Geoinformatica
https://doi.org/10.1007/s10707-019-00376-9

Online flu epidemiological deep modeling on disease contact network

Liang Zhao[1] · Jiangzhuo Chen[2] · Feng Chen[3] · Fang Jin[4] · Wei Wang[5] · Chang-Tien Lu[6] · Naren Ramakrishnan[6]

Received: 30 August 2018 / Revised: 13 April 2019 / Accepted: 3 July 2019 /
Published online: 25 July 2019

Abstract

The surveillance and preventions of infectious disease epidemics such as influenza and Ebola are important and challenging issues. It is therefore crucial to characterize the disease progress and epidemics process efficiently and accurately. Computational epidemiology can model the progression of the disease and its underlying contact network, but as yet lacks the ability to process of real-time and fine-grained surveillance data. Social media, on the other hand, provides timely and detailed disease surveillance but is insensible to the underlying contact network and disease model. To address these challenges simultaneously, this paper proposes a novel semi-supervised neural network framework that integrates the strengths of computational epidemiology and social media mining techniques for influenza epidemiological modeling. Specifically, this framework learns social media users' health states and intervention actions in real time, regularized by the underlying disease model and contact network. The learned knowledge from social media can then be fed into the computational epidemic model to improve the efficiency and accuracy of disease diffusion modeling. We propose an online optimization algorithm that iteratively processes the above interactive learning process. The extensive experimental results provided demonstrated that our approach can not only outperform competing methods by a substantial margin in forecasting disease outbreaks, but also characterize the individual-level disease progress and diffusion effectively and efficiently.

Keywords Social media · Neural network · Computational epidemiology · Stochastic gradient descent

✉ Liang Zhao
lzhao9@gmu.edu

[1] Department of Information Science and Technology, George Mason University, Fairfax, VA 22030, USA

[2] Biocomplexity Institute of Virginia Tech, Arlington, VA 22203, USA

[3] Department of Computer Science, University at Albany, Albany, NY 12222, USA

[4] Department of Computer Science, Texas Tech University, Lubbock, TX 79409, USA

[5] Microsoft, Redmond, WA 98052, USA

[6] Department of Computer Science, Virginia Tech, Arlington, VA 22203, USA

1 Introduction

Epidemics such as Ebola and seasonal influenza pose a serious threat to global public health. The recent Ebola outbreak in West Africa led to 27,055 cases and 11,142 deaths [32]; Seasonal influenza is estimated to result in from 3 to 5 million cases of severe illness and about 250,000 up to 500,000 deaths each year [33]. These diseases share two important characteristics: (1) frequent local and global travels often facilitate the spread of epidemic at a large spatial scale, through close contacts between people; and (2) they spread rapidly. For example, during the 2009 H1N1 pandemic the initial case occurred in Mexico in March 2009, but by the beginning of November 2009, more than 6,000 people had died from H1N1 influenza [28]. In order to implement effective public health measures to mitigate such fast-developing epidemics, it is crucial to characterize the disease and the evolution of the ongoing epidemic efficiently and accurately. To address this issue, recent research in both computational epidemiology and social media mining have achieved important progress and demonstrated their usefulness in dealing with different aspects of the problem.

In the field of computational epidemiology, individual-based network epidemiological techniques have been developed to study the spatio-temporal dynamics of the spread of epidemics. These simulate disease transmission at the individual level, including a consideration of interventions such as vaccinations, school closures, and quarantine. High-performance simulation systems have been developed that are capable of simulating epidemics using network-based models. Such simulations focus on the evolution of an epidemic, enabling planners to: (i) forecast the spatio-temporal spread of the disease; (ii) estimate important epidemic measures such as the peak time; and (iii) evaluate the effectiveness of intervention strategies.

Currently, computational epidemiology suffers from the following challenges. 1) *The lack of spatially fine-grained surveillance data for model tuning*. Existing work mostly relies on surveillance data such as that provided by the Centers for Disease Control and Prevention (CDC) [12] in the United States to estimate the model parameters. However, the CDC surveillance data only provides state-level spatial information, which is insufficient for accurate diffusion modeling within a state. 2) *Difficulties in tracking the dynamics of contact networks in real time*. Interventions such as school closures and vaccination drives play an important role in mitigating epidemics by changing people's infectivity and vulnerability and altering the contact network structure. As yet, however, these approaches lack effective ways to monitor the impact of ongoing interventions during the current season in real time. 3) *High cost and low timeliness of retraining*. Although existing approaches generally rely on batch training based on the CDC surveillance data, the CDC surveillance data is only updated weekly, with a delay of at least one week, and thus can never catch up with the real time disease spread.

Social media, on the other hand, can capture timely and ubiquitous disease information from social sensors (i.e., social media users) [13]. Social media-based approaches can be classified into two categories: (i) aggregate-level disease surveillance and (ii) detailed health-informatics analysis. The first category assumes that self-reported symptoms from social media users are reliable signals reflecting the aggregate-level trend of a particular outbreak. Among these, some focus on detecting or tracking current influenza outbreaks while others aim to forecast the severity of the outbreak. The second category focuses on detailed modeling of the social media contents as well as their relevance to health informatics, disease geoinformatics, and health behaviors. However, social media mining approaches suffer from three major drawbacks. First, as a crucial determinant of the disease diffusion pattern, real contact networks are basically unobservable. Estimating social contact networks merely based on the location of social media users is neither accurate nor sufficient. Second, they are generally only capable of characterizing the health information of individual social media users, not the whole demographic population. Third, they typically only employ the disease information retrieved from social media and do not incorporate disease model knowledge.

Although computational epidemiology can model the progress of a disease and the underlying disease contact network among individuals, it suffers from a lack of timely and fine-grained surveillance data. Social media mining, on the other hand, provides spatiotemporal surveillance with good timeliness and geographical details, but is unable to observe the underlying contact network and disease progress model. In order to overcome the above-mentioned challenges, we propose a novel online semi-supervised neural network framework that integrates the strengths of individual-based epidemic simulation and social media mining techniques, namely **S**oc**I**al **M**edia **N**ested **E**pidemic **S**imula**T**ion (**SimNest**). SimNest is a novel bispace framework for influenza epidemics modeling and prediction that combines computational epidemiology and social media data using an interactive mapping process, as shown in Fig. 1. Specifically, the health states and interventions actions of social media users are not only identified via their posts by neural network, but also regularized in an unsupervised manner by the disease model in computational epidemiology. The user health states and parameterized disease model learned from social media is then invoked to provide the computational epidemic model with individual-level surveillance and optimized disease model parameters. This interactive learning process between social media and computational epidemiology creates a consistent stage between these two spaces. The main contributions of our study are summarized as follows:

- **A novel integrated framework is proposed for computational epidemiology and social media mining**: Existing approaches from computational epidemiology and social

