

Vision: Towards Real Time Epidemic Vigilance through Online Social Networks

Introducing SNEFT – Social Network Enabled Flu Trends

Lingji Chen [†] Harshavardhan Achrekar ^{*} Benyuan Liu ^{*} Ross Lazarus [‡]

[†] Scientific Systems Company Inc
500 West Cummings Park
Woburn, MA 01801
lingji.chen@ssci.com

^{*} Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854
{hdachrekar, bliu}@cs.uml.edu

[‡] Department of Population Medicine
Harvard Medical School
Boston, MA 02101
ross.lazarus@channing.harvard.edu

ABSTRACT

Our vision is to achieve faster and near real time detection and prediction of the emergence and spread of an influenza epidemic, through sophisticated data collection and analysis of Online Social Networks (OSNs) such as Facebook, MySpace, and Twitter. In particular, we present the design of a system called SNEFT (Social Network Enabled Flu Trends), which will be developed in a 12-month SBIR (Small Business Innovation Research) project funded by the National Institutes of Health (NIH). We describe the innovative technologies that will be developed in this project for collecting and aggregating OSN data, extracting information from it, and integrating it with mathematical models of influenza. One of the monitoring tools used by the Centers for Disease Control and Prevention (CDC) is reports of Influenza-Like Illness (ILI) cases; these reports are authoritative but typically have a delay of one to two weeks due to the largely manual process. We describe the SNEFT prototype in the context of predicting ILI cases well in advance of the CDC reports. We observe that OSN data is individually noisy but collectively revealing, and speculate on other applications that can potentially be enabled by OSN data collection and analysis.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Design, Measurement, Performance

Keywords

Flu trends, online social networks

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond (MCS), June 15, 2010, San Francisco, USA.

Copyright 2010 ACM 978-1-4503-0155-8 ...\$5.00.

1. INTRODUCTION

Seasonal influenza epidemics result in about three to five million cases of severe illness, and about 250,000 to 500,000 deaths worldwide each year [16]. Other pandemics such as the H1N1 influenza, though relatively benign in the 2009-2010 season, may in the coming years change into a devastating one similar in scale to the deadly 1918 “Spanish Flu” epidemic. Reducing the impact of these threats is of paramount importance for public health authorities, and studies have shown that effective measures can be taken to contain the epidemics, if early detection can be made [11, 19].

How will we know when an epidemic is emerging? How can the transition from a “normal” baseline situation to an epidemic be detected in real time? The Centers for Disease Control and Prevention (CDC) monitors Influenza-Like Illness (ILI) cases by collecting data from sentinel medical practices, collating reports and publishing them [4]. Although highly authoritative because diagnoses are made and reported by doctors, the system is almost entirely manual, resulting in a 1-2 weeks delay between the time a patient is diagnosed and the moment that data point becomes available in aggregate ILI reports. Public health authorities need the earliest possible warning to ensure effective intervention, and therefore more efficient and timely methods of estimating influenza incidence are urgently needed.

It would be beneficial for public health monitoring if every individual who becomes ill with influenza creates a public announcement of her condition such as “I am down with flu” and “I’m coughing and sneezing and feeling sick.” Although many of these data would be “noisy” *individually*, in the sense that not everyone who sneezes and coughs is infected with an influenza virus, in *aggregate*, they provide a source of data that can be transformed into a large scale picture of the underlying epidemic pattern in time and space. If these narratives were available in real-time, an automated system based on aggregation and processing would be far more timely than the current manual CDC system, permitting far earlier detection and warning of an impending epidemic.

