Social Media based Simulation Models for Understanding Disease Dynamics

Ting Hua[1], **Chandan K Reddy**[1], **Lei Zhang**[1], **Lijing Wang**[1],
Liang Zhao[2], **Chang-Tien Lu**[1], **Naren Ramakrishnan**[1]
[1] Virginia Tech, [2] George Mason University
tingh88@vt.edu

Abstract

In this modern era, infectious diseases, such as H1N1, SARS, and Ebola, are spreading much faster than any time in history. Efficient approaches are therefore desired to monitor and track the diffusion of these deadly epidemics. Traditional computational epidemiology models are able to capture the disease spreading trends through contact network, however, one unable to provide timely updates via real-world data. In contrast, techniques focusing on emerging social media platforms can collect and monitor real-time disease data, but do not provide an understanding of the underlying dynamics of ailment propagation. To achieve efficient and accurate real-time disease prediction, the framework proposed in this paper combines the strength of social media mining and computational epidemiology. Specifically, individual health status is first learned from user's online posts through Bayesian inference, disease parameters are then extracted for the computational models at population-level, and the outputs of computational epidemiology model are inversely fed into social media data based models for further performance improvement. In various experiments, our proposed model outperforms current disease forecasting approaches with better accuracy and more stability.

1 Introduction

The seasonal flu kills 290 to 650 thousand people every year, according to the Centers for Disease Prevention and Control (CDC) and the World Health Organization (WHO) [1]. Flu is not only "deadly" but also "expensive". For example, in the United States, it causes significant economic loss up to $87 billion annually. Further more, because of modern transportation, these diseases can spread much faster and hit larger population. In March 2009, swine flu first occurred in Mexico and California, and soon reached all parts of the world as a result of airline travel [Girard *et al.*, 2010]. How to efficiently monitor and track the dynamics of ongoing epidemic diseases is one of the most crucial challenges in the field of public health. Currently, two related research branches have been working on this challenge, namely, social media mining and computational epidemiology.

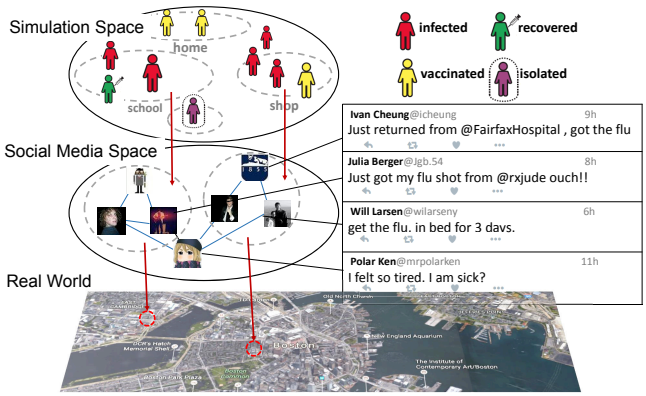

Figure 1: SMS model consists of "social media space" and "simulation space". Both of them can be considered as subsets of the real world.

Computational epidemiology models usually utilize social contact network to simulate the flu spreading process. Specifically, each person in such a system will be assigned geographical, social, behavioral, and demographic attributes (e.g., age and income) [Bisset *et al.*, 2009]. And the social contact network is simulated through assigning daily activities and locations for each node (person) in the network [Bisset *et al.*, 2009; Barrett *et al.*, 2009]. The epidemic dynamics are then modeled as diffusion processes across the network, which enables the computation of infectious time and location for all individuals. However, they are highly dependent on surveillance data provided by the Centers for Disease Control and Prevention (CDC) to estimate parameters, which results in following two limitations.[2] 1) *Low effectiveness*. CDC surveillance data is updated once per week, with at least one week delay in real-time disease transmission. Such outdated data can hardly achieve good performance in monitoring the rapidly spreading epidemics. 2) *Insufficient accuracy*. CDC provides surveillance data at the state-level, with not much detailed information for subregions such as counties. The granularity of these data is too fine to tune accurate pa-

[1] http://www.who.int/mediacentre/factsheets/fs211/en/

rameters for model estimation.

On the other hand, social media users may report their symptoms through online posts, which are known to be the best signals for early disease detection, even before diagnoses [Krieck *et al.*, 2011]. Several attempts have been made to track disease outbreaks through studying the relationship between the aggregate volume of flu-related social media posts and CDC data [Achrekar *et al.*, 2012; Hirose and Wang, 2012; Culotta, 2010]. They usually first identify flu-related tweets by keyword selection and then try different regression models to correlate the tweet volume and CDC statistics [Hirose and Wang, 2012; Culotta, 2010]. However, most social media mining techniques are purely data-driven methods, and do not have a clear understanding of the underlying social contact network in the disease diffusion. As later demonstrated in Section 4, social media mining methods are "short-sighted" in nature. They are good at real-time detection and short-term prediction, since they can utilize the most up-to-date social media data. However, they perform poorly in long-term disease forecasting, because they ignore the inherent features of the disease and therefore fail to model their spreading process.

