

**Online Social Network Sensors for Influenza
Outbreaks**

ARCHIVES

by

Katie Elizabeth Everett

B.S., Massachusetts Institute of Technology (2012)

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document
in whole or in part in any medium now known or hereafter created.

Author
Department of Electrical Engineering and Computer Science
September 9, 2013

Certified by
Prof. Munther Dahleh
Associate Department Head of Electrical Engineering and Computer
Science
Thesis Supervisor

Accepted by
Prof. Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Online Social Network Sensors for Influenza Outbreaks

by

Katie Elizabeth Everett

Submitted to the Department of Electrical Engineering and Computer Science
on September 9, 2013, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Previous research has shown strong correlations between postings on the online social network Twitter where users complain of influenza-like symptoms, and clinical data on actual influenza rates. In addition, previous research has shown that more popular individuals in a real-life social network are infected with influenza earlier than average individuals. We collect all flu-related tweets during the 2012-2013 influenza season in order to compare the timing of flu-related tweets from more popular users compared to less popular users. No difference is seen in flu tweet timing between Twitter users with a high number of followers compared to users with a low number of followers. Restricting the Twitter network to bidirectional edges (mutual followings) performs slightly better, but is still not significant. Future work should focus on identifying edges in online social networks that indicate that two users regularly come into close physical proximity.

Thesis Supervisor: Prof. Munther Dahleh

Title: Associate Department Head of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank Munzer Dahleh and Ammar Ammar for their supervision of this project.

Contents

1	Motivation	11
2	Background	13
3	Methods	17
4	Results and Discussion	19
4.1	Experiment #1	19
4.2	Experiment #2	21
4.3	Experiment #3	21
5	Future Work	25

List of Figures

4-1	Experiment #1	20
4-2	Experiment #2	22
4-3	Experiment #3 - Boston	24
4-4	Experiment #3 - Chicago	24
4-5	Experiment #3 - Houston	24

Chapter 1

Motivation

Worldwide, seasonal influenza causes three to five million cases of severe illness and 250,000 to 500,000 deaths each year.[14] Seasonal influenza reoccurs each year in regular cycles, but the geographic location, timing and size of each outbreak vary, which makes it difficult to produce reliable and timely estimates of influenza activity using traditional time series models. [17] In addition to seasonal influenza, a pandemic outbreak like the 1918 “Spanish Flu” could cause millions of fatalities if a new strain of the influenza virus were to develop, against which no prior immunity existed.[15]

Early detection of influenza activity, if followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza.[15] For example, research suggests that targeted mass prophylactic use of antiviral drugs can contain an epidemic if administered sufficiently early [6][11], which requires early detection to appropriately distribute the medications, which are limited in supply. Rapid, targeted administration of vaccines is a critical element of the World Health Organization’s pandemic influenza response plan. [15]

Chapter 2

Background

State of the art flu-tracking is real-time at best. The traditional surveillance approach used by the Centers for Disease Control and Prevention (CDC) is almost entirely manual, leading to a 1-2 week delay between the time a patient is diagnosed at a strategically selected medical practice, and the time that data point becomes available in aggregate reports. Several innovative surveillance systems have been proposed to monitor influenza activity in real-time. For example, Google Flu Trends [8] aggregates live online search queries for keywords relating to influenza, Espino et al. [4] propose monitoring call volumes to telephone triage advice lines, and Magruder et al. [12] track over the counter drug sales.

Research shows that flu-related Tweets (messages posted on the online social network Twitter) give a good approximation for the prevalence of influenza-like illness. Moreover, Twitter data is publicly available in real-time, unlike any of the previously mentioned data sources. Through Twitter's Streaming API, it is easy to collect all tweets containing flu-related keywords such as "flu", "sick", "influenza", "headache", "sore throat", and "fever". Invariably, this simple keyword-based collecting leads to some false negatives ("I'm feeling a bit under the weather today."), and a large number of false positives ("OMG I have such Bieber fever!!"). Recent papers solve this problem by learning document classification models on manually labeled training sets, resulting in strong correlations with ILI prevalence reported by the CDC. [3] [9] [16] [1]

