate approach, or by scalability issues. By contrast, the primary contribution of this paper is a fine-grained analysis of the interplay among human mobility, social structure, and disease transmission. Our framework allows us to make predictions about likely events of contagion between specific individuals without active user participation.

Conclusions and Future Work

This work is the first to take on prediction of the spread of infectious diseases throughout a real-world population with fine granularity. We focus on self-reported symptoms that appear in people's Twitter status updates, and show that although such messages are rare, we can identify them with systematically high precision and high recall.

The key contribution of this work is a scalable probabilistic model that demonstrates that the health of a person can be accurately inferred from her location and social interactions observed via social media. Furthermore, we show that future health states can be *predicted* with consistently high accuracy more than a week into the future. For example, over 10% of cases of sickness are predicted with 90% confidence even a week before they occur. For predictions one day into the future, our model covers almost 20% of cases with the same confidence.

An early identification of infected individuals is especially crucial in preventing and containing devastating disease outbreaks. Important work by Eubank et al. (2004) shows that by far the most effective way to fight an epidemic in urban areas is to quickly confine infected individuals to their homes. However, this strategy is truly effective only when applied early on in the outbreak. The *speed* of targeted vaccination ranks second in effectiveness. This paper shows that finding some of these key symptomatic individuals, along with other people that may have already contracted the disease, can be done effectively and in a timely manner through social media.

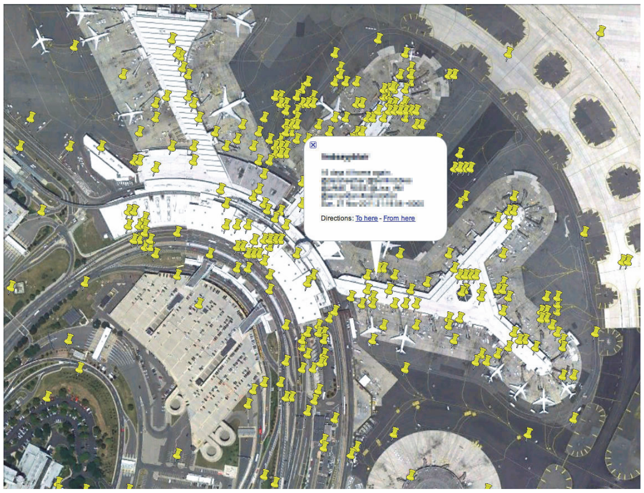


Figure 5: Visualization of a sample of Twitter users (yellow pins) at an airport. The highlighted person X says he will be back in 16 days and mentions specific friends for whom this message is relevant. We immediately see the people at the airport who could have come into contact with X. This work shows that we can accurately predict the health of X from his co-location with other individuals and the heath of his friends. However, additional information can be inferred using methods developed by previous work (Crandall et al. 2010; Backstrom and Leskovec 2011; Cho, Myers, and Leskovec 2011; Sadilek, Kautz, and Bigham 2012). It can be expected that putting all this information together will yield even stronger and more comprehensive predictions about the spread of an infection.

In future work, we will focus on larger geographical areas (including airplane travel), while maintaining the same level of detail (*i.e.*, social ties between concrete individuals and their fine-grained location). This will allow us to model and predict the emergence of global epidemics from the day-to-day interactions of individuals, and subsequently answer questions such as *"How did the current flu epidemic in city A start and where did it come from?"* and *"How likely I am to catch a cold if I visit the mall?"*

For example, Fig. 5 illustrates an instance in our dataset, where a sick person at an airport posts a message, and we can see other people nearby with whom he could have come into contact. Prior work has developed a repertoire of powerful AI techniques for revealing hidden social ties and predicting user location—two features heavily leveraged by our public health model. Therefore, there are opportunities for great synergy in these areas.

Finally, while this paper concentrates on "traditional" infectious diseases, such as flu, similar techniques can be applied to study mental health disorders, such as depression, that have strong contagion patterns as well.

Acknowledgements

We thank Michael Paul for providing his datasets; and Walter Lasecki and Wendy Beatty for their insightful comments. This research was partly funded by ARO grant W911NF-08-1-0242, ONR grant N00014-11-10417, OSD grant W81XWH-08-C0740, and a gift from the Kodak Company. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any of these organizations.

References

Anderson, R., and May, R. 1979. Population biology of infectious diseases: Part I. *Nature* 280(5721):361.

Asur, S., and Huberman, B. 2010. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, 492–499. IEEE.

Backstrom, L., and Leskovec, J. 2011. Supervised random walks: predicting and recommending links in social networks. In *WSDM 2011*, 635–644. ACM.

Chen, P.; David, M.; and Kempe, D. 2010. Better vaccination strategies for better people. In *Proceedings of the 11th ACM conference on Electronic commerce*, 179–188. ACM.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: User movement in location-based social networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Chunara, R.; Andrews, J.; and Brownstein, J. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene* 86(1):39–45.

Clayton, D.; Hills, M.; and Pickles, A. 1993. *Statistical models in epidemiology*, volume 41. Oxford university press Oxford.

Collier, N.; Son, N.; and Nguyen, N. 2011. OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*.

Crandall, D.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences* 107(52):22436.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 115–122. ACM.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Eubank, S.; Guclu, H.; Anil Kumar, V.; Marathe, M.; Srinivasan, A.; Toroczkai, Z.; and Wang, N. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180–184.

Freifeld, C.; Chunara, R.; Mekaru, S.; Chan, E.; Kass-Hout, T.; Iacucci, A.; and Brownstein, J. 2010. Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS medicine* 7(12):e1000376.

Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Grenfell, B.; Bjornstad, O.; and Kappey, J. 2001. Travelling waves and spatial hierarchies in measles epidemics. *Nature* 414(6865):716–723.

Gruzd, A.; Wellman, B.; and Takhteyev, Y. 2011. Imagining Twitter as an imagined community. In *American Behavioral Scientist, Special issue on Imagined Communities*.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *ICML 2005*, 377–384. ACM.

Krieck, M.; Dreesman, J.; Otrusina, L.; and Denecke, K. 2011. A new age of public health: Identifying disease outbreaks by analyzing tweets. *Proceedings of Health Web-Science Workshop, ACM Web Science Conference*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a Social Network or a News Media? In *WWW*.

Lafferty, J. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 282–289. Morgan Kaufmann.

Lampos, V.; De Bie, T.; and Cristianini, N. 2010. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases* 599–602.

Newman, M. 2002. Spread of epidemic disease on networks. *Physical Review E* 66(1):016128.

Paul, M., and Dredze, M. 2011a. A model for mining public health topics from Twitter. *Technical Report. Johns Hopkins University. 2011.*

Paul, M., and Dredze, M. 2011b. You are what you tweet: Analyzing Twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding your friends and following them to where you are. In *Fifth ACM International Conference on Web Search and Data Mining*. (Best Paper Award).

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*.

Snow, J. 1855. *On the mode of communication of cholera*. John Churchill.

Sutton, C., and McCallum, A. 2006. *An introduction to conditional random fields for relational learning*. Introduction to statistical relational learning. MIT Press.

Tumasjan, A.; Sprenger, T.; Sandner, P.; and Welpe, I. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185.