As discussed above, computational epidemiology models can capture the diffusion patterns of disease spread through detailed simulation of the real world, but their "intelligence" has not been fully developed due to the limitations of CDC data in effectiveness and accuracy. In contrast, social media mining methods can utilize the most updated user-provided data, but lack the global knowledge in disease modeling. Zhao et al. [Zhao *et al.*, 2015] proposed a hybrid solution by considering the two factors through one optimization goal. In this paper, we further propose a novel **S**ocial **M**edia based **S**imulation (SMS) model, a framework consist of both graphical model using text mining and computational simulation system.

Specifically, as shown in Figure 1, the proposed SMS model considers online posts from users in social media space, as well as underlying social contact network in computational simulation space. In the social media space, SMS model infers users' health status through their posts. First, SMS model is able to identify infected users through tweets such as "4th day with flu". Second, this model is also capable of identifying potential patients in their incubation period through tweet such as "I felt so tired. I am sick?" These individual posts are then analyzed and aggregated into population-level parameters for simulation space. Based on the detailed social contact network, the disease propagation process is optimized in the simulation. After that, the outputs of computational part are fed into the social media space as the prior knowledge for learning in the next iteration. Such iterative feedback mechanism benefits the learning for both the spaces, and therefore perfectly tackles the challenges that previous social media mining methods and computational epidemiology models can not handle. The major contributions of this paper are summarized as follows.

- **A unified framework that jointly models social media mining and epidemiology simulation is proposed**. The proposed SMS model will collect and analyze the most updated data from social media, and at the same time, is capable of inferring the underlying propagation process like a standard computational model.
- **A "dual-space" learning model is developed for mining the disease diffusion patterns.** Our SMS model consists of two spaces: social media space and simulation space. Different methodology is adopted in different spaces for optimal performance. Meanwhile, information is efficiently shared across the spaces with carefully-designed learning strategies.
- **A novel learning algorithm consisting of multiple inference technologies is developed**. A variety of learning approaches are incorporated into the SMS model, including Gibbs sampling, maximum likelihood estimation, and numerical optimization.
- **Extensive experiments were performed to demonstrate the effectiveness of the proposed the SMS model**. The SMS model is tested on large-scale datasets and is compared with four existing state-of-the-art algorithms. With extensive quantitative and qualitative experimental results, the SMS model shows significant improvement over both social media mining methods and computational epidemiology models.

[2]http://www.cdc.gov/flu/weekly/fluviewinteractive.htm

2 The Proposed SMS Model

The overall goal of this paper can be formally defined as: using the social media data streams $\mathcal{U}$ as inputs, estimate the health states $S_{\mathcal{V},t}$ at each time stamp t for the population $\mathcal{V}$ in the targeted region. To achieve this goal, our proposed SMS model integrates two spaces (**simulation space** and **social media space**) within one framework, as shown in Figure 2. In this section, we first introduce the independent learning process within each space, and then present the information sharing mechanism between the two spaces.

2.1 Learning in Social Media Space

Social media data is defined by $\mathcal{D} = \bigcup_{u \in \mathcal{U}, t \in \mathcal{T}} D_{u,t}$, where $D_{u,t}$ is the post of user u at time t. Note that, multiple posts of user u within time interval t are integrated as one document. In the well-known SEIR model, each person is assumed to be in one of the following states: susceptible (S), exposed (E), infectious (I), and recovered (R). Generally speaking, the individual shows no symptoms in the susceptible (S) and recovered (R) status, will be infected but not yet infectious in the exposed (E) status, and suffers from severe symptoms in the infectious (I) status. Social media users will not post content related to disease in status S and R, since no symptoms are shown in these two status. Therefore, our work assumes a user can be in one of following three health states: healthy (S and R of SEIR), exposed (E), and infectious (I).

As shown in Figure 2 and Algorithm 1, SMS model learns health status from social media data through a specially designed Bayesian graphical model. The generative process of words in our model for social media posts consists of three stages. First, the health status s is chosen from per-document multinomial distribution with prior μ: $s = 0$ indicates the

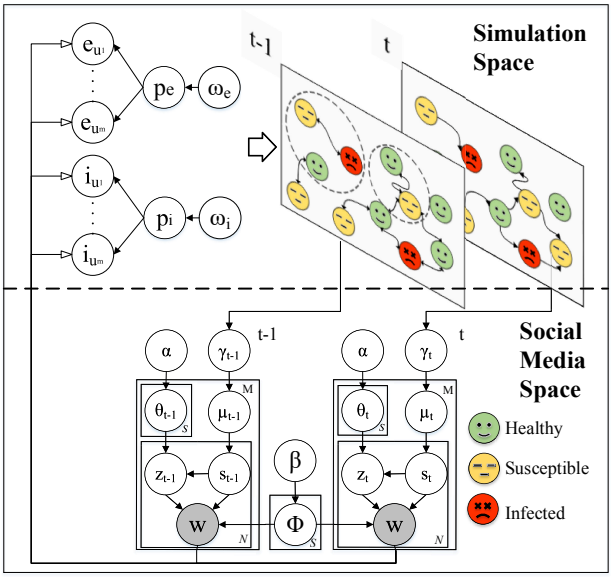

Figure 2: Overall Framework of the SMS model.