We can characterize the spread of influenza as a random process over a network comprising edges representing real-life physical proximity between two people. Influenza spreads primarily from person to person through respiratory droplets from coughing and sneezing. [7] It can also spread indirectly, if, for example, a sick person sneezes into their hands, touches a doorknob, then a healthy person touches the same doorknob shortly afterwards. We can consider a network where people are nodes, and edges connect two people who came within close enough physical contact that influenza spreading was possible. Let us refer to this type of network as a “flu-spreading” network. We can now consider a random process over this network, where at each timestep, a person can be either susceptible to influenza, infected with influenza, or recovered from influenza (and therefore immune to this season’s strain). As we move forward in time, whenever an edge connects an infected node with a susceptible node, with some probability influenza is transmitted across that edge.

From a theoretical perspective, the problem of outbreak detection can be formalized as follows: Given a network and a dynamic process spreading over it, we wish to select a set of nodes to detect the process as efficiently as possible. Several objective functions may be relevant, such as minimizing detection time (so that you identify an outbreak as early as possible), minimizing population affected by undetected outbreaks, or detection likelihood (maximize the probability that we detect an outbreak at all). Leskovec et al show that optimizing any of these objective functions is NP-hard, so we cannot expect to find the optimal solution for a large network.[10] However, they also show that these objective functions exhibit *submodularity*, a diminishing returns property stating that adding an extra node u to a set S of sensors has less benefit than adding u to a sensor set S' , if $S' \subset S$. It is known that for submodular optimization functions on networks, a $1 - 1/e$ approximation can be obtained using the simple greedy strategy of always adding the node that increases the objective function the most.[13] Leskovec et al. give another algorithm that is up to 700 times faster than the simple greedy strategy, and also prove online bounds on the quality of a sensor set chosen by any algorithm.[10]

Even finding an approximate solution to the outbreak detection problem requires

calculations across the entire network topology, which is impractical on a network as large as Twitter. Instead, we can use techniques that only require local knowledge of the network structure to sample sets of nodes that are more central than average.

The *friendship paradox* states that, if you are an average person, your friends have more friends than you do.[5] This was originally an observation about human social networks, but it can also be shown mathematically to be true. Consider a social network represented by an undirected graph where each person is a node and each friendship is an edge. Let u be a node picked uniformly at random from the network. The expected degree of u is simply μ , the mean degree of the network. But, if we then pick v uniformly at random from the neighbors of u , we can show that the expected degree of v is not the mean degree, but $(1 + \sigma^2)\mu$, where σ^2 is the variance of degree in the network.

For graphs such as social networks, this suggests a clever method for sampling individuals who have more friends than average. First, pick a group of people at random, then for each member of the “random” group, select one of their friends at random, creating the “friends” group. The “friends” group will have higher than average degree. In social networks, higher degree people are more central, or in social terms, more popular than average.

Now we have seen that in theory, central nodes in a social network structure are infected earlier by a dynamic process spreading over the network, and that it is possible to sample nodes that are more central than average, using only local information about the network. But, how does it relate to influenza spreading in practice?

During the 2009 H1N1 swine flu outbreak, Christakis and Fowler [2] used the friendship paradox technique to sample Harvard students, by selecting the “random” group from the student directory, and asking each of these students to name a friend, creating the “friend” group. They monitored both groups for the duration of the flu season, and tracked when students had flu symptoms. As hypothesized, the “friends” group acted as sentinels in the network, getting sick two weeks earlier on average than the “random” group.

This suggests that the edges in real-life social networks, at least in a fairly closed population such as a college campus, are a good approximation for the edges in the flu-spreading network on that population. In addition, it agrees with our prediction that more central members of the network should become infected earlier than average.

Therefore, the goal of this project is to answer the following question: Do more central Twitter users tweet about having the flu earlier than average Twitter users? If so, then we can use Twitter as a novel data source for influenza outbreak prediction, rather than simply for outbreak detection.

