Modeling Spread of Disease from Social Interactions

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

Research in computational epidemiology to date has concentrated on coarse-grained statistical analysis of populations, often synthetic ones. By contrast, this paper focuses on fine-grained modeling of the spread of infectious diseases throughout a large real-world social network. Specifically, we study the roles that social ties and interactions between specific individuals play in the progress of a contagion. We focus on public Twitter data, where we find that for every health-related message there are more than 1,000 unrelated ones. This class imbalance makes classification particularly challenging. Nonetheless, we present a framework that accurately identifies sick individuals from the content of online communication. Evaluation on a sample of 2.5 million geo-tagged Twitter messages shows that social ties to infected, symptomatic people, as well as the intensity of recent co-location, sharply increase one's likelihood of contracting the illness in the near future. To our knowledge, this work is the first to model the interplay of social activity, human mobility, and the spread of infectious disease in a large real-world population. Furthermore, we provide the first quantifiable estimates of the characteristics of disease transmission on a large scale without active user participation—a step towards our ability to model and predict the emergence of global epidemics from day-to-day interpersonal interactions.

Introduction

Given that five of your friends have flu-like symptoms, and that you have recently met eight people, possibly strangers, who complained about having runny noses and headaches, what is the probability that you will soon become ill as well? This work explores how accurately such questions can be answered across a large sample of people participating in online social media (see Fig. 1).

Imagine Joe is about to take off on an airplane and quickly posts a Twitter update from his phone. He writes that he has a fever and feels awful. Since Joe has a public Twitter profile, we know who some of his friends are, and from his GPS-tagged messages we see some of the places he has recently visited. Additionally, we can infer a large fraction of the hidden parts of Joe's social network and his latent locations by applying the results of previous work, as we discuss below. In the same manner, we can identify other people who are likely to be at Joe's airport, or even on the same flight. Using both the observed and inferred information, we can now monitor individuals who likely came into contact with Joe, such as the passengers seated next to him. Joe's disease may have been transmitted to them, and vice versa, though they may not exhibit any symptoms yet. As people travel to their respective destinations, they may be infecting others encountered along the way. Eventually, some of the people will tweet about how they feel, and we can observe at least

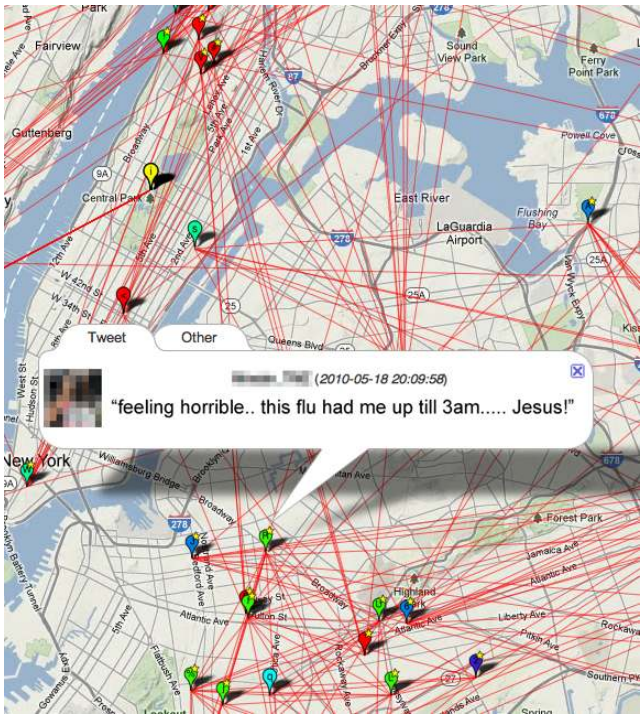


Figure 1: Visualization of a sample of friends in New York City. The red links between users represent friendships, and the colored pins show their current location on a map. We see the highlighted person complaining about her health, and hinting about the specifics of her ailment. This work investigates the impact of such a person on the health of her friends, and of people around her.

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

a fraction of the population that actually contracted the disease.

The example just given illustrates our vision of how public health modeling may look like in the near future. This paper reports on our initial progress towards making this vision a reality.

Traditionally, public health is monitored via surveys and by aggregating statistics obtained from healthcare providers. Such methods are costly, slow, and may be biased. For instance, a person with flu is recorded only after he or she visits a doctor's office and the information is sent to the appropriate agency. Affected people who do not seek treatment, or do not respond to surveys are virtually invisible to the traditional methods.