user of the corresponding post is healthy (S and R of SEIR); $s = 1$ implies the user is exposed but has not been confirmed as infected (e.g., "I feel so tired all day even with 9 hours sleep"); and $s = 2$ denotes the user has been infected, with words such as "get the flu, in bed." Second, after choosing the value for s, topic z is drawn from K-dimensional topic mixture θ_s. Different from other topic models, each document here is associated with S topic distributions. This scheme enables the prediction of the health status based on the extracted topics. Finally, a word is generated from word distribution $\phi_{s,z}$, conditioned on both topic z and health status s.

ALGORITHM 1: Generation process of words in social media space of SMS model.

for *each label* $s = 1, 2, ..., S$ **do**
  for *each topic* $z = 1, 2, ..., K$ **do**
    Draw $\phi_{s,z} \sim Dir(\beta)$;
  end
end
for *each time stamp* $t = 1, 2, ..., T$ **do**
  for *each document* $D_{u,t} = 1, 2, ..., U$ **do**
    Draw $\mu_{u,t} \sim Dir(\gamma)$;
    for *each label* $s = 1, 2, ..., S$ **do**
      Draw $\theta_{u,t,s} \sim Dir(\alpha)$;
    end
    for *each word* w *in document* $D_{u,t}$ **do**
      Draw $s \sim Multi(\mu_{u,t})$;
      Draw $z \sim Multi(\theta_{u,t,s})$;
      Draw $w \sim Multi(\phi_{s,z})$;
    end
  end
end

Besides, a multinomial variable $S_{u,t} = (h_{u,t}, e_{u,t}, i_{u,t})$ is defined for the health status of each user u at time t in social media space. In this vector, only one element equals to 1, and other two elements equal to 0. Specifically, $h_{u,t} = 1$ indicates the user u is healthy, $e_{u,t} = 1$ denotes the user u is exposed to the disease, and $i_{u,t} = 1$ means the user u became infectious. $S_{u,t}$ can be viewed as a "summary" of variable s that: variable s indicates the status for each word, while $S_{u,t}$ indicates the health status for each user. Therefore, the values of elements in $S_{u,t}$ can be computed through posterior distribution of variable s: the s-th ($s = 0, 1, 2$) element in variable $S_{u,t}$ is 1, when the maximal element in posterior distribution of s is the one with index s.

2.2 Learning in Simulation Space

Simulation space is a contact network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ is the targeted population, $\mathcal{E}$ is the edge set, and $\mathcal{W}$ are weights for edges. Specifically, node $v_1 \in \mathcal{V}$ in the network denotes an individual, who has a contact with another individual v_2 through edge $(v_1, v_2) \in \mathcal{E}$, with contact duration equal to $w(v_1, v_2)$. Under the contact network $\mathcal{G}$, person v_2 can be infected by person v_1 with probability $p(w(v_1, v_2), \tau)$, where τ is the transmission probability per contact time unit. Similar to health status of social media users, we assume each person v in the simulation world is associated with three status: healthy (S and R), exposed (E), and infectious (I). Incubation period $p_E(v)$ and infectious period $p_I(v)$ denote the duration of exposed status and infectious status for person v, respectively.

To minimize the inconsistency of social media space and simulation space, the hidden health states calculated by the simulation should be consistent with those obtained from social media. Although it is impossible to map each person v in the simulation space to a specific user u in the social media space, linking the two spaces at the population level is practical and sufficient for our task. Specifically, we compare the social media users with simulated people within the same region (e.g., counties or states), which can be formalized by the following loss function:

$$\begin{aligned}\mathcal{L} = \min_{\tau} \sum_{t=1}^{T} || \sum_{v=1}^{V} I_{v,t}(\mathcal{G}, p_E, p_I, \tau) - \sum_{u=1}^{U} I_{u,t}||^2 \\ + \sum_{t=1}^{T} || \sum_{v=1}^{V} E_{v,t}(\mathcal{G}, p_E, p_I, \tau) - \sum_{u=1}^{U} E_{u,t}||^2.\end{aligned} \tag{1}$$

$I_{v,t}(\mathcal{G}, p_E, p_I, \tau)$ is the overall infectious state of simulation results at time t, and $E_{v,t}(\mathcal{G}, p_E, p_I, \tau)$ is the corresponding incubation state. Here, the transmission probability τ is the parameter needed to be optimized to achieve the best performance.