Indeed, many individuals have already been creating these and similar electronic announcements every day, posting them on Facebook, MySpace, Twitter, and other Online Social Networks (OSNs). OSNs have become popular platforms for people to make connections, share information, and interact with each other. To date, OSNs boasts more

than half a billion users worldwide. Currently the two largest sites are the Facebook with about 400 million users and MySpace with approximately 200 million users. Twitter has about 80 million users and is growing fast. For a large number of users, logging on to their OSN accounts and interacting with friends have become an important part of their daily activities. They use OSNs to talk about their daily events, health status, entertainment, etc. At any time, tens of millions of users are logged on, with each user on average spending tens of minutes daily on the sites. Each week billions of pieces of content (blog posts, web links, photos, etc) are posted and shared between friends.

Data collected from OSNs represent a previously untapped data source for detecting the onset of an epidemic and predicting its spread. Google Flu Trends uses web search terms such as “influenza complication” and “cold remedy” collectively to predict the onset of the annual influenza season 1-2 weeks ahead of the CDC ILI data (and 3 or more weeks ahead of the reports). It can therefore be expected that the “I am down with flu” and “get well soon” messages exchanged between OSN users and their friends provide even earlier and more robust predictions. There have been efforts in utilizing twitter data for predicting “national mood” [9] and diseases in cities [21]. The latter application includes flu; however, the application appears to be inactive, and in our opinion, it lacks adequate statistical analysis, the danger of which has been discussed in [20].

Success for influenza prediction using novel data sources have been reported in literature, e.g., transportation data is used in [1] and currency tracing data is used in [3]. Our SNEFT work represents another approach, and it may have advantages in timeliness of early detection. Since the major OSN sites offer users with real-time search capability, SNEFT will provide a snapshot of the current epidemic condition and a preview of what is coming next, *on a daily or even hourly basis*. This will greatly enhance the responsiveness to influenza epidemic by public health authorities. Such a system may become a valuable tool for public health authorities, for example by becoming part of the CDC’s BioSense program [12].

Also, the user demographic and geographic location information obtained from OSN sites present significant opportunities for us to carry out fine-grained analysis based on people’s age, gender, location, etc. Many OSN users provide their demographic and location information on their profiles. Moreover, a growing number of mobile Twitter clients are tagging their tweets with latitude and longitude of the device at the time the tweet was generated. Such geotagged content will help the proposed work by improving the availability and accuracy of the geographic information tied to influenza related OSN posts.

This paper is organized as follows. Section 2 describes the overall system design and the four functional units. Section 3 describes in detail the data collection process. The mathematical problems of detection and prediction are formulated and described in Section 4. Conclusions are drawn in Section 5 with a discussion of future work.

2. SYSTEM DESIGN

The system architecture of SNEFT is schematically shown in Figure 1 and the main components of the system are listed as follows.

Data collection: CDC ILI reports and other influenza re-

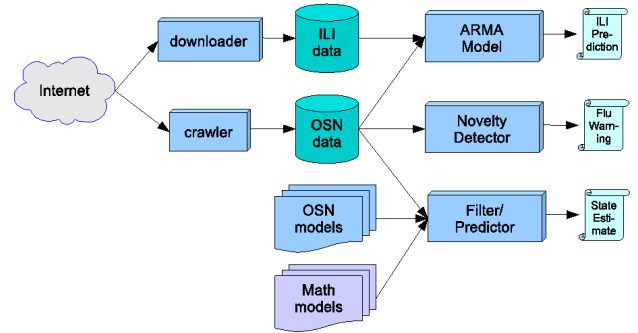


Figure 1: The system architecture of SNEFT.

lated data will be downloaded from the corresponding web-sites. We will choose a list of keywords that are likely to be of significance, and use OSN public search interfaces to collect relative keyword frequencies. Relevant information will be retrieved and stored in a spatio-temporal database for further data analysis. The data collected include user profile status update, blog entries, and tweets (for Twitter) that contain the search keyword, and the corresponding time stamps and user demographic (age, gender, etc) and geographic location information. We will only collect the data that are publicly available on the online social network sites. The collected data will be inserted into a database for aggregate queries. Only aggregated anonymous information (e.g., keyword frequencies for a given geographical location) will be obtained. No user identities or identity connections will be used in the study.