Recently, digital media has been successfully used to significantly reduce the latency and improve the overall effectiveness of public health monitoring. Perhaps most notably, Google Flu Trends[1] models the prevalence of flu via analysis of geo-located search queries (Ginsberg et al. 2008).

Twitter itself has been recently shown to accurately assess the overall prevalence of flu independently in a number of countries with accuracy comparable to current state of the art methods including Google Flu and Center for Disease Control and Prevention (CDC) statistics[2] (Lampos, De Bie, and Cristianini 2010; Culotta 2010; Signorini, Segre, and Polgreen 2011). However, 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.

By contrast, our work takes a bottom-up approach, where we take into account the fine-grained interactions between individuals. We apply machine learning techniques to the difficult task of detecting ill individuals based on the content of their Twitter status updates. We are then able to estimate the physical interactions between healthy and sick people via their online activities, and model the impact of these interactions on public health.

As a result, this work is one of the first steps 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*. Specifically, we investigate the following research question: *What roles do co-location and social ties play in the spread of infectious diseases from person to person?* Our answers to this question provide a solid stepping stone for further research.

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 March 2011, approximately 200 million accounts are registered on Twitter.[3]

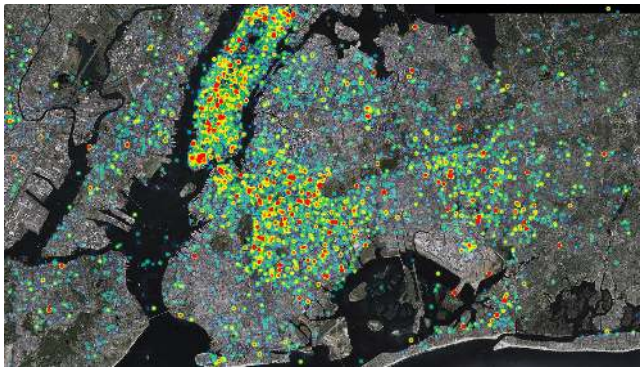


Figure 2: A snapshot of a heatmap animation of Twitter users' movement within New York City that captures a typical distribution of geo-tagged messaging on a weekday afternoon. The hotter (more red) an area is, the more people have recently tweeted from that location.

Using the Twitter Search API[4], we collected a sample of public tweets that originated from the New York City (NYC) metropolitan area shown in Fig. 2. 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. In order to avoid exceeding Twitter's query rate limits and subsequently missing some tweets, we distributed the work over a number of machines with different IP addresses that asynchronously queried the server and merged their results. Twitter does not provide any guarantees as to what sample of existing tweets can be retrieved through their API, but a comparison to official Twitter statistics shows that our method recorded the majority of the publicly available tweets in the region. 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.[5] 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*. The social network of the 6,237 geo-active users is shown in Fig. 3.

Methodology and Models

In this section, we first present our method for automatic detection of Twitter messages that suggest the author con-

[1] http://www.google.org/flutrends

[2] http://www.cdc.gov/datastatistics/

[3] http://www.bbc.co.uk/news/business-12889048

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

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

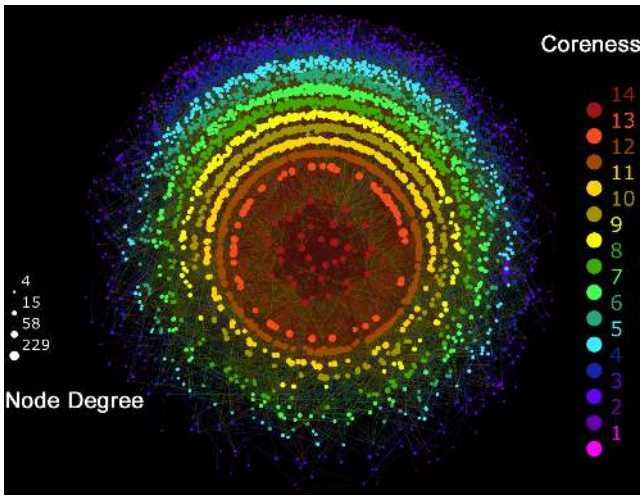


Figure 3: Visualization of the social network consisting of the geo-active users. Edges between nodes represent friendships on Twitter. The image has been created using LaNet-vi package implementing k-core decomposition (Beiró, Alvarez-Hamelin, and Busch 2008). The coreness of nodes is color-coded using the scale on the right. The degree of a node is represented by its size shown on the left. We see that there are relatively few important "hubs" in the central area, and a large number of less connected individuals on the fringes.

tracted an infectious disease.[6] We then discuss how we leverage this framework in order to develop our model of public health.