2.3 Interaction between two spaces

The key to the information transferring from social media space to simulation space is to find a way to aggregate individual-level social media posteriors into population-level parameters. In Equation (1), p_E and p_I are input parameters required by the simulation space. The specific incubation period $p_E(v)$ and infectious period $p_I(v)$ for each individual v can be viewed as observations from multinational distributions $multi(p_E)$ and $multi(p_I)$. As mentioned above,

although it is unrealistic to link each user u in social media space to each individual v in simulation space, the estimation based on population-level is sufficient for our task. The maximum likelihood solutions for p_E is thus calculated as the expectation of social media users' incubation period $n_E^t/|\mathcal{U}|$, where n_E^t denotes the number of users whose incubation period is equal to t days. The estimation of parameter p_I can be calculated in a similar manner.

Conversely, the simulation outputs can also be used to improve the learning performance in social media space. On one hand, in social media space, the ideal values for Dirichlet prior γ of healthy status s should reflect the health status of the population. On the other hand, the simulation outputs include health status of the population. Specifically, two transition parameters, the incubation rate $\rho_{t,e}$ and the infectious rate $\rho_{t,i}$ are defined to denote the ratio of exposed and infectious persons among the entire population, respectively. These values are calculated as shown in Equations (2) and (3):

$$\rho_{t,e} = \sum_{v=1}^{\mathcal{V}} E_{v,t}(\mathcal{G}, p_E, p_I, \tau)/\mathcal{V}, \tag{2}$$

$$\rho_{t,i} = \sum_{v=1}^{\mathcal{V}} I_{v,t}(\mathcal{G}, p_E, p_I, \tau)/\mathcal{V}, \tag{3}$$

where $E_{v,t}(\mathcal{G}, p_E, p_I, \tau)$ and $I_{v,t}(\mathcal{G}, p_E, p_I, \tau)$ are outputs from simulation space, as mentioned in Equation (1). Gamma prior for a Dirichlet parameter of healthy status s (s can be e or i) at epoch t is therefore computed as follows:

$$\gamma_{t,s} \sim Gamma(\sigma\rho_{t,s}, \sigma), \tag{4}$$

where the mean is proportional to the simulation output parameter $\sigma\rho_{t,s}$, while parameter σ controls the consistency of the prior.

3 Model Inference

Although exact inference of posterior distributions for hidden variables in the SMS model is generally intractable, the solution can be estimated through approximate inference algorithms, such as variational expectation [Blei *et al.*, 2003; Hoffman *et al.*, 2013; 2010], Gibbs sampling [Griffiths and Steyvers, 2004; Porteous *et al.*, 2008; Casella and George, 1992], maximum likelihood estimation [Christopher, 2007; Bock and Aitkin, 1981], and numerical optimization [Qin *et al.*, 2009; Wright and Nocedal, 1999]. First, Gibbs sampling is used for the inference of the proposed text mining model in social media space, as this approach can yield more accurate estimation compared to variational inference in LDA-like graphical model. Second, maximum likelihood estimation (MLE) is adapted to estimate the incubation period p_E and infectious period p_I. And the operations in the simulation space are optimized through Nelder-Mead method [Lagarias *et al.*, 1998; Nelder and Mead, 1965].

Using Algorithm 1 and the graphical model in Figure 2, the joint distribution of SMS model in social media space can be represented as Equation (5):

$$\begin{aligned} &P(\mathbf{w}, \mathbf{z}, \mathbf{s}|\alpha, \gamma, \beta) \\ &= \prod_{m=1}^{M}\prod_{n=1}^{N} p(w_{mn}|s_{mn}, z_{mn}) \\ &\prod_{m=1}^{M}\prod_{n=1}^{N} p(z_{mn}|\theta_m^{s_{mn}}) \prod_{m=1}^{M}\prod_{n=1}^{N} p(s_{mn}|\mu_m) \\ &\prod_{m=1}^{M} p(\mu_m|\gamma) \prod_{s=1}^{S}\prod_{m=1}^{M} p(\theta_m^{s_{mn}}|\alpha)p(\gamma|\varphi, \sigma). \end{aligned} \tag{5}$$

The key to this inferential problem is to estimate the posterior distributions of the following hidden variables: (1) topic assignment indicator z_{mn} for words; (2) label assignment indicator s_{mn} for words; (3) topic mixture proportion θ_{msz} and label mixture proportion μ_{ms}. The last term $p(\gamma|\varphi, \sigma)$ of Equation (5) is as follows:

$$p(\gamma|\varphi, \sigma) = \prod_s \frac{\sigma^{\sigma\varphi_s}\gamma_s^{\sigma\varphi_s - 1}\exp(-\sigma\gamma_s)}{\Gamma(\sigma\varphi_s)}, \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function. From the joint distribution, the full conditional distribution for a word term $i = (m, n)$ can be derived, where i denotes word n in document m. As a special case of Markov chain Monte Carlo, Gibbs sampling iteratively samples one instance at a time, conditional on the values of the remaining variables.

