Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions

Sean Brennan **Adam Sadilek** **Henry Kautz**
Department of Computer Science
University of Rochester
Rochester, NY 14627
sbrennan@u.rochester.edu, {sadilek, kautz}@cs.rochester.edu

Abstract

Monitoring and forecast of global spread of infectious diseases is difficult, mainly due to lack of fine-grained and timely data. Previous work in computational epidemiology has shown that mining data from the web can improve the predictability of high-level aggregate patterns of epidemics. By contrast, this paper explores how individuals *contribute* to the global spread of disease. We consider the important task of predicting the prevalence of flu-like illness in a given city based on interpersonal interactions of the city's residents with the outside world. We use the geo-tagged status updates of traveling Twitter users to infer properties of the flow of individuals between cities. While previous research considered only the raw volume of passengers, we estimate a number of latent variables, including the number of sick (symptomatic) travelers and the number of sick individuals to whom each traveler was exposed. We show that AI techniques provide insights into the mechanisms of disease spread and significantly improve predictability of future flu outbreaks. Our experiments involve over 51,000 individuals traveling between 75 cities prior and during a severe ongoing flu epidemic (October 2012 - January 2013). Our model leverages the text and interpersonal interactions recorded in over 6.5 million online status updates without any active user participation, enabling scalable public health applications.

Introduction

Recent research in computer science and computational epidemiology has demonstrated that analysis of social media can reveal important public health information [Lampos *et al.*, 2010; Paul and Dredze, 2011; Chunara *et al.*, 2012; Sadilek *et al.*, 2012b]. Prior work concentrated on two broad areas: (1) capturing aggregate health trends, and (2) modeling the health of particular individuals. The primary goal of the first line of work is to estimate the current rate of influenza in a given country. The second area—enabled by the recent surge in pervasiveness of online social media—places emphasis on predicting which specific individuals will become afflicted with flu tomorrow. Both areas have made significant progress in recent years. However, our understanding of the *emergence* of global epidemic patterns from everyday interactions between people is limited.

This paper makes the first steps towards revealing the underlying mechanisms of disease transfer that ultimately produce the epidemics we observe. We develop computational techniques that combine data mining of online microblogs, machine learning that extracts latent information from the data, and statistical analysis that reveals associations between fine-grained behavior of concrete individuals and population-level flu prevalence. Our methods enable us to shed additional light on important questions in public health that have been either too expensive or outright impossible to answer. In the process, we draw parallels to work done in other scientific fields, including epidemiology and immunology, and show how our methods complement previous results and bring new insights.

It has been shown that the overall intensity of airplane travel is associated with the speed and severity of the spread of influenza [Grais *et al.*, 2003; Colizza *et al.*, 2006; Brownstein *et al.*, 2006; Ruan *et al.*, 2006; Hollingsworth *et al.*, 2007; Nicolaides *et al.*, 2012]. However, all work to date used only coarse-grained aggregate statistics to guide the simulations or to estimate the magnitude of these effects. By contrast, we can now begin to model the actual number of susceptible, infectious, and infected individuals traveling between specific airports by leveraging online social media.

Going beyond data mining, we infer *latent* features that provide better predictions than alternative models. Specifically, we learn a language model that classifies individuals as either "healthy" or "sick" (symptomatic) based on the text of their online messages. Since we face a steep class imbalance problem, where the number of healthy people overwhelms the number of infected, we formulate the classification problem as a specific instance of support vector machine (SVM) learning. We directly optimize the area under the ROC curve to achieve high precision and high recall [Joachims, 2005].

Another latent feature captures *meetings* between people. We leverage the GPS tags associated with the messages to estimate who met who, while tracking the health state of everyone involved. As we will see, modeling the health state significantly improves prediction accuracy, making our forecasts more actionable in real-world settings. This demonstrates that