It should be noted that we do not attempt to distinguish postings reflecting true influenza infections from those reflecting influenza-like illness, which is difficult for even a physician to do without running a diagnostic test. Instead, we build on the body of research work showing a strong correlation between online social network postings about influenza-like symptoms, and medical records of clinical influenza rates. [3] [9] [16] [1] In any case, being able to monitor and predict influenza-like illness is still valuable.

Chapter 3

Methods

Twitter.com is a micro-blogging service that allows users to post messages of 140 characters or fewer. A person can subscribe to the feed of messages posted by another user by “following” that user. Following relationships on Twitter are unidirectional - User A can follow User B, and User B may or may not follow User A. The high message posting frequency enables up-to-the-minute analysis of an outbreak. As compared to search engine query logs, Twitter messages are longer, more descriptive, and generally more publicly available. Twitter profiles often contain semi-structured meta-data (city, state, gender, age), enabling a detailed demographic analysis. Despite the fact that Twitter appears targeted to a young demographic, it in fact has quite a diverse set of users. The majority of Twitter’s nearly 10 million unique visitors in Feb 2009 were 35 years of older. A nearly equal percentage of users are between ages 55 and 64 as are between 18 and 24.

We maintain a persistent connection to the Twitter Streaming API from December 5, 2012 until March 1, 2013, collecting all tweets containing any of the following keywords: flu, influenza, headache, sore throat, fever, cough. For each tweet, the API returns a username and timestamp in addition to the text of the post. From the author’s username we can get additional profile details that are attached to each user, including number of followers, number of friends (or followings), profile creation date, and location. We exclude retweets, which are tweets originally posted by one user that is “retweeted”, or forwarded, by another user. Since retweets do not indicate a

new case of ILI, they are removed from analysis.

We run three experiments on this dataset. In all the experiments, we select a subset of users who are highly central, and determine whether those users tweeted about flu symptoms earlier, on average, than random users. In our first experiment, we compare users in the highest quartile for number of followers to random users. In our second experiment, we take the intersection of the follower and friend lists to get a list of mutual followers. We then compare users in the highest quartile for number of mutual followers to random users. In our third experiment, we perform a similar analysis as the second experiment, but restricted to users with a specific city listed in the location field of their profile.

Chapter 4

Results and Discussion

4.1 Experiment #1

For our first experiment, we partition Twitter users into two cohorts - low followers, who have fewer than 100 followers, and high followers, who have at least 100 followers. We then examine the cumulative occurrence of flu tweets from December 2012 to February 2013, during the seasonal influenza outbreak in the United States (where the majority of English language tweets originate). The graph in Figure 4-1 plots the fraction of all (eventual) flu tweets for the cohort of Twitter users that have happened by the date on the x-axis.

If the seasonal influenza outbreak were affecting high follower users earlier than low follower users, we would expect to see the red curve shifted to the left. Instead, we see that the two curves are essentially identical (Kolmogorov-Smirnov statistic = 0.0119, p-value=1.00), indicating that there is no difference in influenza timing between the two cohorts. This implies that most edges on Twitter do not indicate that the two users come into close physical proximity on a regular basis, and are therefore not of use for influenza outbreak prediction. This result is not surprising, given that Twitter users often follow celebrities and news outlets, which are not people they would regularly come into close physical proximity with, which is necessary for spreading influenza.