$$p(z_{mn} = k|\mathbf{w}, \mathbf{z}_{\neg i}, \mathbf{s}) = \frac{n^v_{sz_{\neg i}} + \beta}{\sum_{v=1}^{V}(n^v_{sz_{\neg i}} + \beta)} \frac{n^{z_{\neg i}}_{ms} + \alpha}{\sum_{z=1}^{K}(n^{z_{\neg i}}_{ms} + \alpha)} \tag{7}$$

In the above equation, V is the size of the vocabulary, K is the number of topics, $n^v_{sz_{\neg i}}$ is the number of topic z and label s assigned to term v in the scope of the entire data set, without current instance i and its topic assignment. $n^{z_{\neg i}}_{ms}$ is the number of words selecting label s and topic z in document m except current instance i

$$p(s_{mn} = s|\mathbf{w}, \mathbf{z}, \mathbf{s}_{\neg i}) \propto \frac{n^v_{s_{\neg i}z} + \beta}{\sum_{v=1}^{V}(n^v_{s_{\neg i}z} + \beta)} \frac{n^z_{ms_{\neg i}} + \alpha}{\sum_{z=1}^{K}(n^z_{ms_{\neg i}} + \alpha)} (n_m^{s_{\neg i}} + \gamma). \tag{8}$$

Similar to the inference of z, $n^v_{s_{\neg i}z}$ is the number of topic z and label s assigned to term v in the scope of the entire data set, without current instance i and its label assignment, $n^z_{ms_{\neg i}}$ is the number of words choosing label s and topic z in document m except current instance i, and $n_m^{s_{\neg i}}$ is the number of words (remove instance i) choosing label s in document m.

Parameters Φ_{szv}, θ_{msz}, and μ_{ms} are multinomial distributions with Dirichlet priors, and can be easily computed according to Bayes rule and the definition of Dirichlet prior.

The optimal values of transmission rate τ are searched using Nelder-Mead optimization method, since solving for τ with respect to loss function $\mathcal{L}$ in Equation (1) is a non-convex and non-differentiable problem.

SMS model is based on semi-supervised learning. In the training process, SMS model is fed with labeled tweets (health states). The trained model $\mathcal{M}$ of text part in social media space of SMS contains the distribution of words ϕ_s for

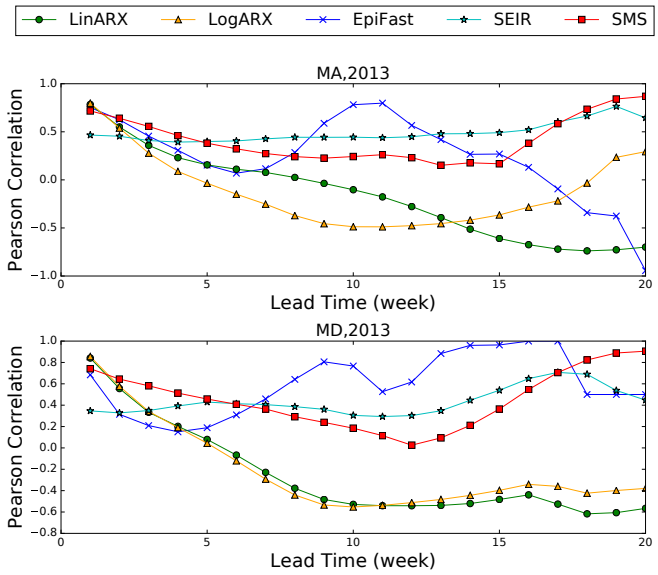

Figure 3: Performance comparison in terms of Pearson correlation for MA and MD, 2013.

health state s. With trained model $\mathcal{M}$, SMS model can be used to estimate the posterior distributions of health state $\tilde{s}$ of unlabeled Twitter streams. In order to achieve this goal, we follow the approach introduced in [Steyvers *et al.*, 2004] to run the inference process on the new documents exclusively. Inference for this testing process are given to Equations (7) and (8) with the difference that: the current Gibbs sampler is run with ϕ_s fixed. In the initial stage, the algorithm randomly assigns switch variables to words. Then a number of Gibbs sampling updates are made to estimate the posterior.

$$p(\tilde{s} = s|\tilde{w} = v, \tilde{s}_{\neg i}, \tilde{z}, \mathcal{M}) \propto \phi_{s,v}(n^{s}_{m,\neg i} + \gamma) \qquad (9)$$

4 Experimental Results

In this section, we first describe the data preparation, the metrics used for evaluation, and the settings for all the comparison methods. After that, our proposed SMS model is compared with existing state-of-the-art algorithms on real-world data sets.