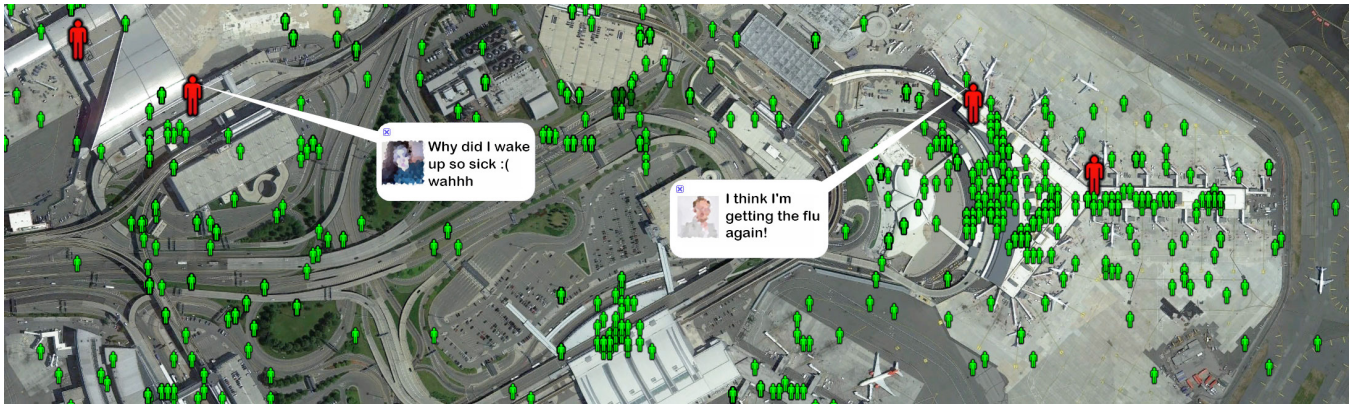

Figure 1: Snapshot of Twitter activity at the JFK airport in New York City. Individuals who indicate sickness are highlighted in red. From the GPS-tagged data, we see who likely came into contact with the infected people. This paper shows that we can accurately predict the prevalence of flu in a city by modeling the flux of healthy and sick travelers while modeling their physical interactions.

AI techniques are essential for solving important sustainability challenges that humanity faces today.

Given that we detect a 20% increase in the influx of sick passengers flying into Boston on a given day, how does the prevalence of influenza-like disease in Boston change in the near future? What role does the intensity of physical contact between the passengers and city residents play in the outbreak? Can these patterns be leveraged to predict future outbreaks? In the remainder of this paper, we propose and evaluate a model that provides *quantitative* answers to such questions on the basis of online social network data (see Figure 1).

The real-time aspect of online social media encompassing a large fraction of the population enables pervasive data collection and analysis at scale. However, the data is noisy and often incomplete. The low signal-to-noise ratio makes detailed modeling challenging. We accept the inherent uncertainty in the data and begin to answer the challenges involved. While traditional survey data is more clear-cut than online logs, it is too affected by biasing factors. For example, infected people who do not visit a doctor participating in a surveillance program, or who provide misleading answers result in unreliable data. We argue in the paper that a *unification* of traditional techniques and scalable data-mining approaches will lead to better models and tools by canceling each others' weaknesses.

As we will see, signals that can be extracted from geo-tagged online data with relative ease—such as the total number of Twitter users flying from New York City to Boston on a given day—do explain a portion of the aggregate patterns we see. However, we show that by including *inferred* attributes—such as the most likely health state of each passenger on each day—a significantly more accurate model is obtained. By viewing Twitter users as noisy proxies for mobility and health of the general population, we are able to model key statistics in real time, including the expected number of *sick* passengers flying from city A to city B.

As a result, we show it is possible to decompose the flu index in a given city into a weighted contribution of a large collection of factors. Some factors capture local interactions of individuals within a city, while others model the dynamics of human travel across cities. The learned features are subsequently leveraged for accurate predictions of future progress of flu outbreaks. Our model provides insights that could help guide individual as well as public health management decisions. Important questions such as "Should I fly tomorrow or next week?" and "Which transit hubs or specific connections should we close to curtail a flu pandemic?" can now be answered in a more data-driven fashion. Specifically, we show the following:

- Raw airline traffic volume between cities estimated from Twitter data explains 56% of the variance in flu prevalence across 75 major US metropolitan areas.
- Accounting for the *health* of individual passengers explains additional 17% of the variance (a total of 73%).
- Modeling physical encounters between sick and healthy individuals explains additional 5% (a total of 78%).
- Our prediction of a city's flu index next week is within 7% of the true value 95% of the time.
- The single most important factor of the prevalence of flu in a given city is the number of sick passengers that have flown into the city over the past 7 days.
- Our results hold in validation on two independent respected flu metrics (CDC and Google Flu Trends).