Novelty detection: Detecting the transition from a “normal” baseline situation to a pandemic in real time is a challenging task. This transition is likely to be represented in the volume and content of OSN data at a very early stage. We will develop novelty detection techniques [5] to monitor OSN data and detect the time when “something is different,” so as to provide a timely warning to public health authorities for further investigation and response.

ILI prediction: We will build an Auto-regressive Moving Average (ARMA) model to predict ILI incidence as a linear function of current and past OSN data and past ILI data. Such a model will be systematically built and verified using historical data, and the prediction will provide a valuable “preview” of ILI cases well ahead of CDC reports.

Integration with mathematical models: Mathematical models of influenza have been developed over the past 50 years and played an important role in furthering our understanding of the dynamics of influenza spread and the effect of intervention [8, 2]. Such models typically have many parameters that are hard to obtain precisely, and are often determined by fitting historical data. We will first build an “OSN sensor model” which describes “what would be observed on OSN if the population is infected as such and such.” We will then integrate real time OSN data with the prediction of mathematical models, to obtain a posterior estimate of the “infected state” of the population. Possible parameter values that are not consistent with OSN observations are weighted less, while those that are consistent are weighted more. Thus OSN data helps to “sharpen” the prediction of mathematical models.

3. OSN DATA COLLECTION

In this section we give a detailed description of the data collection process. Facebook, MySpace, and Twitter each provides a search interface (API) that allows a user to enter any keywords such as “flu” to search for blogs/tweets that contain the keywords. The APIs typically return a list of blog entries across the whole site that contains the keywords in reverse-time order. Each entry contains the following information: publisher profile ID, time stamp of the blog, and the corresponding blog content. The individual fields can be obtained using a HTML content scraper. Given a profile ID, we can retrieve the publisher profile information from the OSN sites, which usually includes the name, age, gender, geographic location, friends/followers, and other information.

Based on the search APIs provided by the OSN sites, we will develop an OSN crawler for each of three OSN sites. In the following, we will describe the design of Facebook crawler in details. While MySpace and Twitter have different search APIs, return result formats, privacy settings, and underlying Web technologies, the architecture and main components of the crawlers for these two sites are expected to be similar to that of Facebook.

The structure of the Facebook crawler is depicted in Figure 2. The function of each component is described as follows.

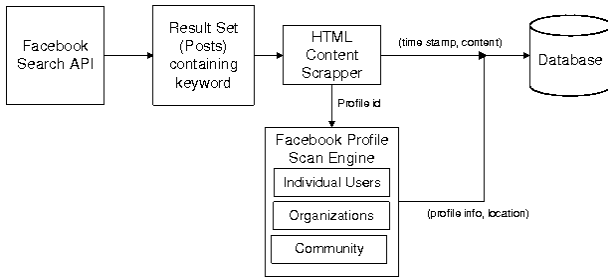


Figure 2: Design of the Facebook data collection engine.

- *Facebook Search Engine:* After signing in Facebook with a valid account, we can enter keywords and search for updates and posts that contain the keywords. The “Post by everyone” option allows us to search for given keywords in updates and posts of users on Facebook.
- *Results Set containing keyword:* When users post updates on Facebook they are given a few options including friends, group, and everyone. The “everyone” option (default setting) makes corresponding updates available to public and the Facebook search engine. All results that show up are available to the public for a limited time period.
- *HTML Content Scraper:* HTML Content scraper is a screen scraper for web pages. We are interested in getting useful information out of posts that are returned from the keyword search. The HTML content is sent as input onto regular expression matcher and techniques of pattern matching are applied to extract relevant content. We are particularly interested in the following three fields: profile ID, time stamp of the post, and the post content.

- *Facebook Profile Scan Engine:* Given a profile ID, we will retrieve the detailed information of the profile, which typically includes, among other things, name, gender, age, affiliations (school, work, region), birthday, location, education history, and friends. Note that a profile may belong to an individual user, an organization, or a community.