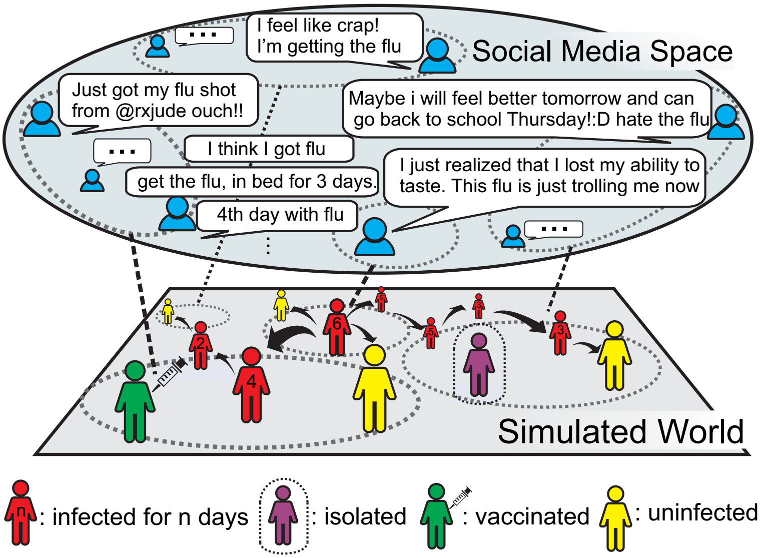

Social Media Space
I feel like crap! I'm getting the flu
...
Just got my flu shot from @rxjude ouch!!
Maybe i will feel better tomorrow and can go back to school Thursday!:D hate the flu
...
I think I got flu
get the flu, in bed for 3 days.
I just realized that I lost my ability to taste. This flu is just trolling me now
4th day with flu
...
Simulated World
2
4
6
5
3
n
: infected for n days
: isolated
: vaccinated
: uninfected

Fig. 1 In SimNest, the simulated world mirrors social media space. The posts of social media users reflect their health, vaccination, or isolation status. This information is mapped to the corresponding spatial subregions in the demographics-based contact network in the simulated world

media mining focus on different but complementary aspects of the problem, with the former focusing on modeling the underlying mechanisms of disease diffusion while the latter provides timely and detailed disease surveillance. The new SimNest framework proposed here utilizes both types of information by integrating their respective strengths.
- **A semi-supervised multilayer perceptron (MLP) has been developed for mining epidemic features**: To achieve a deep integration, we enforce unsupervised pattern constraints derived from the epidemic disease progress model onto the supervised classification. Using this semi-supervised strategy, the sparsity of labeled data can be solved.
- **A new designing an online stochastic training algorithm is presented**: To minimize the inconsistencies between Twitter space and the simulated world, we iteratively optimize model parameters via an online algorithm based on the stochastic gradient descent. This algorithm ingests the social media data streams and updates the model parameters in real time, thus not only reducing the cost of retraining but also ensuring the timeliness of the model.
- **Extensive experiments have been conducted to evaluate the new algorithm's performance**: The proposed SimNest model was evaluated using four real-world dataset and the results compared to those obtained by existing models. The proposed algorithm consistently outperformed the competing methods in multiple metrics. The performance for individual-level epidemics modeling and forecasting were also demonstrated and discussed.

The rest of this paper is organized as follows. Section 2 reviews existing work in this area. Section 3 presents the problem formulation. Section 4 elaborates the mathematical descriptions of the SimNest model, and Section 5 presents the parameter optimization for SimNest. Section 6 introduces the extended functions of SimNest. In Section 7, the extensive experimental results are analyzed. This paper concludes by summarizing the study's important findings in Section 8.

2 Related work

Computational models for epidemiology are important for a number of reasons. Traditionally, computational epidemiology has tended to focus on *compartmental models* where a population is divided into subgroups (compartments) based on people's health status and demographics, with the epidemic dynamics being modeled by ordinary differential equations [25, 29].

Recently, individual-based computational models have begun to be developed to support network epidemiology, where an epidemic is modeled as a stochastic propagation over an explicit interaction network between people. One common approach taken by network epidemiology is to model the interactions between people using random graph models [16, 21]. Here, the closed form analytical results obtained can be applied to study epidemic dynamics, but this relies on the inherent symmetries in random graphs. With no explicit location modeling, it cannot be applied to compute the geographical spread of an epidemic.

Another direction taken by network epidemiology is to develop a realistic representation of a population by considering members' social contact networks, and then using individual-based simulations to study the spread of epidemics within each network [5, 9]. This approach first constructs a synthetic population, where each individual is assigned demographic, geographic, social, and behavioral attributes so that at various aggregate levels the synthetic population is statistically indistiguishable from the real population. The synthetic individuals are also assigned daily activities and physical locations at any moment, so by connecting all those located within close proximity to each other one can construct the corresponding synthetic social contact network for the population [4]. Individual-based simulations model epidemics as diffusion processes across this network, computing who infects whom at what time and at which location [9]. In addition to the synthetic network and disease model, another key component of individual-based epidemic simulations is the associated set of public health and individual interventions carried out to control the epidemic, which can be either pharmaceutical in nature such as vaccination, or non-pharmaceutical such as social distancing. These interventions affect the epidemic evolution by changing the node or edge properties of the network.

There have been a number of influenza epidemic knowledge mining techniques proposed based on the use of social media, and these can be categorized into two threads. The first thread focuses on *aggregate level disease surveillance*. For example, Krieck et al. [23] suggested that self-reported symptoms are the most reliable signal in detecting whether a tweet is relevant to an outbreak or not and then went on to demonstrate that this is because even though people generally do not identify their specific problem until diagnosed by an expert, they readily write about how they feel. Using a similar approach to identify flu-related tweets, researchers have generally concentrated on tracking the overall trend of a particular disease outbreak, typically influenza, by monitoring social media [2, 22, 35–37].

The second thread focuses on *detailed health-informatics semantic analysis*. These approaches typically model the language of the social media messages and their relevance to public health [27, 30] influenza surveillance [15, 19], disease geoinformatics [18], user interactions [11], and health behavior [13, 31]. Paul and Dredze [27] proposed a topic model that captures the symptoms and possible treatments for ailments, and then went on to propose a way to identify the geographical patterns in the prevalence of such ailments. Specific to self-reporting on influenza, Collier et al. [15] categorized five sub-classes of tweets that serve as user behaviour response surveys for influenza outbreaks, while Dredze et al. [18] focused on achieving accurate geographical location identification for influenza outbreak detection and Brennan et al. [11] utilized Twitter user interactions to uncover the health condition of Twitter users. Tackling the problem from a different direction, Chen et al. [13] concentrated on modelling the disease progression in individuals.

3 Problem setup

This paper aims to characterize the spatiotemporal diffusion of influenza epidemics across their underlying social contact networks. Specifically, assume the time are split into T discrete time intervals $\mathcal{T} = \{0, \cdots, t, \cdots, T\}$. We aim to determine for each time interval $t \in \mathcal{T}$ the health states $\mathcal{Z}$ of the people in the population of interest, provided the social media data and demographics of the population as inputs. The occurrence time of each state transition within a time interval will be rounded to the boundaries of this time interval. To address this problem, approaches based on computational epidemiology and social media mining are formulated in turn below.

3.1 Individual-based epidemic simulation

A disease transmits through person-to-person contact. These contacts form a network called a social contact network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, which is a directed, edge-weighted network where nodes $\mathcal{V}$ correspond to individuals in the population. An edge $(v_1, v_2) \in \mathcal{E}$ with weight $\mathcal{W}(v_1, v_2)$ denotes the nodes v_1 and $v_2 \in \mathcal{V}$ has a contact of duration $\mathcal{W}(v_1, v_2)$. During the contact the disease may transmit from node v_1 to v_2 with probability $p(\mathcal{W}(v_1, v_2), \tau)$, where τ, the transmissibility, is the probability of transmission per unit of contact time and is a parameter associated with the disease. We first assume that the contact network $\mathcal{G}$ is constant. In Section 6, we will consider the situation when $\mathcal{G}$ changes due to interventions. To implement such social contact network, we apply the method in EpiFast [9], which follows the way of Episimdemics [7] consisting of two general steps. In Step 1, a simulated world with synthetic individuals is created based on the real-world demographics data. So in such simulated world, the individuals in each household with basic profiles are corresponding to those in each household in the real-world demographics; In Step 2, the social contact networks are derived from the synthetic population based on physical co-location of interacting persons, using the method that is proposed and described in [6]. The specific data and settings we adopted will be introduced in the experiment section.