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. Tweets sent from mobile devices are usually GPS-tagged with accurate location. We leverage these to model meetings among users in the physical world.

Using the Twitter Search API[1], we collected a corpus of

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

Dataset Statistics	
Number of days	95
Number of airports tracked	100
Number of metropolitan areas	75
Users total	632,611
Target users	51,137
Tweets by target users	6,287,446
User flights inferred through tweets	73,460
Number of meetings (target users only)	445,812

Table 1: Summary of statistics of Twitter data relevant to airplane travel collected by our backtracking technique. *Target users* are ones who have tweeted from more than one airport.

public geo-tagged tweets that originated from the 100 busiest commercial airports in the United States. We periodically queried Twitter for all recent tweets within three kilometers of each airport. The collection period was 95 days long and started on October 12, 2012. Altogether, we have logged over 6.5 million tweets authored by more than 630,000 unique users (see Table 1).

Since this work focuses on the effects of fast long-distance travel on global disease transmission, we concentrate on users who had tweeted from two or more airports. We refer to them as *target users*, and our dataset contains 51,137 such individuals. We used the Twitter REST API[2] to collect detailed timelines of target users to obtain a broader context of their health and interactions with others. This enables us to account for users who may indicate sickness before or after tweeting from an airport. We refer to mining this context informally as *backtracking* user timelines.

We evaluate our models against two well-established flu surveillance methods: the official U.S. Center for Disease Control (CDC) statistics[3] (C_f), and Google Flu Trends[4] (G_f). Each index estimates the prevalence of flu across various regions based on hospital records and web search query analysis, respectively [Ginsberg *et al.*, 2008]. Both data sources are published with weekly granularity and with limited geographical resolution. While G_f includes 75 metropolitan areas served by our 100 airports, CDC operates at the resolution of U.S. states at the finest level. By contrast, our methods provide real-time flu signal with spatial resolution limited only by the accuracy of the GPS sensor in people's mobile devices (typically under 100 meters).

Background

Support vector machine (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 ailment and all other tweets. Linear binary SVMs are trained by finding a hyperplane defined by a normal vector w with the maximal margin separating it from the positive and negative datapoints. Finding such a hyperplane is inherently a quadratic optimization problem given by the following objective function that can be solved efficiently and in a parallel fashion using stochastic gradient descent methods [Shalev-Shwartz *et al.*, 2007].

$$\min_w \frac{\lambda}{2}||w||^2 + \mathcal{L}(w, D) \tag{1}$$

where λ is a regularization parameter controlling model complexity, and $\mathcal{L}(w, D)$ is the hinge-loss over all training data D given by

$$\mathcal{L}(w, D) = \sum_i \max\left(0, 1 - y_i w^T x_i\right) \tag{2}$$

Class imbalance, where the number of examples in one class is dramatically larger than in the other class, complicates virtually all machine learning. For SVMs, prior work has shown that transforming the optimization problem from the space of individual datapoints $\langle x_i, y_i \rangle$ in matrix D to one over *pairs* of examples $\langle x_i^+ - x_j^-, 1 \rangle$ yields significantly more robust results [Joachims, 2005].

Regression analysis is a statistical technique of quantifying the relationship between one or more independent variables and a dependent response variable. In this work, we apply a regularized least-squares regression model with elastic net algorithm [Zou and Hastie, 2005]. This formalism encourages grouping of strongly correlated independent variables, and enables variable selection in a principled way.

Related Work

There is ample previous work in computational epidemiology on building relatively coarse-grained models of disease spread by harnessing simulated populations [Newman, 2002; Grais *et al.*, 2003; Eubank *et al.*, 2004; Nicolaides *et al.*, 2012]. 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. Hollingsworth *et al.* (2007) present an interesting analysis of the impact of frequent fliers. By evaluating a SEIR simulation, they argue that these travelers could significantly accelerate an epidemic of a respiratory disease only if they get infected early on in the outbreak. However, these works focus on simulated populations and hypothetical scenarios. By contrast, we address the problem of predicting the health of real world populations in real time—a step towards understanding actual threats and ongoing disease outbreaks.