Detecting Illness-Related Messages

As a first step, we need to identify Twitter messages that indicate the author is infected with a disease of interest at the time of posting. Based on the results of previous work, we expect that health-related tweets are relatively scarce as compared to other types of messages (Culotta 2010; Paul and Dredze 2011a). Given this class imbalance problem, we formulate a semi-supervised cascade-based approach (shown in Fig. 4) to learning a robust support vector machine (SVM) classifier with a large area under the ROC curve (*i.e.*, consistently high precision and high recall). SVM is an established model of data in machine learning (Cortes and Vapnik 1995). We learn an SVM for linear binary classification to accurately 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").

In order to learn such classifier, we ultimately need to effortlessly obtain a high-quality set of labeled training data. We achieve this via the following "bootstrapping" process. We begin by training two different binary SVM classifiers,

[6]In this study, 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).

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.

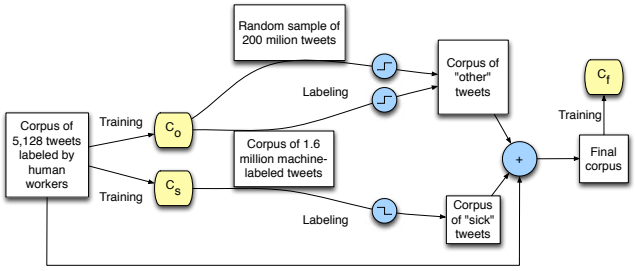


Figure 4: A diagram of our cascade learning of SVMs. The ⎍ and ⎌ symbols denote thresholding of the classification score, where we select the bottom 10% of the scores predicted by C_o (*i.e.*, tweets that are normal with high probability), and the top 10% of scores predicted by C_s (*i.e.*, likely "sick" tweets).

C_s and C_o, using the SVM^{light} package.[7] C_s is highly penalized for inducing false positives (mistakenly labeling a normal tweet as one about sickness), whereas C_o is heavily penalized for creating false negatives (labeling symptomatic tweets as normal). For both classifiers, the misclassification penalty for one direction was always a hundred times larger than in the opposite direction. We train C_s and C_o using 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. After training, we used the two classifiers 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 is to extract tweets that are with high confidence about sickness with C_s,

[7]http://svmlight.joachims.org/

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. 4.

The cascade yields a final corpus with over 700 thousand "sick" messages and 3 million "other" tweets, which we use for training the final SVM C_f. We will discuss how we leverage C_f to model the disease spread below, but first let us describe the feature space and our learning methodology in more detail.

As 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:

$$\Big(i, feel, sick, i\ feel, feel\ sick, i\ feel\ sick\Big).$$

Before tokenization, we convert all text to lower case, strip punctuation and special characters, and remove mentions of user names (the "@" tag). All re-tweets (analogous to email forwarding) have been removed as well, since those messages typically refer to popular news and social games, and rarely describe the current state of the author. 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. When learning the final SVM C_f, we only consider tokens that appear at least three times in the training set.

While our feature space has a very high dimensionality (C_f operates in more than 1.7 million dimensions), with many possibly irrelevant features, support vector machines with a linear kernel have been shown to perform very well under such circumstances (Joachims 2006; Sculley et al. 2011; Paul and Dredze 2011a).

To overcome the class imbalance problem, where the number of tweets about an illness is much smaller than the number of other messages, we apply the ROCArea SVM learning method that directly optimizes the area under the ROC curve, as described in Joachims (2005). Traditional objective functions, such as the 0-1 loss perform poorly under severe class imbalance. For instance, a trivial model that labels every example as belonging to the majority class has an excellent accuracy, because it misses only the relatively few minority examples. By contrast, the ROCArea method works by implicitly transforming the classical SVM learning problem over individual training examples into one over pairs of examples. This allows efficient calculation of the area under the ROC curve from the predicted ranking of the examples.