Each person is assumed to be in one of the following four health states at any time: *susceptible (S), exposed (E), infectious (I), and recovered (R)*, which is known as the SEIR disease model and is widely used in the mathematical epidemiology literature [3, 25]. Associated with each person v is an incubation period $p_E(v)$ and an infectious period $p_I(v)$, each from a distribution. We assume that both are normally distributed, i.e., $p_E(v) \sim \mathcal{N}(\mu_E, \sigma_E)$ and $p_I(v) \sim \mathcal{N}(\mu_I, \sigma_I)$. A person is in the susceptible state until he becomes exposed. If a person v becomes exposed, he remains so for $p_E(v)$ days, during which he is not infectious. Then he becomes infectious and remains so for $p_I(v)$ days. Finally he recovers and remains so. The transition $S \mapsto E$ is probabilistic, but we assume that once person v becomes exposed, $p_E(v)$ and $p_I(v)$ are sampled from the two normal distributions respectively so their values are determined. Hence, given the parameters, let $Z_{v,t}(p_E(v), p_I(v)) \in \{S, E, I, R\}$ denote the health state of person $v \in \mathcal{V}$ for the time interval $t \in \mathcal{T}$. We then have $\mathcal{Z} = \{Z_{v,t}(p_E(v), p_I(v))\}_{v \in \mathcal{V}, t \in \mathcal{T}}$, where $\mathcal{Z}$ stands for peoples' inferred health status based on individual-based epidemic simulations.

3.2 Social media based user health state inference

Social media is a popular way for people to post messages about their everyday feelings, and is commonly treated as a surrogate for the physical world [2]. Taking Twitter as an example, suppose the set of Twitter users who have ever mentioned their flu infectiousness is denoted as $\mathcal{U} \subseteq \mathcal{V}$, which can increase with Twitter data streams. Each user $u \in \mathcal{U}$ posts $n_{u,t}$ tweets in each time interval t (e.g., hour, day), $t = 1, 2, \cdots, T$. Suppose we have a predefined set of keywords $\mathcal{K}$ related to flu. Define $X_{u,t} \in \mathbb{Z}^{|\mathcal{K}| \times 1}$ as the vector of keywords frequencies from the tweet postings of user u at time t. Hence, $X_u = \{X_{u,t}\}_t^{\mathcal{T}}$ denotes the keyword vectors of user u, while $\mathcal{X} = \{X_u\}_{u \in \mathcal{U}}$ denotes the set of all keyword vectors. We are interested in learning a classifier f_W that maps the social media user textual content $X_{u,t}$ to their corresponding health state $Y_{u,t}$:

$$f_W(X_{u,t}) : X_{u,t} \rightarrow Y_{u,t} \tag{1}$$

where $Y_{u,t} = \mathbf{1}[Z_{u,t} = I]$, I stands for "Infectious", and $\mathbf{1}[\cdot]$ stands for the indicator function. Thus, $Y_{u,t} = 1$ signifies that user u's health state $Z_{u,t}$ at time t is infectious (I); and $Y_{u,t} = 0$ that it is not. $Y_u = \{Y_{u,t}\}_t^{\mathcal{T}}$ denotes all the health states of user u. W denotes the parameter set of the classifier.

There are three main challenges when using either individual-based epidemic simulation or social media mining techniques individually: (1) There is as yet no surveillance data that is sufficiently real-time and fine-grained to permit the detailed progress of the epidemic simulation to be linked consistently with the physical world. (2) The person-to-person disease contact network and disease model is opaque to social media data. (3) The fast-streaming and time-evolving nature of huge social media data requires efficient updating of the trained model. Traditional batch-based training suffer from high expense and poor timeliness.

In order to overcome the above-mentioned challenges in each of the above threads when used individually, we propose simultaneously using both types of information by deeply integrating the strengths of individual-based epidemic simulation and social media mining techniques in our new framework, **S**oc**I**al **M**edia **N**ested **E**pidemic **S**imula**T**ion (**SimNest**), which is elaborated in the following section.

4 SimNest model

As shown in Fig. 2a, SimNest learns the users' health status from social media posts based on a multilayer feature representation. In addition to considering each time point individually, SimNest utilizes a disease progress model from computational epidemiology to constrain the temporal pattern of two aspects of health status: (1) constraining the infectious period to

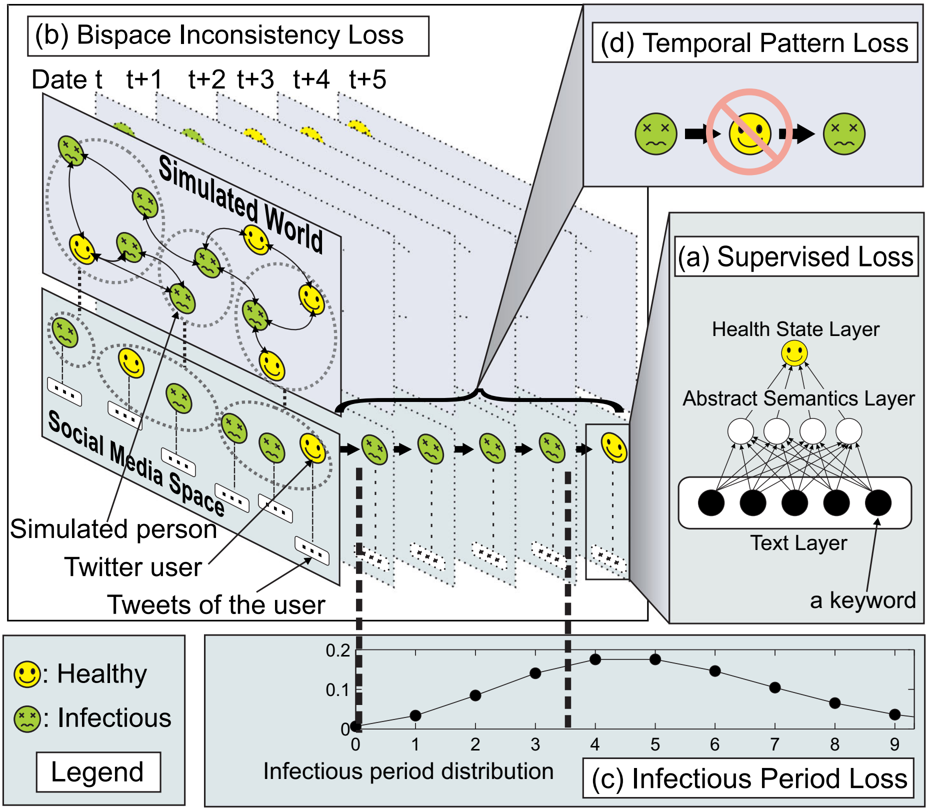


Fig. 2 The illustration of the SimNest model

follow a probability distribution, as depicted in Fig. 2c, and (2) resisting temporally discontinuous health states, as shown in Fig. 2d. Figure 2b illustrates how mapping social media users' health states into a demographics-based synthetic contact network can be used to implement interactive learning between these two spaces. Simulation model parameters are thus adjusted by the social media surveillance data while the weights of the multilayer-based health state model are regularized by the underlying synthetic disease contact network.