A seminal case study in traditional epidemiology focused on recovering detailed contagion events by examining passengers and crew who had been delayed aboard an airliner [Moser *et al.*, 1979]. People's location, encounters, and health states were manually reconstructed by telephone surveys, analysis of airline data, and medial examination. Such efforts do not scale with 643 million individuals in almost 9 million flights that take place annually in the U.S. alone. In this work, we show that similar results can be achieved *at scale* by applying automated methods.

Researchers have concentrated on capturing the overall *trend* of a particular disease outbreak, typically influenza, by monitoring social media [Culotta, 2010; Lampos *et al.*, 2010;

[2]https://dev.twitter.com/docs/api

[3]http://www.cdc.gov/flu/weekly/

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

Chunara *et al.*, 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 [Collier *et al.*, 2011; Paul and Dredze, 2011]. In our previous work, we have shown that future health of an individual can be accurately *predicted* from geo-tagged tweets on the basis of his or her interactions with already infected people, including online friends and encountered strangers [Sadilek *et al.*, 2012b].

Brownstein *et al.* (2006) find that the regional influenza spread is associated with the volume of air travel in November prior to the flu season. Taking advantage of the sudden reduction in air travel after the September 11 terrorist attacks, the researchers show that the following flu season has been delayed. While Brownstein et al. present an excellent analysis of the impact of the bulk travel volume on the ensuing flu season, the mechanisms underlying this impact remain unclear—a gap our work begins to fill. Their model explains 60% of the yearly variance of influenza spread, whereas our approach explains 73% of *weekly* signal—a significantly harder problem.

Eubank *et al.* (2004) are beginning to leverage more fine-grained information, including people's activities. They developed a simulation tool (EpiSims) that leverages synthetic—but statistically realistic—human mobility to study the spread of infectious diseases over a metropolitan area. They show their simulation-based approach is a viable alternative to the classical models formulated using differential equations. Eubank et al. demonstrate that their methods enable accurate modeling of a hypothetical spread of disease throughout a large, although in many ways artificial, population. This knowledge can in turn be used to seek an optimal emergency response policy. Using simulated scenarios, researchers have shown that some airports specifically play a more important role in the global pandemic than others, either because of the sheer number of passengers passing though them, or their connectedness to other airports [Colizza *et al.*, 2006; Epstein *et al.*, 2007]. Simulations are used by prior work to fill in missing data as well as to test possible future outcomes of a contagion.

The primary contribution of this paper is a connection between local interactions between individuals and aggregate patterns of disease spread without resorting to synthetic data. We propose a novel way of modeling the emergence and evolution of global epidemics from the behavior and mobility of tens of thousand of individuals. We concentrate on the scalability of our models both in terms of computational complexity and human effort required. By applying machine learning and statistical analysis to fine-grained online data, our framework allows monitoring and predictions of flu prevalence in a timely fashion and without active user participation.

Looking at an even 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.

Methodology and Models

In this section, we first discuss a method for automatic detection of Twitter users afflicted with an infectious disease using text classification.[5] We then verify that the flu signal inferred from geo-tagged Twitter data agrees with established measures of flu prevalence. Subsequently, we propose and evaluate a statistical model that captures people's fine-grained mobility, interactions, and health. The structure of this model provides interesting insights into the mechanisms of global spread of disease from the activities of individuals. Finally, we leverage the statistical model to predict future flu activity in specific cities. The following subsections describe each of these steps in detail.

Identifying Sick Individuals

In order to quantify the impact of sick passengers on the spread of flu, we first need to infer people's health status, 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. Building on previous work, we learn a linear support vector machine binary classifier C while directly optimizing the area under the ROC curve [Joachims, 2005; Paul and Dredze, 2011; Sadilek *et al.*, 2012a]. 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. Since the SVM operates in the space of English phrases containing up to three words, it significantly outperforms less sophisticated keyword-matching approaches and achieves 0.98 precision and 0.97 recall [Sadilek *et al.*, 2012a]. For example, C is not confused by a message such as "I am sick of this homework, it gives me a headache!" even though it contains potentially misleading keywords.