Modeling 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 $T \in \{1, 4, 12\}$ hours. We use the $100m$ 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 a person is ill up to two days after they write a "sick" tweet. It is important to note that the relationships among friendship, co-location, and health are consistent over a wide range of duration of contagiousness (from 1 to 7 days). Most infectious illnesses produce influenza-like symptoms that stop within a few days, and thus within these temporal bounds.

To quantify the effect of social ties on disease transmission, we leverage users' Twitter friendships. 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, as we will see, having social ties to infected people significantly increases your chances of becoming ill in the near future. 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.

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 may be sufficient.

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 capable of *inferring* the missing information.

Experiments and Results

In this section, we evaluate our approach, compare the results of our model with an established baseline, and discuss insights gained.

Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
sick	0.9579	sick of	−0.4005
headache	0.5249	you	−0.3662
flu	0.5051	of	−0.3559
fever	0.3879	your	−0.3131
feel	0.3451	lol	−0.3017
cough	0.3062	who	−0.1816
feeling	0.3055	u	−0.1778
coughing	0.2917	love	−0.1753
throat	0.2842	it	−0.1627
cold	0.2825	her	−0.1618
home	0.2107	they	−0.1617
still	0.2101	people	−0.1548
bed	0.2088	shit	−0.1486
better	0.1988	smoking	−0.0980
being	0.1943	i'm sick of	−0.0894
being sick	0.1919	so sick of	−0.0887
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: Top twenty most significant negatively and positively weighted features of our SVM model.

Evaluation of the final SVM C_f described in the previous section on a held-out test set of 700,000 tweets shows 0.98 precision and 0.97 recall. This evaluation run also allows us to choose an optimal threshold on the classification score that separates the normal tweets from sick tweets. Table 2 lists the most significant features C_f found. Table 3 shows examples of tweets that C_f identified as "sick". We now apply C_f to modeling the spread of infectious diseases throughout the sampled population of NYC described above.

The correlation between the prevalence of infectious diseases predicted by our model and the predictions made by Google Flu Trends specifically for New York City is 0.73. The official CDC 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. Furthermore, the only overlap between our predictions and those of Google is for May 18 through 23, 2010. Thus, the correlation between the two needs to be interpreted with this context in mind.

Figures 5a and 5b show the impact of co-location and friendship with infected people on a given day on one's health the following day. We analyze both the individual and joint effects of the two factors on disease transmission. For brevity, we include plots only for a 1-day lag, since other time offsets result in a similar relationship.

Looking at co-location effect alone, we observe a definite exponential relationship between probable physical encounters and ensuing sickness. All three curves in Fig. 5a consistently fit $f(x) = C\, e^{(0.055x)}$, where C is a constant that captures the length of time overlap T (note that $C \simeq 0.011/T$; thus the larger the slack the smaller the effect). For instance, having 40 encounters with sick individuals with a 1-hour slack makes one ill with 20% probability. With a more lenient slack, such as 4 hours, one needs over 80 encounters to reach the same level of risk.

In Fig. 5b, we see that the number of sick friends also has an exponential effect on the probability of getting sick: $f(x) = 0.003\, e^{(0.413x)}$. By contrast, the number of friends in any health state (*i.e.*, the size of one's friend list) has *no* impact on one's health. In fact, the conditional probability of getting sick given n friends (the blue line in Fig. 5b) is virtually identical to the prior probability of getting sick (the black line).

We have discussed the latent influence of friendships earlier. This is quantitatively shown as the green line in Fig. 5b, where we *subtract out* the effect of co-location from the influence of social ties. We do this by counting only sick friends who have not been encountered. Comparison with the red curve shows that for smaller numbers of friends ($n \leqslant 6$), co-location has a weak additional effect over the proxy effect of social ties. However, for larger n, the residual impact of friendships plateaus, and co-location begins to dominate.