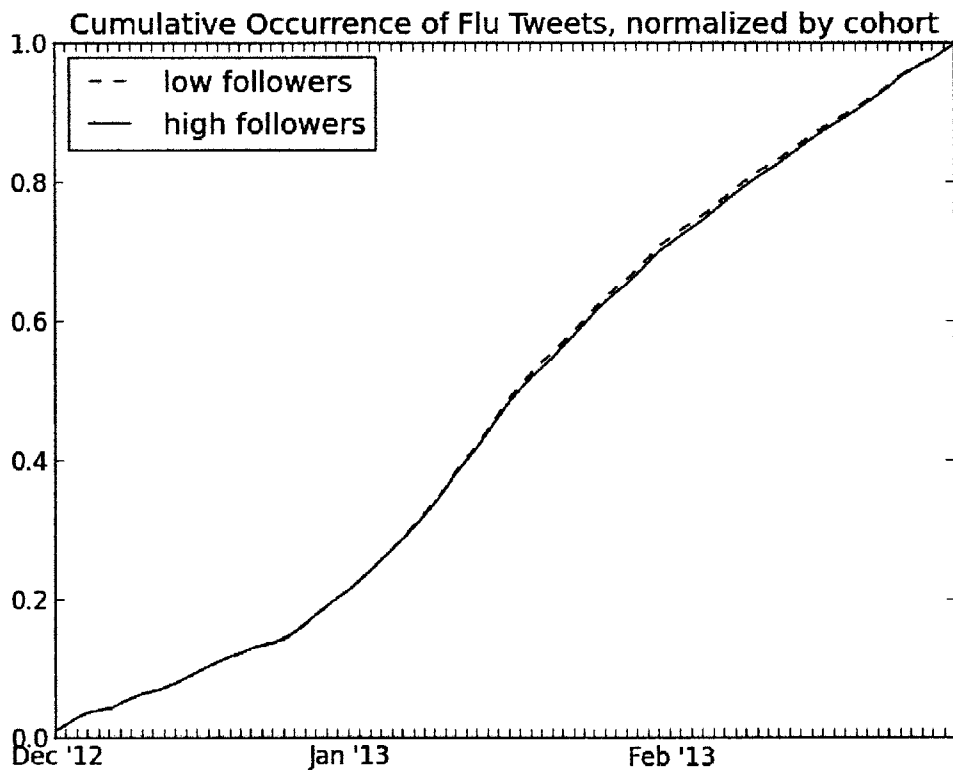


Figure 4-1: Experiment #1

4.2 Experiment #2

In our second experiment, in order to identify users who regularly come into close physical proximity, we will try restricting the Twitter network to the edges that are bidirectional, known as mutual followings. A mutual following, where User A follows User B, and User B follows User A, is more likely to indicate a social relationship between the two users. We perform a similar partitioning of Twitter users into two cohorts, those with a low number of mutual followers, and those with a high number of mutual followers. Figure 4-2 shows the cumulative occurrence of flu tweets in these two cohorts over the course of influenza season.

Again, if users with a high number of mutual followings were affected by the seasonal influenza outbreak earlier than users with a low number of mutual followings, we would expect to see the solid curve shifted to the left of the dashed curve. While there is a slight shift visible that we did not see in the first experiment, this difference is not statistically significant (Kolmogorov-Smirnov statistic = 0.071, p-value = 0.979), implying that our heuristic of restricting to mutual followings was insufficient to identify edges in the Twitter network that indicate that the two users regularly come into the close physical proximity required for influenza transmission.

4.3 Experiment #3

It is possible that even if mutual followings on Twitter indicate a social relationship between two users, that the users are living far apart and might have regular contact on online social networks, but not regular face-to-face contact. Twitter users can list a city as their profile location. Unfortunately, due to rate limits on the Twitter API, we are not able to make a sufficient number of profile location queries to restrict our dataset to mutual edges in which both users list the same city as their location. Instead we will count all mutual followers as edges, as in Experiment #2, but analyze flu timing in each city separately. We choose Boston, Chicago, and Houston for this experiment because they are large cities with only one word in their title, which makes