Validating Twitter Health Signal (T_f)

A major deficiency of the C_f and G_f measures is their coarse (weekly) temporal granularity. This is problematic as the spread of influenza-like illnesses is a highly dynamic process [Moser *et al.*, 1979]. We propose and evaluate a novel measure of *daily* flu intensity on day d in region r using our SVM-annotated twitter data

$$T_f(d,r) = \sum_{u \in \text{UsersTweetingAt}(d,r)} \Pr(u \text{ indicates sickness on day } d)$$

where we convert SVM predictions into sickness probabilities [Platt and others, 1999]. T_f approximates the expected number of sick users on a given day in a given region.

To evaluate the relevance of T_f, we collected geo-tagged tweets from five diverse metropolitan areas: New York City, Los Angeles, Boston, Seattle, and San Francisco. Figure 2 shows that T_f is strongly correlated with official statistics:

[5] 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). We will use the term "flu" to refer to this class of ailments.

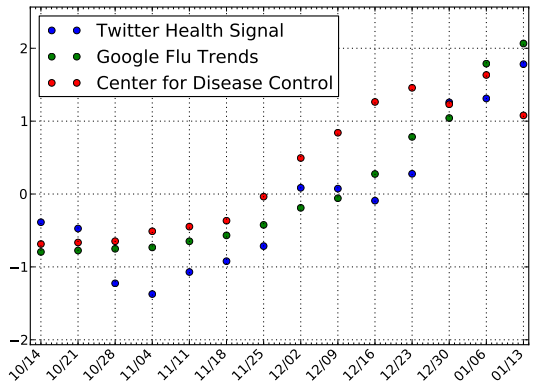

Figure 2: Three independent measures of flu prevalence show mutual agreement over a fourteen week time period. T_f (blue) is our proposed Twitter flu index.

CDC index C_f ($R = 0.80$, p-value: 0.002) and Google Flu Trends G_f ($R = 0.87$, p-value: 0.0002).

This agrees with prior work that finds a strong relation between signals relevant to flu-like disease mined from Twitter, official statistics, and Google Flu Trends in a number of countries [Culotta, 2010; Lampos *et al.*, 2010]. Our experiments below use all three signals (C_f, G_f, and T_f) as dependent variables in order to establish the predictability of official coarse-grained statistics as well as fine-grained metrics on the basis of social media data.

Understanding the Global Spread of Disease

We estimate the expected fraction of sick passengers p^s flying from airport x to y in a given time slice t:

$$\mathbb{E}[p^s(t, x \to y)] = \frac{f^s(t, x \to y)}{f(t, x \to y)}, \tag{3}$$

where f denotes the total number of Twitter users flying between two airports and f^s is the number of sick users on the same route.

Since official RITA[6] statistics for recent months are not available at submission time, we estimate the deviation in the number of passengers directly from Twitter data. We determine that a user flew from airport A_x to A_y on day d if he tweeted at A_y on day d and appeared at A_x on either day d or $d-1$. Since we collect tweets within a three kilometers of each airport, some of the users are not airplane passengers. They could simply live nearby. However, by considering target users who appear in more than one airport, we obtain a high-precision dataset of travelers.

Similarly, the influx of sick twitter users to airport x at time t is given by

$$\mathcal{I}(t, x) = \frac{\sum_{i \in \text{Airports} \setminus x} f^s(t, i \to x)}{\sum_{i \in \text{Airports} \setminus x} \frac{1}{D} \sum_{d=1}^{D} f^s(d, i \to x)}. \tag{4}$$

[6]Research and Innovative Technology Administration (RITA) manages detailed airplane travel statistics for the U.S. Department of Transportation: http://www.transtats.bts.gov/

Note that we express the influx as a fraction of a "typical day" baseline to account for differences in the traffic volume across geographical areas and times of day. While $\mathbb{E}[p^s]$ and $\mathcal{I}$ measure flow of passengers between airports, city-level statistics are calculated by aggregating over all airports in the corresponding metropolitan area.

Human contact is a crucial factor in the transmission of infectious diseases. The GPS coordinates embedded in the tweets allow us to estimate physical encounters (*meetings*) between specific individuals. We say two users met if they appear within 100 meters of each other within one hour. In general, the granularity of the data does not allow detection of specific meetings, but we do quantify the increased chance of spreading an illness either by direct or indirect (*e.g.*, as touching a seat at the gate) contact between people. As a result, we can propagate health risk among users in our model.