To ensure the underlying health states in the contact network $\mathcal{G}$ are consistent with those gathered from social media data D, SimNest simultaneously optimizes the contact network, disease progress model parameters p_I and p_E, and the social media-based health state inference $f_W(\cdot)$. Among all the keyword vectors $\mathcal{X}$, we are given a set of labeled samples $\mathcal{X}_1 = \{X_{u,t}\}_{u\in\mathcal{U}_1,t\in\mathcal{T}}$ with corresponding class label $\mathcal{Y}_1 = \{Y_{u,t}\}_{u\in\mathcal{U}_1,\mathcal{T}}$, and unlabeled samples $\mathcal{X}_2 = \{X_{u,t}\}_{u\in\mathcal{U}_2,t\in\mathcal{T}}$, where $\mathcal{U}_2 = \mathcal{U} - \mathcal{U}_1$ is the set of all the unlabeled users. Mathematically, the SimNest model is formulated as jointly minimizing four loss functions: (A) Supervised loss, (B) Bispace consistency loss, (C) Infectious duration loss, and (D) Temporal proximity loss, as illustrated below.

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_1(\mathcal{Y}_1, \mathcal{X}_1, W) + \mathcal{L}_2(\mathcal{X}_2, \mathcal{G}, p_E, p_I, W)\\ &+\mathcal{L}_3(\mathcal{X}_2, p_I, W) + \mathcal{L}_4(\mathcal{X}_2, W)\end{aligned} \tag{2}$$

These different loss functions are illustrated in Fig. 2. In the following subsections, we will discuss each in turn.

4.1 Supervised loss

To build an effective mapping $f_W(\cdot)$ between tweet texts and user health states, which is an abstract concept, we substantialize it by applying a deep data representation, namely multilayer perception:

$$\begin{aligned}f_W(x) = s(h^{(1)}) &= s\left(\sum\nolimits_{j=1}^{m} W_j^{(2)} s\left(h_j^{(2)}\right) + W_0^{(2)}\right),\\ h_j^{(2)} &= \sum\nolimits_{i=1}^{|\mathcal{K}|} W_{j,i}^{(1)} x_i + W_{j,0}^{(1)}\end{aligned} \tag{3}$$

Here, apart from the input layer that is the tweet text and the output layer that is the user health state, another hidden layer represents the abstract semantics, where m represents the number of hidden layer features. $W = W^{(1)} \cup W^{(2)}$, where $W^{(1)} \in \mathbb{R}^{|\mathcal{K}|\times m}$ is the weight matrix for the mapping from the text layer to the abstract semantics layer, $W^{(2)} \in \mathbb{R}^{m\times 1}$ is the weight vector for the mapping from the abstract semantics layer to the user health status layer and $s\ (\cdot)$ is the sigmoid function. $h^{(1)} = \sum_{j=1}^{m} W_j^{(2)} s(h_j^{(2)}) + W_0^{(2)}$.

A common way to learn W is to define a loss function over the training data, and then obtain the best W by minimizing the loss of misclassification towards labels:

$$\mathcal{L}_1 = \min_W \sum_u^{\mathcal{U}_1} \sum_t^{\mathcal{T}} \left\| f_W(X_{u,t}) - Y_{u,t} \right\|^2 \tag{4}$$

4.2 Bispace consistency loss

To sufficiently benefit from the complementary advantages of individual-based epidemic simulation and social media data, the inner inconsistency of the integrated model must be minimized. The hidden health states in the individual-based epidemic simulation need to be consistent with the observations from social media. On the other hand, the intelligence gleaned from the social media data also needs to correspond to the hidden disease progression across the hidden contact network. Expressing this more formally, our goal can be formulated in terms of the following loss function:

$$\mathcal{L}_2 = \min_{\Theta, W} \sum_{v}^{\mathcal{V}} \sum_{t}^{\mathcal{T}} \left\| Q_{v,t}(\mathcal{G}, p_E, p_I) - f_W(X_{v,t}) \right\|^2 \tag{5}$$

where $Q_{v,t}(\mathcal{G}, p_E, p_I) = \mathbf{1}[Z_{v,t}(p_E(v), p_I(v)) = I]$, and I stands for the "infectious" state, as noted in Section 3. $\Theta = \{\mathcal{G}, p_E, p_I\}$ are the parameters of the individual-based epidemic simulation and $p_E(v) \sim \mathcal{N}(\mu_E, \sigma_E)$ and $p_I(v) \sim \mathcal{N}(\mu_I, \sigma_I)$ are the incubation and infectious duration distributions for person v, respectively.

Although it is not possible to link the corresponding person to a specific user in Twitter, and not everybody posts tweets, the specific spatial subregion (e.g., blocks, counties, etc.) of Twitter user $u \in \mathcal{U}$ and simulated individual $v \in \mathcal{V}$ can be known. Hence, the above loss function can be transformed to a fine-grained spatial subregion:

$$\mathcal{L}_2 = \min_{\Theta, W, \lambda_1} \sum_{l,t}^{L,\mathcal{T}} \left\| \lambda_1 \sum_{v}^{\mathcal{V}_l} Q_{v,t}(\mathcal{G}, p_E, p_I) - \sum_{u}^{\mathcal{U}_{2,l}} f_W(X_u, t) \right\|^2 \tag{6}$$

where $\mathcal{U}_{2,l}$ denotes the Twitter users in location l, $\mathcal{V}_l$ denotes the people in location l, and λ_1 is the parameter scaling the person count in the individual-based epidemic simulation down to the count of social media users in that location.

4.3 Infectious period loss

Existing social media mining techniques typically do not assume a specific disease progression model and hence cannot take advantage of important knowledge patterns. In contrast, SimNest borrows the appropriate disease progression model from the epidemic simulation to regularize the patterns in the huge unlabeled social media data. This not only greatly mitigates the problem of label data sparsity, but also improves the timeliness and generalization of the modeling. In particular, the infectious duration d_u for a Twitter user $u \in \mathcal{U}$ will depend on the flu outbreak's specific characteristics as well as his or her general state of physical health, which is denoted as normal distribution here as it is one of the commonly used distributions for infection duration [14, 20]. Moreover, the maximum likelihood of normal distribution can be equivalent to a squared loss which is consistent with other loss terms in our objective terms and hence easier for optimization. However, the user could customize appropriate distribution freely according to the disease type and SimNest widely accommodates exponential family distributions. The normal distribution-based is as follows:

$$\left[\sum\nolimits_{t}^{\mathcal{T}} f_W(X_{u,t}) \right] = d_u \sim p_I(u) = \mathcal{N}(u|\mu_I, \sigma_I) \tag{7}$$

where the infectious duration d_u is calculated as $\sum_t^{\mathcal{T}} f_W(X_{u,t})$ because $f_W(X_{u,t}) = 1$ when infectious and $f_W(X_{u,t}) = 0$, otherwise. d_u is sampled from $p_I(u)$, the probability distribution of the infectious duration of the user u, which is a Gaussian distribution $\mathcal{N}(u|\mu_I, \sigma_I)$ with the mean μ_I and standard deviation σ_I. By maximizing the likelihood function for the observations, we can obtain the following objective function:

$$\max \prod_{u}^{\mathcal{U}_2} N(d_u|\mu_I, \sigma_I) = \max \sum_{u}^{\mathcal{U}_2} \log N \left(\sum_{t}^{\mathcal{T}} f_W(X_{u,t})|\mu_I, \sigma_I \right)$$

which can be transformed to the following formulation by considering Eq. 1:

$$\mathcal{L}_3 = \min_{W,p_I} \frac{1}{2\sigma_I^2} \sum_u^{\mathcal{U}_2} \left\| \sum_t^{\mathcal{T}} f_W(X_{u,t}) - \mu_I \right\|^2 + \frac{|\mathcal{U}_2|}{(2\pi\sigma_I^2)^{1/2}} \tag{8}$$

4.4 Temporal proximity loss

Another important intrinsic pattern in the health status modeling is that the states in the neighboring time points should be similar. Moreover, a person who is recovering from the flu typically cannot get the flu again in the same flu season, as illustrated in Figure 2(D). Thus, the infectious dates are temporally consecutive, leading to the following loss function for the proximity of the neighbor states:

$$\mathcal{L}_4 = \min_W \sum_u^{\mathcal{U}_2} \sum_t^{\mathcal{T}} \left\| f_W(X_{u,t}) - f_W(X_{u,t+1}) \right\|^2 \tag{9}$$

This loss function discourages repeated transition between “healthy” and “infectious” like shown in Section D in Fig. 2, which is unreasonable in real world. And it encourages a continuous status of infectiousness.