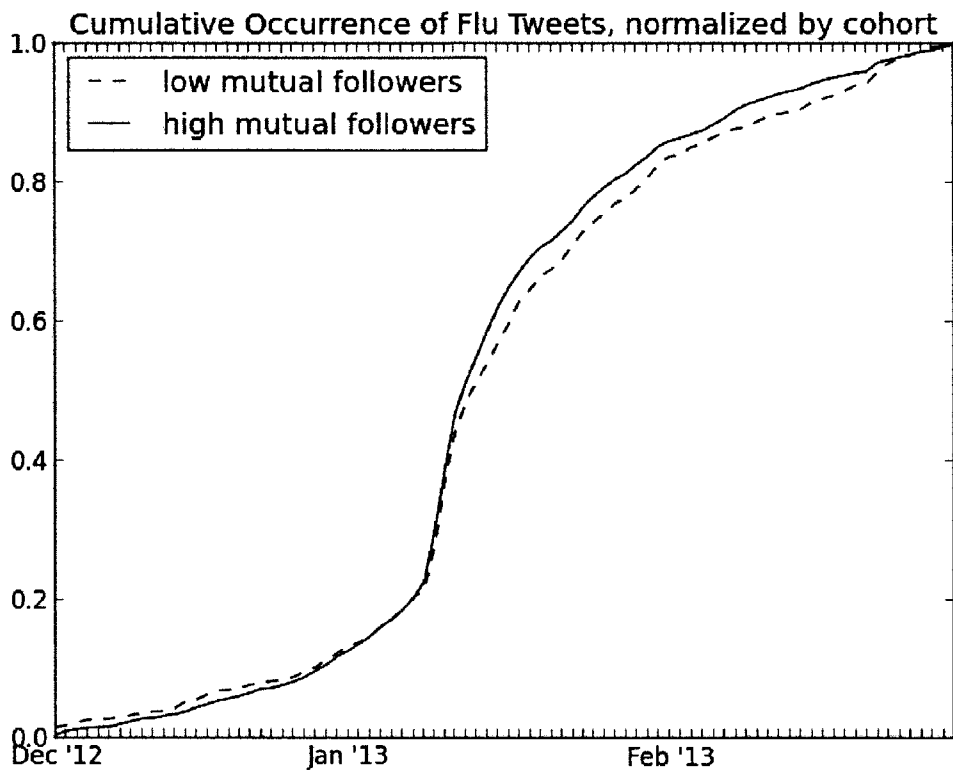


Figure 4-2: Experiment #2

parsing location strings easier.

We see once more in Figure 4-3 that there is no significant difference in influenza outbreak timing between our cohorts, indicating that our methods from this project are not useful for influenza outbreak prediction. We will discuss alternative approaches that may prove more fruitful in the next section.