The information collected in this process will be aggregated and fed to a database for storage, indexing, and querying for later analysis. The most important fields of each record include the profile information (name, gender, age, etc), location, time stamp, and blog content.

The OSN data collection is affected by the following factors. *Search rate limit:* OSN sites often impose rate limits for their search APIs. The number of search requests originating from a given IP address or account are counted against the search rate limiter. *Return results limit:* The search APIs provided by the OSN sites only returns a limited number of blog entries or results within a limited time frame. For example, the search API of Twitter returns only the most recent 10 day’s tweets that contain the given keywords. *User activity pattern:* The number of posts varies at different time of a day. Our results show that there exist significant disparities in user activities between different hours of the day, days of a week, and special holidays such as Valentine’s Day, Mother’s Day, Thanksgiving, and Christmas. Therefore, it is necessary to carefully schedule the search time to guarantee that we obtain the complete set of blog posts containing the keywords, and there is no gap in the collected data.

Because of these constraints, if we wait for too long between adjacent searches, the number of blogs posted during the time interval may grow beyond what is allowed by the return limit and some of the results will be lost, leaving a gap in our collected data. To ensure that we retrieve the complete set of blogs containing the search keywords, we propose the following data collection scheme:

- To mitigate the search rate limit constraint, we will launch multiple concurrent search sessions from different IP addresses. These current search sessions will coordinate among themselves and collect data at different time intervals so that each session is within the search rate limit.
- The number of active search sessions will be dynamically adjusted according to the volume of the returned search results. To achieve this, we first estimate the volume of search results using an exponentially weighted moving average (EWMA) scheme. Denote the estimated average and current search result volume at search round k by $v(k)$ and $u(k)$, respectively, the moving average of the search result volume is computed as follows:

$$v(k) = \alpha v(k-1) + (1 - \alpha)u(k)$$

where α is the smoothing factor that reflects the weight of the previous estimate.

- Based on the estimated search result volume, sessions that are still active compute the required search rate to collect the data. If the required rate exceeds the rate limit, new search sessions will be triggered to share the load. When the search result volume becomes lighter, the number of active search sessions will be reduced.

In Twitter, mobile clients can tag their tweets with the location (latitude and longitude) of the device at the time the tweet was generated. Such geotagged content will improve the availability and accuracy of the geographic information tied to influenza related OSN posts. With the fast growing OSN access from mobile phones, it is likely that other OSN sites may provide similar geotagging functionality. Together with user demographic information obtained from user profiles, our system will allow us to carry out fine-grained analysis based on people’s age, gender, location, etc.

4. DETECTION AND PREDICTION

In this section we provide mathematical problem formulation for detection and prediction.

Novelty Detection

A novelty detector that monitors OSN data and issues a warning when “something is different” will be a valuable tool in monitoring influenza epidemic. We briefly describe the mathematical problem as follows. For ease of exposition let us “stack up” all relative frequencies of the chosen list of keywords at a given time t for a given region, and denote the vector by $y(t)$, $t = 1, 2, \dots, N$. Let us also assume that we have obtained some CDC ILI reports that captured the transition from a normal baseline situation to an epidemic situation, and we have identified the transition time around $t = k^*$. Then we can pose the following problems:

1. Find certain characterization of $y(t)$, say $c(t)$, that can be automatically computed, such that from $c(1), c(2), \dots, c(k^*)$ to $c(k^* + 1), \dots, c(N)$, there is a distinctive “jump.” Note that $y(t)$ is of high dimension because the list of keywords can be long, but ideally $c(t)$ should be a scalar that can be easily “thresholded.”
2. Identify a small subset of $y(t)$ for the above novelty detection task. As a hypothetical example, the relative frequency of “get well” messages alone may be sufficient. More realistically, the combination of several important keyword frequencies may be needed to detect the change occurring around time $t = k^*$.
3. Verify the above novelty detector on “test data,” e.g., data for a different season that is not used in constructing the novelty detector.