5 Online training algorithm

To efficiently solve the optimization problem presented in Eq. 2, we propose an online parameter optimization framework. The new framework adopts an alternating minimization approach [8], where all the variables are fixed except for the one being updated.

5.1 Solving for W

The process of solving W is based on stochastic gradient descent (SGD) [8]. Training with SGD makes it possible to handle very large databases since every update involves one (or a pair) of examples, and grows linearly in time with the size of the dataset. The convergence of the algorithm is also ensured for low enough values of threshold error. We elaborate the partial derivatives of the loss function in Eq. 2 with respect to the weight matrix W. This can be decomposed into the partial derivatives of each of the sub-loss functions $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$ in Equations 4, 6, 8, and 9, respectively.

$$\frac{\partial \mathcal{L}_{1,u,t}}{\partial W_{j,k}^{(1)}} = (f_W(X_{u,t}) - Y_{u,t}) s'\left(h^{(1)}\right) W_j^{(2)} s'\left(h_j^{(2)}\right) X_{i,k}^{(l)} \tag{10}$$

where $s'(x) = s(x) \cdot (1 - s(x))$.

$$\frac{\partial \mathcal{L}_{1,u,t}}{\partial W_j^{(2)}} = (f_W(X_{u,t}) - Y_{u,t}) \cdot s'(h^{(1)}) \cdot s\left(h_j^{(2)}\right) \tag{11}$$

where $\mathcal{L}_{1,u,t} = \mathcal{L}_1(f_W(X_{u,t}), Y_{u,t})$.

$$\begin{aligned} \frac{\partial \mathcal{L}_{4,u}}{\partial W_{j,k}^{(1)}} = & \left(f_W(X_{u,t}) - f_W(X_{u,t+1})\right) \cdot (s'\left(h^{(1)}\right) W_j^{(2)} s'\left(h_j^{(2)}\right) X_{u,t,k}^{(n)} \\ & - s'(\tilde{h}^{(1)}) W_j^{(2)} s'\left(\tilde{h}_j^{(2)}\right) X_{u,t+1,k}^{(n)} \Big) \end{aligned} \tag{12}$$

where $\mathcal{L}_{4,u} = \sum_t^{\mathcal{T}} \mathcal{L}_4(X_{u,t}, X_{u,t+1}, W)$.

$$\frac{\partial \mathcal{L}_{4,u}}{\partial W_j^{(2)}} = (f_W(X_{u,t}) - f_W(X_{u,t+1})) \cdot \left(s'\left(h^{(1)}\right) s\left(h_j^{(2)}\right) - s'\left(\tilde{h}^{(1)}\right) s\left(\tilde{h}_j^{(2)}\right)\right) \tag{13}$$

The derivative of $\mathcal{L}_3$ with respect to W is calculated as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_{3,u}}{\partial W_{j,k}^{(1)}} &= \sum_t^T \frac{\partial \mathcal{L}_{3,u,t}}{\partial s(h_t^{(1)})} \frac{\partial s\left(h_t^{(1)}\right)}{\partial h_t^{(1)}} \frac{\partial h_t^{(1)}}{\partial s(h_t^{(2)})} \frac{\partial s\left(h_t^{(2)}\right)}{\partial h_t^{(2)}} \frac{\partial h_t^{(2)}}{\partial W_{j,k}^{(1)}} \\ &= \sum_t^T \left(\sum_i^T f_W(X_{u,i}) - \mu_I\right) s'\left(h_t^{(1)}\right) W_j^{(2)} s'\left(h_{t,j}^{(2)}\right) X_{u,t,k}^{(n)} \end{aligned} \tag{14}$$

where $\mathcal{L}_{3,u} = \sum_t^{\mathcal{T}} \mathcal{L}_{3,u,t}$, and $\mathcal{L}_{3,u,t} = \mathcal{L}_3(X_{u,t}, W, p_I)$.

$$\begin{aligned} \frac{\partial \mathcal{L}_{3,u}}{\partial W_j^{(2)}} &= \sum_t^{\mathcal{T}} \frac{\partial \mathcal{L}_{3,u,t}}{\partial s(h_t^{(1)})} \frac{\partial s(h_t^{(1)})}{\partial h_t^{(1)}} \frac{\partial h_t^{(1)}}{\partial W_j^{(2)}} \\ &= \sum_t^T \left(\sum_i^T f_W(X_{u,i}) - \mu_I\right) s'\left(h_t^{(1)}\right) s\left(h_{t,j}^{(2)}\right) \end{aligned} \tag{15}$$

Similarly, the derivative of $\mathcal{L}_2$ is as follows:

$$\frac{\partial \mathcal{L}_{2,l,t}}{\partial W_{j,k}^{(1)}} = \sum_u^{\mathcal{U}_{2,l,t}} \left(\sum_v^{\mathcal{U}_{2,l,t}} f_W(X_{v,t}) - \sum_v^{\mathcal{V}_{l,t}} Q_v(p_E, p_I)\right) \cdot s'\left(h_u^{(1)}\right) W_j^{(2)} \cdot s'\left(h_{u,j}^{(2)}\right) \cdot X_{u,k} \tag{16}$$

where $\mathcal{L}_{2,l,t} = \mathcal{L}_2(X_{l,t}, W, \mathcal{Z})$.

$$\begin{aligned} \frac{\partial \mathcal{L}_{2,l,t}}{\partial W_j^{(2)}} &= \sum_u^{\mathcal{U}_{2,l,t}} \sum_v^{\mathcal{U}_{2,l,t}} f_W(X_{v,t}) s'\left(h_u^{(1)}\right) s\left(h_{u,j}^{(2)}\right) \\ &\quad - \sum_u^{\mathcal{U}_{2,l,t}} \sum_v^{V_{r,t}} Q_v(p_E, p_I) s'\left(h_u^{(1)}\right) s\left(h_{u,j}^{(2)}\right) \end{aligned} \tag{17}$$

5.2 Solving for Θ

Solving for $\Theta = \{\mathcal{G}, p_E, p_I\}$ with respect to the loss function $\mathcal{L}_2$ is a nonconvex and non-differentiable problem, so a numerical optimization algorithm such as the Nelder-Mead method [8] can be adopted to solve it.