While features involving *total* users flying are simple to calculate, features involving *sick* users require machine learning in order to scale. As we will see, while features representing flight volume do explain a sizable fraction of the variance in official flu statistics, features inferred from noisy data provide a significant improvement and make accurate predictions possible.

There is an inherent time lag between contagion events and an observed outbreak. This is due to complex mechanisms involving incubation periods of the disease, the intensity of human contact, prevention measures instituted by authorities, and other factors. To model this delay, we formulate some of our features over a spectrum of time intervals and time offsets—ranging from a day to several weeks. Our model captures fine-grained geographical patterns (*e.g.*, daily flu predictions for a specific city) as well as coarse-grained aggregate patterns (*e.g.*, overall flu prevalence in the U.S.).

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 underestimated, but evaluation by prior work suggests that the latter effect is small [Sadilek *et al.*, 2012a]. Our observations also represent only a small fraction of air travelers. RITA estimates that there were roughly 643 million enplaned passengers in 2012, meaning we capture roughly 1/1,000 of them. However, the results in this paper indicate that by modeling the latent effects, such a sampling ratio is sufficient to accurately model public health across a large number of metropolitan areas. We note that currently used methods suffer from similar challenges as well. For example, infected people who do not visit a doctor, or do not respond to surveys are virtually invisible to the traditional methods.

About 12% of all U.S. adults use Twitter, but the user base skews towards younger, more urban people as compared to the general population [Smith and Bruenner, 2012]. Since the demographics of airplane travelers—who we approximate by Twitter users—are biased in similar ways, the skewing effect is mitigated [POST, 2000]. We believe our methods of automated text analysis coupled with fine-grained location

data is a valuable complement to traditional survey-based approaches to human behavior modeling. Users behave naturally because they are unaware they participate in a study. Moreover, the GPS data allows us to infer physical encounters even between individuals who have not noticed each other. For example, they are simply strangers sitting next to each other at the gate, waiting for their flight to begin boarding.

Experiments and Results

Flight Volume. It has been shown that aggregate air traffic volume can be a significant indicator of disease spread [Brownstein *et al.*, 2006; Colizza *et al.*, 2006]. We begin with regression analysis that considers only passenger features without regard for health state of each individual. Namely, we focus on the *total* number of users we observe flying between two airports (f). We compute these statistics across a spectrum of daily, weekly, and monthly time intervals and offsets as described above. We find that these features (which do not require AI techniques) account for 56% of the variance in the official Google Flu Trends data. As we will see, this is not enough to do reliable prediction that can guide important decisions.

Modeling Latent Health Features. By augmenting the model with features $\mathcal{I}$ that leverage *latent* information inferred by the SVM, the regression model explains 73% of the variance in Google Flu Trends. This is a significant improvement of 17 percentage points of variance explained, a nearly 30% improvement over previously considered methods. Interestingly, by modeling *only* the latent features and none of the non-sickness features, the amount of variance explained remains 73%. This shows that the raw passenger volume data does not explain any additional signal and is completely superseded by the latent factors. Statistics on sick travelers in the seven days preceding the day for which we make a prediction are the most dominant—explaining 71% of the variance alone.

Modeling Meetings. Thus far, we have ignored the effect of physical encounters between individuals. We now introduce an additional feature $\mathcal{M}(t, x)$ that counts the number of meetings between people traveling to airport x at time t. These exposure events are counted along the entire journey of each traveler including the final destination. Including $\mathcal{M}$ in the model explains additional 5% of the variance. Crucially, accounting for meetings *without* regard for health state of each individual does not explain any additional variance.

Predicting Disease Spread. We now turn to predicting future levels of flu prevalence in a specific area based on the flow of airplane travelers and their interactions. We learn a linear regression model that predicts T_f on a given day by leveraging features mentioned above (f, $\mathcal{I}$, and $\mathcal{M}$) over the preceding three days. Since the model operates with standardized variables, a *single* set of parameters captures all the cities in our dataset.

We compare the prediction made by our model that explicitly takes into account health of individual travelers to a baseline model that is oblivious people's health states. Otherwise, the baseline model uses the same type of features. We evaluate the model by predicting the last seven days in the dataset (a peak of this season's flu epidemic), while training on the remainder of the data. The health-conscious model is significantly more accurate—achieving 93% accuracy within a 95% confidence interval. By contrast, the baseline model is only 88% accurate at the same confidence level.