Many techniques of novelty detection parallel those for anomaly detection [5]. We will investigate histogram based method and clustering based method as a starting point, and move to more sophisticated methods as need arises.

ILI Prediction

In view of the lag inherent in CDC’s ILI reports, and the success of Google Flu Trends [13], it can be expected that a timely prediction of ILI percentage will be made by utilizing OSN data. More specifically, let $z(t)$ denote ILI percentage as reported by CDC, $t = 1, 2, \dots, N$. The OSN data and the ILI data are collected at different time scales; this multi-scale aspect will be fully investigated in the proposed effort, but in this section, for simplicity, we use the same scale to describe the approach. Our objective is to fit a model of the form

$$z(t+1) = \alpha_0 z(t) + \dots + \alpha_p z(t-p) + \beta_0^T y(t) + \dots + \beta_q^T y(t-q),$$

so that prediction of the *future* ILI percentage can be made *now*.

Since $y(t)$ is of high dimension, we will identify a low dimensional subset of $y(t)$ to use in the above ARMA model, instead of $y(t)$ itself. One brute-force way of achieving this is by exhaustive search: Pick a subset and perform the fitting, and we can obtain a metric of the “fitting error,” e.g., sum of squared errors. Then the best subset can be chosen.

Note that for each fitting, the problem of deciding p and q , the number of delayed values, as well as the coefficient vector α and β , have to be solved. There are standard approaches to these problems, see for example [18].

Nonlinear Filtering with Mathematical Models

For concreteness of the discussions that follow, consider a simple Susceptible-Infectious-Removed (SIR) model [2], where the dynamics of the population in each compartment is described by

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad R = N - S - I, \quad (1)$$

with N being the total population, β the transmission rate and γ the recovery rate.

We will denote by $x(t)$ the “state” of the population, which in this case is given by $x = [S, I]^T$. We will also denote by θ the parameter vector used in the model, which in this case is given by $\theta = [\beta, \gamma]^T$. In a more sophisticated *patch model* [8, 2] that accounts for geographical spread of infection, x would consist of the susceptible and the infected populations in each patch, and θ would include additional parameters such as travel rate, birth rate and death rate. In a *stochastic model*, $x(t)$ would be a random vector instead of a deterministic one.

For the purpose of the proposed research, a stochastic mathematical model gives the following transition probability regarding the disease spread $\text{Prob}(x(t+1)|x(t), \theta)$, where for simplicity we have used a discrete time formulation. Usually the parameter θ is not known precisely and can be estimated from historical data [17, 15, 22, 14]. Let $p_0(\theta)$ denote the prior distribution of the parameter θ . Starting with $\text{Prob}(x(0))$ and proceeding according to (1), the distribution of future state $x(t+1)$ can be predicted at time t .

By observing OSN data $y(1), y(2), \dots, y(t)$, we can improve upon the prediction. First we need to establish a “sensor model” for the observed OSN data $y(t)$, i.e., the probability $\text{Prob}(y(t)|x(t))$. We can then solve the filtering problem of obtaining the posterior distribution of $x(t)$ based upon the observations until time t .

This posterior prediction is “sharper” in the sense that some values of the initial state $x(0)$ and the parameter θ that are not consistent with the observed values $y(t)$ are “discounted” in the distribution of the future state.

An accurate sensor model $\text{Prob}(y(t)|x(t))$ is hard to build because record of $x(t)$ is usually not available. However, we believe that even a coarse sensor model will provide valuable information. For example, based on an estimate of the percentage of households who have both teenagers and computers, and analysis of OSN user behaviors, we may come up with an estimate of a typical y value, *given, for example, that x contains 200 infected people in a chosen geographical region*. We further postulate that it’s the *relative*, not *absolute*, amplitude of y that matters the most in the filtering solutions.