5.3 Solving for p_I, λ_1

The sufficient statistics μ_I and σ_I of the infectious period distribution p_I have the following analytical solution:

$$\mu_I = \frac{1}{|\mathcal{U}_2|} \sum_u^{\mathcal{U}_2} \sum_t^{\mathcal{T}} f_W(X_{u,t}) \tag{18}$$

$$\sigma_I = \left(\sum_u^{\mathcal{U}_2} \sum_t^{\mathcal{T}} f_W(X_{u,t} - \mu_I)/|\mathcal{U}_2|\right)^{1/2} \tag{19}$$

Solving for λ_1 according to the loss function $\mathcal{L}_2$ in Eq. 6 yields the following analytical solution:

$$\lambda_1 = \sum_{l,t}^{L,\mathcal{T}} \sum_u^{\mathcal{U}_{2,l}} f_W(X_u, t) / \sum_{l,t}^{L,\mathcal{T}} \sum_v^{\mathcal{V}_l} Q_{v,t}(\mathcal{G}, p_E, p_I) \tag{20}$$

Utilizing the above alternating optimization process, SimNest is trained and utilized to forecast the spatiotemporal epidemic diffusion progress online as illustrated in Algorithm 1. Specifically, the unlabeled data set $\mathcal{X}$ is continually updated by the social media data streams, with the most out-dated information (which can be as much as three months old) being replaced by the newly-arriving data. Then, the weight matrix W is optimized via SGD until convergence. Utilizing the optimized infectious period distribution as the input for the simulation process, the epidemic simulation parameter p_E is optimized by minimizing the inconsistencies with social media data. Finally, the population's health status $\mathcal{Z}$ is predicted. The optimized parameter p_E is then utilized for the next-step optimization of weight matrix W with the updated unlabeled data. Therefore, as the data is streaming, the parameters are being optimized with the newest data and the predicted health status $\mathcal{Z}$ streams out.

Algorithm 1: Online algorithm for SimNest.

```
Input: Data matrix $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, Twitter data stream $\mathcal{C}$, contact network $\mathcal{G}$.
Output: the population's predicted health status $\mathcal{Z}$.
Set the learning rate $\eta = 0.5$. Initialize weight matrix $W$ as matrix of random values between $-1$ and 1;
repeat
    Update unlabeled data set $\mathcal{X}_2$ by Twitter data stream;
    repeat
        Randomly select a labeled sample $(X_{u,t}, Y_{u,t})$;
        $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_1(X_{u,t}, Y_{u,t}, W)}{\partial W}$;
        Randomly select an unlabeled sample $X_u$;
        $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_3(X_u, p_I, W)}{\partial W}$;
        Randomly select an unlabeled sample $X_v$;
        for $i \leftarrow 1$ to $T$ do
            $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_4(X_{v,i}, X_{v,i+1}, W)}{\partial W}$
        end
        Randomly select a user $u$ from a location $l \in L$;
        $W \leftarrow W - \eta \cdot \frac{\partial \mathcal{L}_2(X_{u,t}, \mathcal{G}, p_E, p_I, W)}{\partial W}$;
        $\mu_I \leftarrow \frac{1}{|\mathcal{U}_2|} \sum_u^{\mathcal{U}_2} \sum_t^{\mathcal{T}} f_W(X_{u,t})$;
        $\sigma_I \leftarrow (\sum_u^{\mathcal{U}_2} \sum_t^{\mathcal{T}} f_W(X_{u,t} - \mu_I)/|\mathcal{U}_2|)^{1/2}$;
    until converge;
    $p_E, \mathcal{Z} \leftarrow \min \sum_t^{\mathcal{T}} \sum_l^{L} \left\| \lambda_1 \sum_v^{\mathcal{V}_l} Q_{v,t}(\mathcal{G}, p_E, p_I) - \sum_u^{\mathcal{U}_{2,l}} f_W(X_{u,t}) \right\|^2$;
    $\lambda_1 \leftarrow \sum_{l,t}^{L,\mathcal{T}} \sum_u^{\mathcal{U}_{2,l}} f_W(X_u, t) / \sum_{l,t}^{L,\mathcal{T}} \sum_v^{\mathcal{V}_s} Q_{v,t}(\mathcal{G}, p_E, p_I)$
until the end of data stream;
```

6 Extensions

6.1 Dynamics of contact network

Interventions are typically the most common and effective ways for both the government and individuals to reduce the potential impact of a disease outbreak, influencing the epidemic diffusion largely by changing the people-people contact network. These can be categorized into two types: (1) Pharmaceutical (PI) and (2) Non-pharmaceutical (NPI). PI interventions, such as administering antivirals and vaccines, can change the characteristics (e.g., disease transmissibility) of the person nodes in the social contact network, while NPI interventions are those actions that effectively change the contact network structure, including school closures, quarantine and sequestration. Therefore, both types of interventions can result in changes in the social contact network.

The SimNest framework accommodates these heterogeneous dynamics of contact networks effectively via two aspects: (1) Timely intervention action monitoring based on social media data; and (2) Intervention substantialization through the epidemic simulation process. Take vaccination as an example. First, tweets like "I just got flu shot, it still hurts." that mention their user $\mathcal{U}_l$'s vaccinations from each subregion $l \in L$ are identified by the text classifiers. In our experiments, we achieved a 78% identification accuracy based on the cross-validation results. For example, Fig. 3 shows the users who got the flu shots that were identified through their Twitter postings during Jan 2011 and Jan 2013 in Virginia. It clearly demonstrates both yearly and weekly periodicity, with a peak around November each year. The relative vaccination ratio in different subregions can then be estimated as $r_l = |\mathcal{U}_l|/\lambda_1|\mathcal{V}_l|$, where $|\mathcal{V}_l|$ is the size of the population in subregion l and λ_1 is the population size scaling factor from the physical world to the Twittersphere, as calculated by Eq. 20. Next, in the epidemic simulation SimNest substantializes the vaccinations by reducing the transmissibility $p(\mathcal{W}(v_1, v_2))$, $(v_1 \in \mathcal{V}_l$ or $v_2 \in \mathcal{V}_l)$ for $r_l \cdot |\mathcal{V}_l|$ random individuals in region l by some ratio, which can either be set by domain knowledge or from the literature.

6.2 Heterogeneous surveillance data

The SimNest framework is also sufficiently flexible to incorporate multiple surveillance data sources. In our basic problem definition, we only utilized social media data as a fine-grained surveillance data. However, SimNest also allows the addition of heterogeneous surveillance data sources such as the CDC [12] surveillance data for the United States, and the parallel PAHO [26] surveillance data for Latin America. Taking the CDC surveillance data as an example, which reports state-level weekly aggregate data, in order to be comparable,

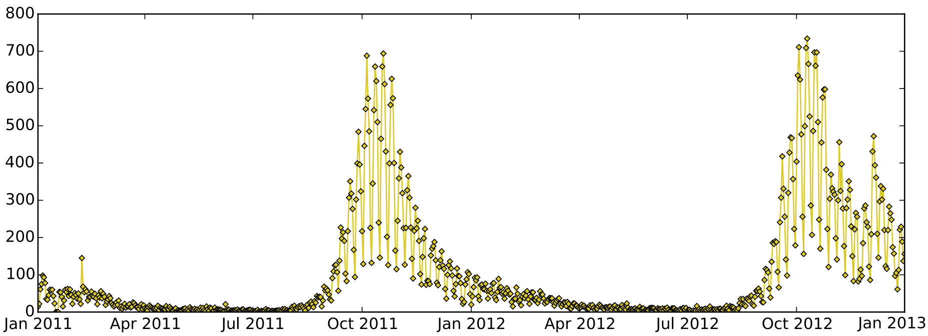


Fig. 3 Counts of Twitter users in Virginia who got flu shot

SimNest aggregates the predicted user health states into state-level weekly data and inserts the following loss function into Eq. 2, yielding the following:

$$\mathcal{L}_c = \min_{W,\lambda_2} \sum_{i}^{T'} \|\lambda_2(a_e - a_s + 1) \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} f_W(X_{u,t}) - C(i)\|^2$$

where $C(i)$ denotes the additional surveillance data for the ith time interval. Assume τ' denotes the time interval between two consecutive data points of C, and τ is the interval of time step of the discrete simulation system. T' is defined as the number of timepoints of the surveillance data such that $T' = \lfloor T \cdot \tau'/\tau \rfloor$, $a_s = \lfloor i \cdot \tau'/\tau \rfloor$, $a_e = \lfloor (i+1) \cdot \tau'/\tau \rfloor - 1$. λ_2 is the scaling parameter.