4.1 Datasets

Twitter data used in this paper consists of two parts: training set $\mathcal{D}_1$ and testing set $\mathcal{D}_2$. The training set $\mathcal{D}_1$ was collected using the following steps: 1) **Twitter stream data collection**. Twitter data streams were retrieved through REST API using flu related keywords, such as "flu", "h1n1", and "influenza". The keyword lists are provided by Paul and Dredze [Paul and Dredze, 2012]. 2) **Identify tweets health status**. We asked human annotators to create the labels for the tweets. Each annotator selected a label from status "healthy", "exposed", and "infected" for each tweet. A label was confirmed only if it was chosen by at least 2 annotators.

The testing data set $\mathcal{D}_2$ was created as follows, which shares the same users $\mathcal{U}$ with $\mathcal{D}_1$. 1) **Extract users**. Users $\mathcal{U}$ of tweets in the training set $\mathcal{D}_1$ were extracted from data streams. 2) **Retrieve tweets**. Retrieve posts belonging to authors $\mathcal{U}$, which were published two weeks before and after the time span of dataset $\mathcal{D}_1$. 3) **Geocoding**. Conduct geocoding on tweets to identify location information such as GPS tags using Carmen geocoder[1]. 4) **Data clean**. Remove retweets and only keep tweets within the targeted regions. We collected data in Maryland (MD) and Massachusetts (MA) from August 2012 to July 2014. 70% of the tweets were assigned with locations. It should be noted that $\mathcal{D}_1$ and $\mathcal{D}_2$ share the same set of users.

4.2 Labels and Evaluation Metrics

In this paper, the ground truth influenza data used for validation is provided by the Centers for Disease Control and Prevention (CDC), which contains the percentage of weekly physician visits related to influenza-like illness (ILI) for most regions in the United States.

In this paper, three different widely used metrics for evaluating the prediction performance are adopted: Pearson correlation, mean squared error (MSE), and peak-time error. **Pearson correlation** is the covariance of predicted results and the ground truth divided by their deviation product. It measures the linear relationship between variables, with values varying between +1 and -1. The larger Pearson correlation value implies the stronger positive linear correlation between two variables. **Mean squared error (MSE)** is the mean of squared error between the predicted results to the ground-truth class labels. **Peak-time error** is the difference between predicted peak time (the week with largest infected population) and the actual peak time. A smaller peak-time error indicates better forecasting performance.

4.3 Comparison Methods

The proposed SMS model is compared with four other models, including 2 social media mining methods (LinARX and LogARX) and 2 computational epidemiology models (SEIR and EpiFast).

LinARX [Achrekar *et al.*, 2011] uses standard autoregressive exogenous model to explore the dependence between influenza-like illness (ILI) visits and social media data time series. The orders of LinARX for the Twitter data time series and CDC time series are set as 2 and 3 based on cross-validation.

LogARX [Achrekar *et al.*, 2012] evolved from LinARX, where an additional logit function transformation is introduced in order to enforce 0-1 classification boundary for ILI visit percentage. The orders of LogARX for both time series (CDC and social media time series) are set as 2 based on cross-validation.

SEIR [Murray, 2002] models epidemic dynamics into four health states: susceptible (S), exposed (E), infectious (I), and recovered (R). The volume of the positive tweets classified was fed into above mentioned LinARX model. The orders of the LinARX model for both time series (Twitter data and CDC data) were set as 2 based on cross-validation.

[1]https://github.com/mapbox/carmen

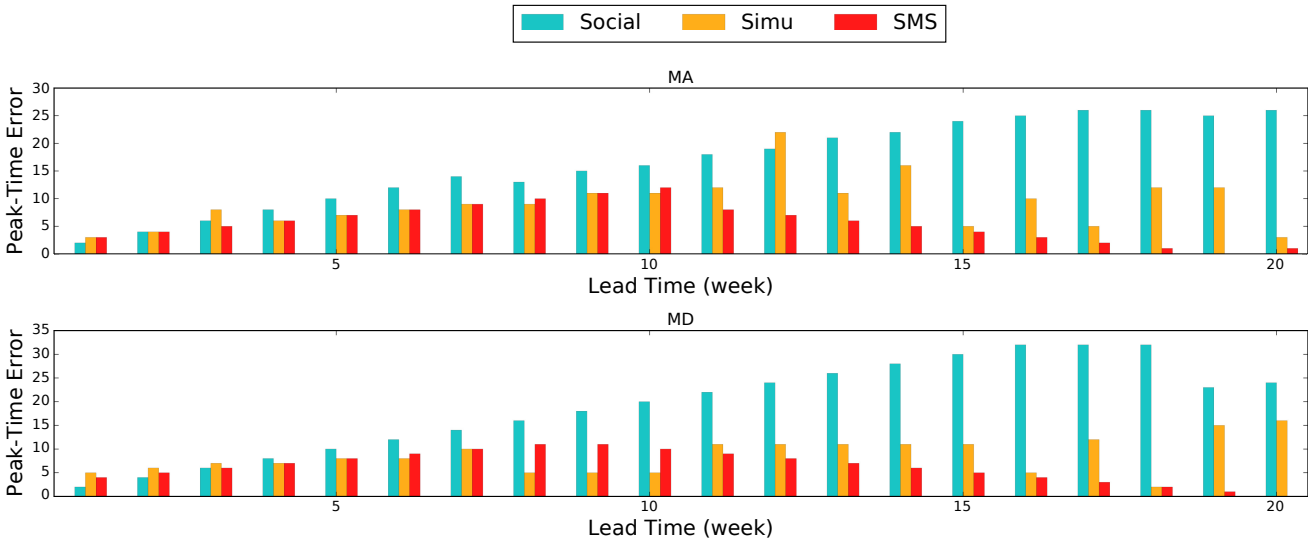

Figure 4: Performance in terms of peak time in MA and MD states for 2013 data.