The nonlinear filtering and prediction problem can be solved using Sequential Monte Carlo methods (particle filters) [10]. The problem with unknown parameter θ can be dealt with by augmenting the state space through introducing a random walk, or by Expectation Maximization [7].

The result of filtering/prediction is an improved estimate of the underlying state x . This will enhance the inference conducted and actions taken that are traditionally associated with the use of mathematical models of influenza. Our approach provides a new framework of incorporating new types of data into established mathematical models, and new applications may emerge from it.

5. CONCLUSIONS AND FUTURE WORK

In this paper we described our vision to achieve faster and near real time detection and prediction of the emergence and spread of an influenza epidemic, through sophisticated data collection and analysis of OSNs such as Twitter, Facebook and MySpace. In particular, we presented the design of a system called SNEFT, which will be developed in a 12-month SBIR project funded by NIH. We described the innovative technologies that will be developed in this project for collecting and aggregating OSN data, extracting information from it, and integrating it with mathematical models of influenza. We presented in detail the ongoing OSN data collection work, and provided mathematical problem formulations for the detection and prediction tasks.

We observe that OSN data is individually noisy but collectively revealing. This is not unlike the accelerometer chip data in a laptop: Individually it is random, but collectively they can be used to detect earthquakes [6]. We believe that sophisticated data extraction and analysis of OSNs have the potential to be used in disaster relief, supply chain management, as well as epidemic vigilance.

6. ACKNOWLEDGMENTS

This project has recently been selected for funding by the National Institutes of Health (NIH) under the Small Business Catalyst Awards for Accelerating Innovative Research program. The contract is currently under negotiation with the expected start date in June, 2010. The authors would like to thank the reviewers of the NIH and ACM MCS Workshop for valuable feedbacks.

7. REFERENCES

- [1] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Broeck, V. Colizza, and A. Vespignani. Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(1):45+, Sep 2009.
- [2] F. Brauer, L. J. S. Allen, P. V. den Driessche, and J. Wu, editors. *Mathematical epidemiology*. Springer, 2008.
- [3] D. Brockmann. Human mobility and spatial disease dynamics. In H. G. Schuster, editor, *Reviews of Nonlinear Dynamics and Complexity*. Wiley-VCH, June 2009.
- [4] Centers for Disease Control and Prevention. FluView, a weekly influenza surveillance report, 2009.
- [5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, September 2009.
- [6] E. S. Cochran. The quake-catcher network, 2010.
- [7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.
- [8] O. Diekmann and J. A. P. Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley, 2000.
- [9] P. Dodds and C. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, July 2009.
- [10] A. Doucet, J. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. New York: SpringerVerlag, 2001.
- [11] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsrithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437:209–214, 2005.
- [12] C. for Disease Control and Prevention. Biosense.
- [13] 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*, 457:1012–1014, 2009.
- [14] M. Halloran, N. Ferguson, S. Eubank, I. Longini, D. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. Germann, D. Wagener, R. Beckman, K. Kadau, C. Barrett, C. Macken, D. Burke, and P. Cooley. Modeling targeted layered containment of an influenza pandemic in the united states. *Proc Natl Acad Sci USA*, 105:4639–4644, 2008.
- [15] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proc Natl Acad Sci USA*, 101(42):15124–15129, 2004.
- [16] F. Jordans. WHO working on formulas to model swine flu spread, 2009.
- [17] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science*, 300(5627):1966–1970, 2003.
- [18] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, 2009.
- [19] I. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. Cummings, and M. Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.
- [20] E. Morozov. Swine flu: Twitter’s power to misinform. Foreign Policy magazine website post.
- [21] P. Watson, J. Geraci, D. Greenblatt, A. Chadha, and C. McMahon. Realtime disease detection for your city from messages on twitter, 2009.
- [22] H. J. Wearing, P. Rohani, and M. J. Keeling. Appropriate models for the management of infectious diseases. *PLoS Med*, 2(7):e174, 07 2005.