The derivative of $\mathcal{L}_c$ with respect to W is as follows:

$$\frac{\partial \mathcal{L}_{c,i}}{\partial W^{(1)}_{j,k}} = \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} \left(\lambda_2 \alpha \cdot \sum_{l,p=a_s}^{L,a_e} \sum_{v}^{\mathcal{U}_{2,l,p}} f_W(X_{v,p}) - C(i) \right) \cdot s'\left(h_t^{(1)}\right) W_j^{(2)} s'\left(h_{t,j}^{(2)}\right) X^{(n)}_{u,t,k} \tag{21}$$

where $\alpha = (a_e - a_s + 1)$.

$$\begin{aligned} \frac{\partial \mathcal{L}_{c,i}}{\partial W^{(2)}_{j}} &= \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} \lambda_2 \alpha \sum_{l,p=a_s}^{L,a_e} \sum\nolimits_{v}^{\mathcal{U}_{2,l,p}} f_W(X_{v,p}) s'\left(h_t^{(1)}\right) s\left(h_{t,j}^{(2)}\right) \\ &\quad - C(i) \sum\nolimits_{l,t=a_s}^{L,a_e} \sum\nolimits_{u}^{\mathcal{U}_{2,l,t}} s'\left(h_t^{(1)}\right) s\left(h_{t,j}^{(2)}\right) \end{aligned} \tag{22}$$

In addition, the analytical solution of the scaling factor λ_2 is as follows:

$$\lambda_2 = \sum\nolimits_{i}^{T'} M_i \cdot C(i) / \sum\nolimits_{i}^{T'} M_i^2 \tag{23}$$

where $M_i = (a_e - a_s + 1) \sum_{l,t=a_s}^{L,a_e} \sum_{u}^{\mathcal{U}_{2,l,t}} f_W(X_{u,t})$.

7 Experiments

In this section, the performance of the proposed SimNest model is evaluated using real world data. After describing the experimental setup, the effectiveness of the SimNest model on state-level influenza epidemic forecasting is demonstrated by comparing its performance with those of 7 comparison methods. The new model's ability to forecast events in fine-grained geographical subregions is also evaluated. This section concludes by presenting a case study on the dataset for Connecticut.

7.1 Experiment setup

This subsection presents the data preparation, label set and performance metrics.

7.1.1 Dataset

Twitter data The input Twitter data in this paper was preprocessed by the following process to retain the flu-related tweets. First, we queried the Twitter API using flu-related keywords and retrieved the data for the period Jan 1, 2011 to Apr 15, 2015 for the entire United States. The flu-related keywords were extracted by domain expert on epidemiology based on the

Table 1 Real-world census data and Twitter datasets

	Demographics	Twitter		
state	population size	#connections	#tweets	#users
CT	3,518,288	175,866,264	9,513,741	10,257
MA	6,593,587	332,194,314	19,785,147	15,005
MD	5,699,478	285,159,648	20,754,218	19,758
VA	7,882,590	407,976,012	15,899,713	14,302

glossary of influenza[1] which included terms such as “flu”, “influenza”, and “h1n1”, among others. Then the retrieved tweets were classified according to whether or not they were flu-related. Those tweets unrelated to flu or did not talk about the status of the author himself/herself will be filtered out by the classifier. For the classifier, we adopted LibShortText [34], a logistic regression model specially designed for classifying short text like tweets. The classifier was trained on an existing labeled training set provided by Lamb et al. [24]. This training set formed our labeled tweets set, consisting of the tweets $\mathcal{X}_1$ and their labels $\mathcal{Y}_1$ as discussed in Section 4. The input features $\mathcal{K}$ of this model were the disease keywords provided by Paul and Dredze [27]. The preprocessed Twitter dataset is denoted as $\mathcal{D}$.

The authors $\mathcal{U}_2$ of the preprocessed tweets set $\mathcal{D}$ were extracted and the tweets they posted during the two weeks before and after their tweets in $\mathcal{D}$ were retrieved via Twitter API. After removing retweets, this Twitter data set was geocoded and only those tweets sent from within the location of interest retained to form the unlabeled Twitter data set $\mathcal{X}_2$ defined in Section 4. Four states, namely Connecticut (CT), Massachusetts (MA), Maryland (MD), and Virginia (VA) were utilized for this performance evaluation. The Carmen geocoder [18] was utilized to resolve the location of each tweet into a tuple containing information at the country, state, county, and city level. About 70% of the tweets in our dataset were assigned with a location by Carmen.

U.S. census data The household structure and demographics utilized in the construction of our social contact network are derived from U.S. Census data.[2] In these datasets, each person has attributes including age, income, gender, and household size, while each location has attributes including coordinates, land use, and business type. To generate the contact network, we utilized the actual demographic data for each region. In this paper, we focus on four regions, Connecticut (CT), Massachusetts (MA), Maryland (MD), and Virginia (VA). Information about the Twitter data and demographics for the four regions are shown in Table 1.

7.1.2 Labels and metrics

For the proposed model and all the competing methods, the data between Aug 1, 2011 and Jul 31, 2012 was utilized as the training season, while the data between Aug 1 2012 and Jul 31 2014 was used for predicting. The forecasting results for the flu outbreaks were validated against the corresponding influenza statistics reported by the Centers for Disease Control

[1] https://www.cdc.gov/flu/glossary/index.htm

[2] https://www.census.gov/data.html

and Prevention (CDC). The CDC publishes the percentage of the number of physician visits related to influenza-like illness (ILI) weekly for each major region in the United States.

In the experiment, four metrics were adopted, namely mean squared error (MSE), Pearson correlation, p-value, and peak time error. MSE represents the mean value of the squared errors between all the predicted data points and corresponding label points. The Pearson correlation is the covariance of the predicted and label data points divided by the product of their standard deviations. This varies from -1 to 1 and the larger the value, the stronger the positive correlation between the two sets of data points. The p-value indicates how likely the hypothesis of no correlation between the predicted and label data points is to be true; the smaller the p-value, the more statistically significant the Pearson correlation. Lastly, peak time error represents the time interval between the predicted peak time (i.e., the week with the highest number of infected people) and the actual peak time reflected by the CDC label data.

7.1.3 Comparison methods

The performance of the new SimNest model proposed here was compared with those of 8 other methods. Of these, 5 methods originated from social media mining: *Linear Autoregressive Exogenous model (LinARX)* [1], *Logistic Autoregressive Exogenous model (LogARX)* [2], *Multi-variable linear regression model (multiLinReg)* [17], *Linear Regression model (LinReg)* [22], and our two baselines *Semi-supervised MLP+LinARX (semiLinARX)*, and *Semi-supervised MLP+LogARX (semiLogARX)*. The remaining 2 methods came from computational epidemiology: *SEIR* [25] and *EpiFast* [9].