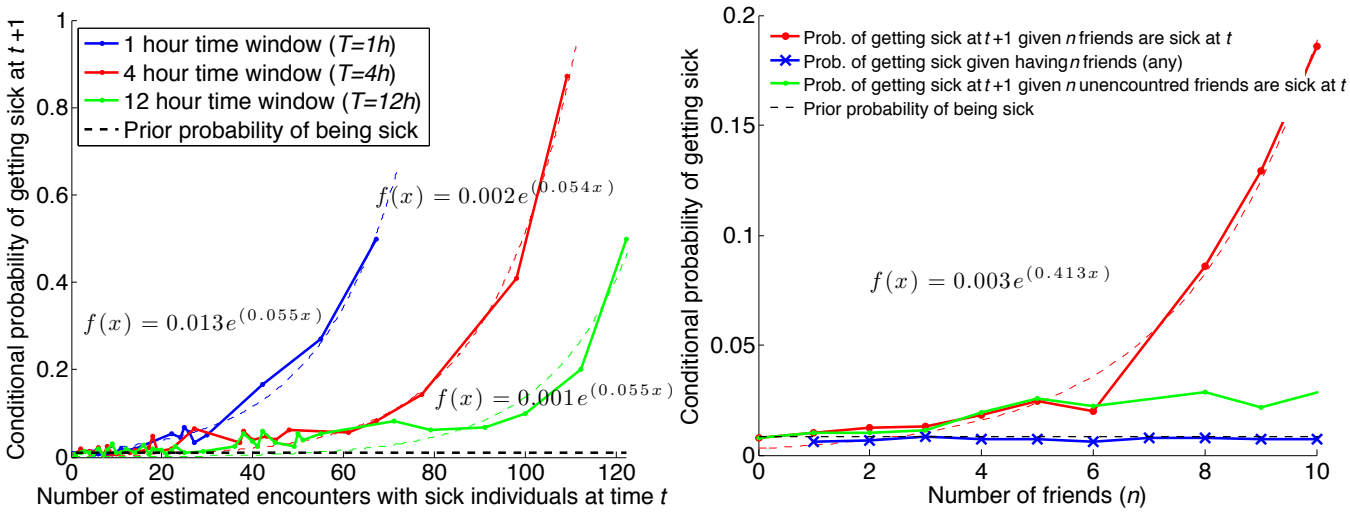
Related Work

Since the famous cholera study by John Snow, much work has been done in capturing the mechanisms of epidemics (Snow 1855). There is ample previous work in computational epidemiology on building models of coarse-grained disease spread via differential equations (Anderson and May 1979), 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 vaccination or containment strategy would have on humanity, a country's defense, 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 assessing and modeling 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 instances 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;

Came home sick today from work with a killer headache and severe nausea, took 2 advil and slept for 6 hours. I feel much better now.
Meh I actually have to go to school tomorrrow.. #sick
Not feeling good at all...that sucks because I plans with my bff and job interviews set up until Tuesday. Stomach is killing me
I'm feeling better today still stuffed up but my nose isn't running like it was yesterday and my cough is better as well it hurts.
Guys I'm sorry. I'm really have to get some rest. I have nausea, headache, is tired, freezing & now have I got fever. Good Night! :-*
It hurts to breathe, swallow, cough or yawn. I must be getting sick, though because my ear feels worse than my throat.
I just sneezed 6 times in a row. i hate being sick.
feeling misserable. stomach hurts, headache, and no, I'm not pregnant.
Been sleep all day smh.... Currently soothing my jimmy frm α headache as I go back to sleep
Just not feeling it today. Looks like man flu has come back for a visit. I need to be well and have work - is that too much to ask?

Table 3: Example tweets that our SVM model C_f identified as "sick". Note the high degree of variability, and sometimes subtlety, in the way different people describe their health.



(a) Effects of co-location alone

(b) Effects of friendship and co-location

Figure 5: Being co-located with ill, symptomatic individuals, and having sick friends on a given day (t) makes one more likely to get sick the next day ($t+1$). On the horizontal axis in **(a)**, we plot the amount of co-location of an asymptomatic user with known sick people on a given day. In **(b)**, we show the number of friends (of an asymptomatic user); either only sick ones or any depending on the curve. The vertical axes show the conditional probability of getting sick the next day. We also plot the prior probability of being sick. For co-location, results for three slack time windows, within which we consider an appearance of two users close together as co-location, are shown (1, 4, and 12 hours).

Lampos, De Bie, and Cristianini 2010; Chunara, Andrews, and Brownstein 2012). Interesting work of Ritterman, Osborne, and Klein (2009) shows that noisy Twitter data is a valuable information channel for predicting public opinion regarding the likelihood of a pandemic. 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. There is a potential for synergy between the work of Paul et al. and ours that would allow us to model the spread of *specific* diseases by leveraging the rich language models.

However, all these works consider only aggregate patterns captured by coarse-grained statistics, whereas the primary contribution of our paper is a more detailed study of the interplay among human mobility, social structure, and disease transmission. Our framework allows us to track—without active user participation—specific likely events of contagion between individuals, and model the relationship between an epidemic and self-reported symptoms of actual users of online social media.