EpiFast [Beckman *et al.*, 2014] simulates disease propagation in a social contact network. Nelder Mead method [Beckman *et al.*, 2014] is adopted to minimize the error between predicted results and actual ILI visit percentage.

4.4 Results

In this section, models were compared by the percentage of ILI visits, with lead times varied from 1 week to 20 weeks. The results are validated in terms of three evaluation metrics introduced above for two states (MA and MD). Due to page limitation, only results of year 2013 are being reported here, and similar patterns can be seen in other years.

Performance on Pearson correlation.

The forecasting performance in terms of Pearson correlation in Massachusetts (MA) and Maryland (MD) is reported in Figure 3. In general, SMS model yields the best overall performance in terms of Pearson correlation, methods based on social media mining (LinARX and LogARX) can achieve better performance than the popular computational epidemiology methods (SEIR and EpiFast) for shorter periods, but computational models show their advantage with larger lead times.

As shown in Figure 3, the Pearson correlations of social media mining methods are high when the lead time is small, for example, less than 2 weeks. However, their performance decreases quickly with the increase in lead time. On the contrary, although the computational epidemiology methods perform worse than social media mining techniques at shorter lead times, they become more stable as the lead time increases. Our SMS model has a comparable initial performance with social media mining methods, and outperform them significantly with a large margin when the lead-time is over 10 weeks.

These results based on Pearson correlation confirm that our proposed SMS model is the best performer over all other methods, social media methods are good at predicting the near future, and the computational models are better for long-term forecasting. These phenomena are inevitable, driven by the underlying characteristics of different methods. Social media mining methods highly rely on real-time data. Such dependence leads to their good performance of predicting outcomes in the near future, but results in their inability to achieve long-term stability. Computational epidemiology methods, on the other hand, use CDC data which inherently was 1-2 weeks of time lag. Thus they are less sensitive to the current data and perform worse than social media approaches in forecasting the near future. SMS model benefits from utilizing real-time data and combining it with long-term progression mechanism, and therefore achieves the best overall performance.

	2013 MA	2013 MD
LinARX	6.65E-04 ± 3.74E-04	8.19E-04 ± 4.01E-04
LogARX	5.51E-04 ± 2.39E-04	5.00E-04 ± 1.79E-04
EpiFast	2.24E-03 ± 9.24E-04	5.14E-03 ± 5.57E-03
SEIR	3.73E-04 ± 5.38E-05	4.61E-04 ± 1.53E-04
SMS	**2.38E-04 ± 6.16E-05**	**2.63E-04 ± 7.51E-05**

Table 1: Performance in terms of mean square error in MA and MD states for 2013 data. The best performers are marked in bold, the corresponding second best performers are marked with underlines.

Performance on MSE and peak-time error.

Table 1 shows the mean squared error (MSE) results for five methods. Each term is the mean value of the MSEs with different lead times plus/minus corresponding standard deviation. SMS model is the best performer, with smallest average mean and the least variance in both cases. Generally, computational epidemiology models are better than social media mining methods. First, computational epidemiology method, SEIR, is the second best performer. Second, computational epidemiology models are more stable. Their standard deviations are much smaller than their mean values, while social media mining methods' deviations have the same orders of magnitude with their corresponding mean values. This is

because computational epidemiology approaches can model long-time disease spreading patterns across contact network, and obtain more robust performance than social media mining methods.

Figure 4 displays the performance of peak-time errors, i.e., the difference between predicted and actual peak time. Peak-time prediction is decided by larger volume data points rather than isolated moments, which requires significant prior knowledge compared to other measurements. Here we made two variations: "**Simu**" is the method that only comprises of computational space, and "**Social**" is the variation that only includes social media mining component. As can be seen from the figure, "Social" method usually starts as the best performer (at lead-time of 1 or 2 weeks), but its errors continue growing with increase in lead-time. Computation-only variation "Simu" generally produces better results at larger lead-times, but are less stable compared to SMS. For example, for the performance in MA, errors of "Simu" method grow dramatically from lead-time of 11 to 12 weeks, and quickly drop again at lead-time 13 weeks. The hybrid version, SMS model, is more stable, which always "smartly" chooses the pattern that can yield better results.