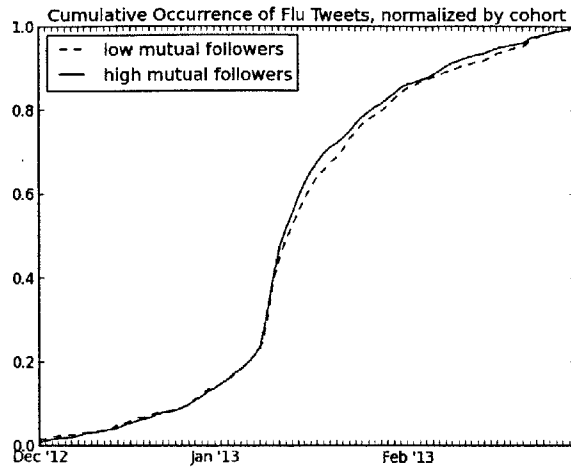


Figure 4-3: Experiment #3 - Boston

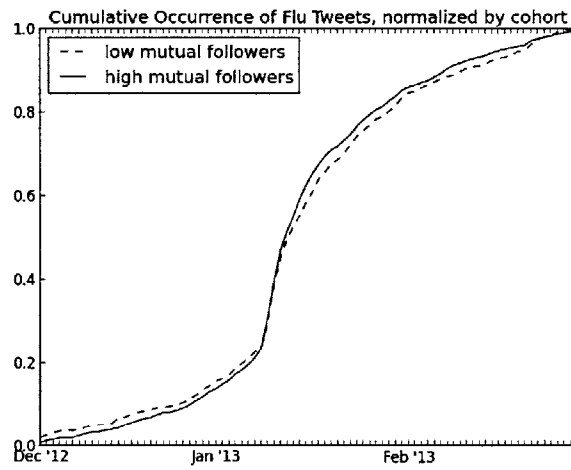


Figure 4-4: Experiment #3 - Chicago

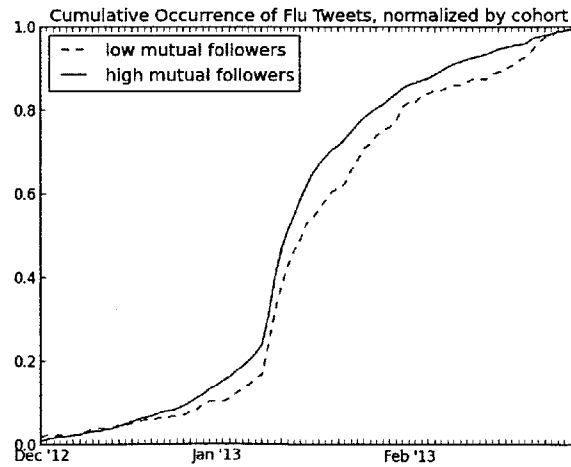


Figure 4-5: Experiment #3 - Houston

Chapter 5

Future Work

Future work should focus on ways of using publicly available real-time data that more accurately approximate the edges that influenza actually spreads over in real life. Here we present several alternative approaches using Twitter data, as well as several approaches using other data sources.

For the majority of Twitter users, most of the follower/following edges are not reciprocal, indicating that these edges primarily spread information, which often flows in one direction, rather than social interactions, which generally flow in both directions. These users may also have some mutual followings, but the mutual followings make up a small fraction of their total number of followings. However, there are some Twitter users who almost exclusively engage in mutual followings. It is possible that these users are using the Twitter platform in a different manner, more for social interactions than for obtaining information. Further investigation could attempt to identify cohorts of these users whose Twitter relationships may more accurately reflect their social interactions in real life, which would be more useful for predicting disease outbreaks.

The online social network Facebook is an obvious alternative source of data. The primary disadvantage compared to Twitter is that most Facebook users set their profiles as private, making data collection much more challenging. In particular, it would be interesting to use photos posted on Facebook as a data source. Facebook users can upload photos, and then "tag" them, which labels the faces in the photograph

with people's names. In addition, facial recognition technology can often identify the people in photos without the need for manual tagging. Users who appear together in a Facebook photo are likely people who regularly interact in real life.

Another possible data source is the location-based platform Foursquare. Foursquare users can use their smartphone to "check in" at a location such as a restaurant, concert, or sporting event. Based on previous a user's previous check-ins, Foursquare makes personalized recommendations to the user about places and events they might be interested in. This geolocation data could be tremendously useful for disease outbreak prediction, but again, privacy concerns and data availability are a major challenge.

Bibliography

- [1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. *The First International Workshop on Cyber-Physical Networking Systems*, pages 702–707, 2011.
- [2] N A Christakis and J H Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.
- [3] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.
- [4] Jeremy U Espino, William R Hogan, and Michael M Wagner. Telephone triage: a timely data source for surveillance of influenza-like diseases. In *AMIA Annual Symposium Proceedings*, volume 2003, page 215. American Medical Informatics Association, 2003.
- [5] Scott L Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477, 1991.
- [6] Neil M Ferguson, Derek a T Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meechai, Sopon Iamsirithaworn, and Donald S Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–14, September 2005.
- [7] Centers for Disease Control and Prevention. Seasonal influenza, 2013.
- [8] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
- [9] Vasileios Lampos, Tijn De Bie, and Nello Cristianini. Flu detector-tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. Springer, 2010.
- [10] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

- [11] Marc Lipsitch, Ted Cohen, Ben Cooper, James M Robins, Stefan Ma, Lyn James, Gowri Gopalakrishna, Suok Kai Chew, Chorh Chuan Tan, Matthew H Samore, David Fisman, and Megan Murray. Transmission dynamics and control of severe acute respiratory syndrome. *Science (New York, N.Y.)*, 300(5627):1966–70, June 2003.
- [12] S Magruder. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL technical digest*, 24(4):349–53, 2003.
- [13] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.
- [14] World Health Organization. WHO Influenza Fact Sheet.
- [15] World Health Organization. Development and implementation of a national deployment and vaccination plan for pandemic influenza vaccines, 2012.
- [16] Michael J Paul and Mark Dredze. A Model for Mining Public Health Topics from Twitter. 2010(May 2009), 2010.
- [17] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 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):e19467, 2011.