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. Pioneering work in this broad area includes Silenzio et al. (2009), which studies characteristics of young lesbian, gay, and bisexual individuals in online social networks. They focus on discovering such members of a community, and design methods for effective peer-driven information diffusion and preventative care, focusing specifically on suicide. Twitter has also been used to monitor the seasonal variation in affect around the globe (Golder and Macy 2011).

Looking at a more global scale, Bettencourt and West (2010) argue for a comprehensive scientific approach to urban planning. They show there are underlying patterns that tie together the size of a city with its emergent characteristics, such as crime rate, number of patents produced, walking speed of its inhabitants, and prevalence of epidemics. The authors argue that cities are the source of many major problems, but also contain the solutions because of their concentrated creativity and productivity.

Since this work leverages social ties and user location, the large body of prior work on inferring and predicting these characteristics becomes relevant. A number of researchers have demonstrated that it is possible to accurately predict people's fine-grained location from their online behavior and interactions (Cho, Myers, and Leskovec 2011; Sadilek, Kautz, and Bigham 2012). Much progress has been made in predicting the social structure of participants in online media, including Twitter, from various types of observed data (Crandall et al. 2010; Backstrom and Leskovec 2011; Sadilek, Kautz, and Bigham 2012). Applying these machine learning techniques will significantly expand the breadth of data available by allowing us to consider not only declared friendships and public check-ins, but also their inferred—though more ambiguous—counterparts.

Conclusions and Future Work

This work is the first to take on modeling 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 high precision as well as high recall. We achieve this by developing an SVM model that is robust even in the presence of strong class imbalance. This is a necessary precondition for further progress, as false negatives and false positives cannot be traded-off against each other in this domain—they both carry equal importance.

We have seen that avoiding encounters with infected people decreases your chances of becoming ill, whereas a large amount of contact with them makes an onset of a disease almost certain (Fig. 5a). Similarly, by interpreting a virtual friendship as a proxy for unobservable phenomena and interactions, we have shown that the likelihood of becoming ill exponentially 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 (Fig. 5b). Additionally, we model the *joint* influence of co-location and social ties, and quantify the latent impact of friendships.

Figure 6: Visualization of a sample of Twitter users (yellow pins) at the Newark Liberty International 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, and additional candidates can be inferred using methods developed by previous work (Crandall et al. 2010; Backstrom and Leskovec 2011; Sadilek, Kautz, and Bigham 2012). Additionally, recent results show that the future location and co-location of the individuals can be *predicted* at various temporal scales with high accuracy (Cho, Myers, and Leskovec 2011; Sadilek, Kautz, and Bigham 2012). Since some people explicitly mention their symptoms, it can be expected that putting all this information together will yield strong predictions about the spread of an infection.

An early identification of infected individuals is especially crucial in preventing and containing devastating disease outbreaks. Important work of 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 agility of targeted vaccination ranks second in effectiveness. This paper shows that finding 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. As our final contribution, we show that the predictions made by our model strongly correlate with Google Flu Trends, currently the state of the art system for monitoring the prevalence of influenza-like illnesses.

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 concerete 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?".

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, as we illustrate in Fig. 6.

References

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

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

Beiró, M.; Alvarez-Hamelin, J.; and Busch, J. 2008. A low complexity visualization tool that helps to perform complex systems analysis. *New Journal of Physics* 10:125003.

Bettencourt, L., and West, G. 2010. A unified theory of urban living. *Nature* 467(7318):912–913.

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*.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

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.

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.

Golder, S., and Macy, M. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.

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.

Joachims, T. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 217–226. 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*.

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

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 (ICWSM 2011)*.

Ritterman, J.; Osborne, M.; and Klein, E. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. *1st International Workshop on Mining 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).

Sculley, D.; Otey, M.; Pohl, M.; Spitznagel, B.; Hainsworth, J.; and Yunkai, Z. 2011. Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Signorini, A.; Segre, A.; and Polgreen, P. 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).

Silenzio, V.; Duberstein, P.; Tang, W.; Lu, N.; Tu, X.; and Homan, C. 2009. Connecting the invisible dots: Reaching lesbian, gay, and bisexual adolescents and young adults at risk for suicide through online social networks. *Social Science & Medicine* 69(3):469–474.

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