5 Conclusion

This paper provides a novel framework for forecasting disease spread on large-scale social contact network. On one hand, similar to social media mining models, the proposed SMS model can analyze the semantic meaning of the social media data and infer users' health status through Bayesian inference model. On the other hand, similar to computational methods, our SMS model can aggregate the individual results into population-level parameters required for simulation. Our extensive experimental results show that the SMS model has obvious advantages over both computational epidemiology models and social media mining methods. For short-term forecasting, SMS model can achieve the best performance using most up-to-date health information from social media data. Also, SMS model can maintain a good long-term (more than 10 weeks) prediction performance, as well as other computational methods, through its powerful simulation component.

6 Acknowledgements

This work is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, by the National Science Foundation via grants DGE-1545362, IIS-1633363, IIS-1619028, IIS-1707498, IIS-1646881, and by the Army Research Laboratory under grant W911NF-17-1-0021. The US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, NSF, Army Research Laboratory, or the U.S. Government.

References

[Achrekar *et al.*, 2011] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707. IEEE, 2011.

[Achrekar *et al.*, 2012] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Online social networks flu trend tracker: a novel sensory approach to predict flu trends. In *Proceedings of the 5th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, pages 353–368. Springer, 2012.

[Barrett *et al.*, 2009] Christopher L Barrett, Richard J Beckman, Maleq Khan, VS Anil Kumar, Madhav V Marathe, Paula E Stretz, Tridib Dutta, and Bryan Lewis. Generation and analysis of large synthetic social contact networks. In *Proceedings of the 41st Winter Simulation Conference (WSC)*, pages 1003–1014. Winter Simulation Conference, 2009.

[Beckman *et al.*, 2014] Richard Beckman, Keith R Bisset, Jiangzhuo Chen, Bryan Lewis, Madhav Marathe, and Paula Stretz. Isis: A networked-epidemiology based pervasive web app for infectious disease pandemic planning and response. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1847–1856. ACM, 2014.

[Bisset *et al.*, 2009] Keith R Bisset, Jiangzhuo Chen, Xizhou Feng, VS Kumar, and Madhav V Marathe. Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd international conference on Supercomputing (ICS)*, pages 430–439. ACM, 2009.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. In *The Journal of Machine Learning Research*, volume 3, pages 993–1022. MIT Press, 2003.

[Bock and Aitkin, 1981] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. volume 46, pages 443–459. Springer, 1981.

[Casella and George, 1992] George Casella and Edward I George. Explaining the gibbs sampler. volume 46, pages 167–174. Taylor & Francis, 1992.

[Christopher, 2007] Bishop Christopher. Pattern recognition and machine learning. pages 93–94. Springer, New York, 2007.

[Culotta, 2010] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics (SOMA)*, pages 115–122. ACM, 2010.

[Girard *et al.*, 2010] Marc P Girard, John S Tam, Olga M Assossou, and Marie Paule Kieny. The 2009 a (h1n1) influenza virus pandemic: A review. volume 28, pages 4895–4902. Elsevier, 2010.

[Griffiths and Steyvers, 2004] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. volume 101, pages 5228–5235. National Acad Sciences, 2004.

[Hirose and Wang, 2012] Hideo Hirose and Liangliang Wang. Prediction of infectious disease spread using twitter: A case of influenza. In *Proceedings of the 55th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, pages 100–105. IEEE, 2012.

[Hoffman *et al.*, 2010] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems(NIPS)*, pages 856–864, 2010.

[Hoffman *et al.*, 2013] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. volume 14, pages 1303–1347, 2013.

[Krieck *et al.*, 2011] Manuela Krieck, Johannes Dreesman, Lubomir Otrusina, and Kerstin Denecke. A new age of public health: Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health Web-Science Workshop, ACM Web Science Conference*, 2011.

[Lagarias *et al.*, 1998] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. volume 9, pages 112–147. SIAM, 1998.

[Murray, 2002] James D Murray. Mathematical biology i: an introduction. In *interdisciplinary applied mathematics*, volume 17. Springer, 2002.

[Nelder and Mead, 1965] John A Nelder and Roger Mead. A simplex method for function minimization. volume 7, pages 308–313. Oxford University Press, 1965.

[Paul and Dredze, 2012] Michael J Paul and Mark Dredze. A model for mining public health topics from twitter. volume 11, pages 16–6, 2012.

[Porteous *et al.*, 2008] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM, 2008.

[Qin *et al.*, 2009] A Kai Qin, Vicky Ling Huang, and Ponnuthurai N Suganthan. Differential evolution algorithm with strategy adaptation for global numerical optimization. volume 13, pages 398–417. IEEE, 2009.

[Steyvers *et al.*, 2004] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[Wright and Nocedal, 1999] Stephen J Wright and Jorge Nocedal. Numerical optimization. volume 35. Springer Science, 1999.

[Zhao *et al.*, 2015] Liang Zhao, Jiangzhuo Chen, Feng Chen, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 639–648. IEEE, 2015.