Conclusions and Future Work

Understanding the subtle mechanisms of epidemics is a major challenge for the field of computational sustainability. This paper explores prediction of the global spread of an infectious disease on the basis of fine-grained social network data—an important instance of the general problem of modeling emergent properties of large real-world dynamical systems. We use geo-tagged status updates of Twitter users as a noisy proxy for the flux of people into and out of a city. Instead of looking at simple statistics, such as the number of airplane passengers flying into a city, we apply machine learning to people's messages and infer important latent variables. This includes the volume of sick passengers, the number of people they physically encountered, and other features. We quantify the impact of a number of otherwise elusive factors on flu outbreaks, and show that flu prevalence in a given area can be accurately predicted by modeling people's mobility, while keeping track of their health over time with fine granularity.

Future work will focus on explicit modeling of the mechanisms underlying individuals' health states. Building on prior work on simulated populations, each person can be in a susceptible, infected, recovered, or other state. While the true state can only be observed when the person visits a doctor's office, we can model this latent variable using AI techniques by leveraging indirect evidence. In this work, we considered two states—healthy (susceptible) and sick (infected)—that are inferred from people's online communication. A promising direction for future work is to formulate important epidemiological problems as probabilistic graphical models. These models would be *informed* by timely evidence extracted from social media and the web at large. This paper makes the first steps in this direction and opens novel challenges. We hope that this will stimulate further collaboration of AI researchers and the broader scientific community on important interdisciplinary problems.

Acknowledgements

This research was partly funded by ARO grant W911NF-08-1-0242, ONR grant N00014-11-10417, OSD grant W81XWH-08-C0740.

References

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

[Brownstein *et al.*, 2006] J.S. Brownstein, C.J. Wolfe, and K.D. Mandl. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states. *PLoS medicine*, 3(10):e401, 2006.

[Chunara *et al.*, 2012] R. Chunara, J.R. Andrews, and J.S. Brownstein. 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, 2012.

[Colizza *et al.*, 2006] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015, 2006.

[Collier *et al.*, 2011] N. Collier, N.T. Son, and N.M. Nguyen. OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2011.

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

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

[Epstein *et al.*, 2007] J.M. Epstein, D.M. Goedecke, F. Yu, R.J. Morris, D.K. Wagener, and G.V. Bobashev. Controlling pandemic flu: the value of international air travel restrictions. *PLoS One*, 2(5):e401, 2007.

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

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

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

[Grais *et al.*, 2003] R.F. Grais, J. Hugh Ellis, and G.E. Glass. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *European journal of epidemiology*, 18(11):1065–1072, 2003.

[Hollingsworth *et al.*, 2007] T.D. Hollingsworth, N.M. Ferguson, and R.M. Anderson. Frequent travelers and rate of spread of epidemics. *Emerging infectious diseases*, 13(9):1288, 2007.

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

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

[Moser *et al.*, 1979] M.R. Moser, T.R. Bender, H.S. Margolis, G.R. Noble, A.P. Kendal, and D.G. Ritter. An outbreak of influenza aboard a commercial airliner. *American Journal of Epidemiology*, 110(1):1–6, 1979.

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

[Nicolaides *et al.*, 2012] C. Nicolaides, L. Cueto-Felgueroso, M.C. González, and R. Juanes. A metric of influential spreading during contagion dynamics through the air transportation network. *PloS one*, 7(7):e40961, 2012.

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

[Platt and others, 1999] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[POST, 2000] POST. Statistical information on air passenger numbers and characteristics http://www.parliament.uk/documents/post/e3.pdf. 2000.

[Ruan *et al.*, 2006] S. Ruan, W. Wang, S.A. Levin, et al. The effect of global travel on the spread of sars. *Mathematical Biosciences and Engineering*, 3(1):205, 2006.

[Sadilek *et al.*, 2012a] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

[Sadilek *et al.*, 2012b] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[Shalev-Shwartz *et al.*, 2007] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814. ACM, 2007.

[Smith and Bruenner, 2012] Aaron Smith and Joanna Bruenner. Pew research centers internet & american life project: Twitter use 2012. *http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx*, 2012.

[Zou and Hastie, 2005] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.