(1) *Linear Autoregressive Exogenous model (LinARX)* [1]: This standard ARX model built the dependence of future visit percentage on the historical time series for the CDC's ILI visit percentage data [12] and the volume of influenza tweet data $\mathcal{D}_{(+)}$. The orders of LinARX for the Twitter data time series and CDC time series were set as 2 and 3, respectively, based on cross-validation.

(2)) *Logistic Autoregressive Exogenous model (LogARX)* [2]: Building on their earlier LinARX model, this method added a logit function transformation to the historical time series to enforce the boundary 0-1 of the value of the ILI visit percentage. The orders of LogARX for the two time series were both set as 2 based on cross-validation.

(3) *Simple Linear Regression model (LinReg)* [22]: This method assumed a linear mapping between the input, the volume of infectious tweets $\mathcal{D}_{(+)}$, the output, and the future ILI visit percentages.

(4) *Multi-variable linear regression model (multiLinReg)* [17]: This method treats a combination of keywords $\mathcal{K}$'s volumes as a multivariate input of the simple regression model.

(5) *EpiFast* [9]: This model followed the definition in Section 3.1, and mainly utilized two parameters to tune, p_E and p_I. These were optimized by minimizing the error of the predicted and the actual ILI visit percentage via the Nelder Mead method [9].

(6) *Semi-supervised MLP+LinARX (semiLinARX)*: This method built the classifier $f_W(\cdot)$ by simultaneously minimizing the loss functions $\mathcal{L}_1$, $\mathcal{L}_3$, and $\mathcal{L}_4$ in Eqs. 3, 8, and 9. It used both labeled set $\mathcal{X}_1 \cup \mathcal{Y}_1$ and unlabeled set $\mathcal{X}_2$, as well as input feature $\mathcal{K}$.

(7) *SEIR* [25]: This model divided the population into four health states, namely susceptible (S), exposed (E), infectious (I), and recovered (R). The epidemic dynamics were

modeled using ordinary differential equations and visit percentage calculated by multiplying the volume of the state "I" by a ratio, which was optimized by cross-validation. The volume of the positive tweets classified was fed into LinARX. The orders of the LinARX model for both time series (Twitter data CDC and surveillance data) were set as 2 based on cross-validation.

(8) *Semi-supervised MLP+LogARX (semiLogARX)*: Using the same semi-supervised MLP as semiLinARX, the volume of the positive tweets classified is fed into LogARX. The orders of LogARX for both time series of Twitter data and CDC surveillance data were again set as 2 based on cross-validation.

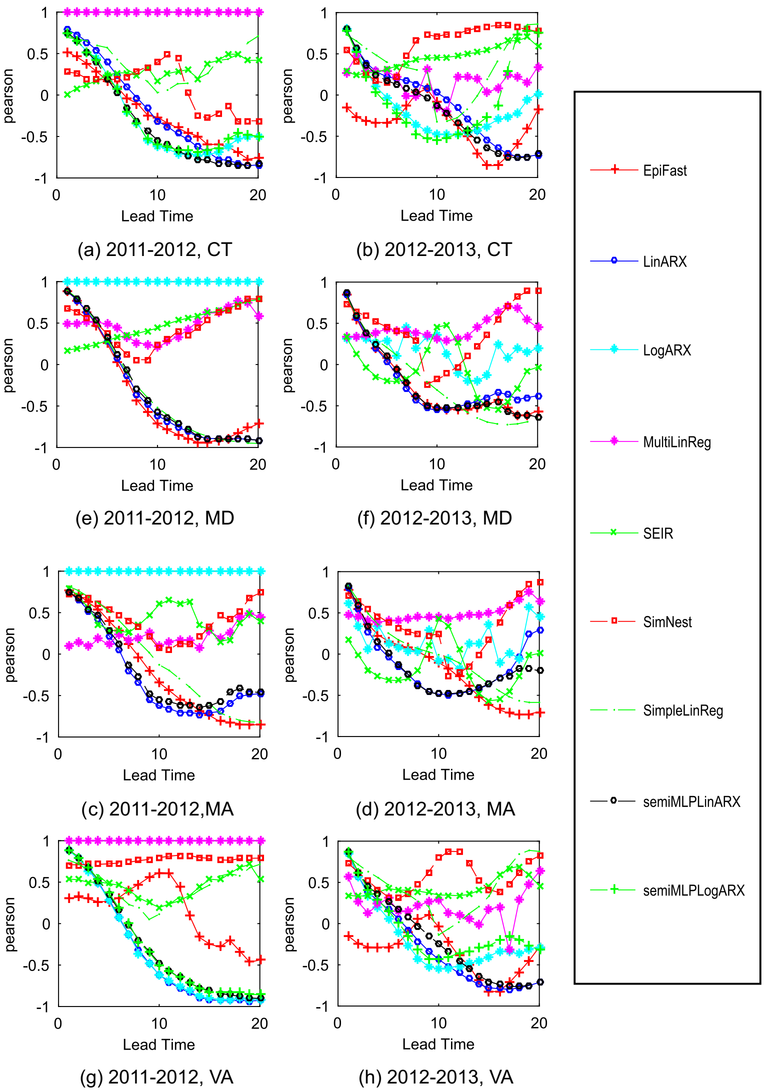
7.2 State-level performance evaluation of influenza epidemics modeling

The performance of each model for forecasting the percentage of ILI visits for each state with different lead times was evaluated. The lead times were varied from 1 week to 20 weeks, so every method was asked to forecast the data from 1 week out to 20 weeks in the future. The performance was evaluated in terms of the 4 different metrics introduced above for all 4 datasets for three seasons, utilizing the training set and test set for each. For the purposes of this research, every season was deemed to start on August 1st and ends on July 31 the following year. In this experiment, our SimNest model included the extensions described above in Section 6.

7.2.1 Performance on the Pearson correlation and p-value

Figures 4 and 5 show the forecasting performance achieved by all 9 of the models in terms of the Pearson correlation and p-value. Overall, the social media-based methods (i.e., LinARX, LogARX, multiLinReg, semiMLPLinARX, semi-MLPLogARX, and LinReg) typically achieved high Pearson correlations of between 0.6-0.95 for short lead times of less than 2 weeks, but the Pearson correlation decreased to below 0 as the lead time increased to 20 weeks. multiLinReg achieved the best performance on training set, but performed poorly on test set, which shows it has over-fitting. The p-values confirmed the statistical significance of the high Pearson correlation for lead times below 2 weeks. The computational epidemiology-based methods (i.e., SEIR and EpiFast) did not perform as well as the social media-based methods for short lead times, but the Pearson correlations did not drop significantly as the lead times increased. For example, SEIR still achieved a Pearson correlation of around 0.6 for lead times as long as 20 weeks. The reasons for this are two-fold. First, social media-based methods benefit from the availability of real-time surveillance data, while computational epidemiology-based methods use CDC data with its inherent 1-2 week time lag. This important difference means that the former enjoy a significant advantage when predicting data points in the very near future. Second, social media-based methods are purely data-driven, while computational epidemiology methods make use of a long-term disease progression mechanism. This makes computational epidemiology less sensitive to current data and more robust in terms of the overall performance.

According to Fig. 4, the proposed new SimNest model achieved the best overall performance in the terms of Pearson correlation, achieving the highest correlation score in two states: Virginia and Connecticut. It also performed second best overall on the training datasets for Massachusetts and Maryland, and was consistent among the top 2 methods in the test datasets for these states. Compared to social media-based methods such as LinARX,



(a) 2011-2012, CT (b) 2012-2013, CT

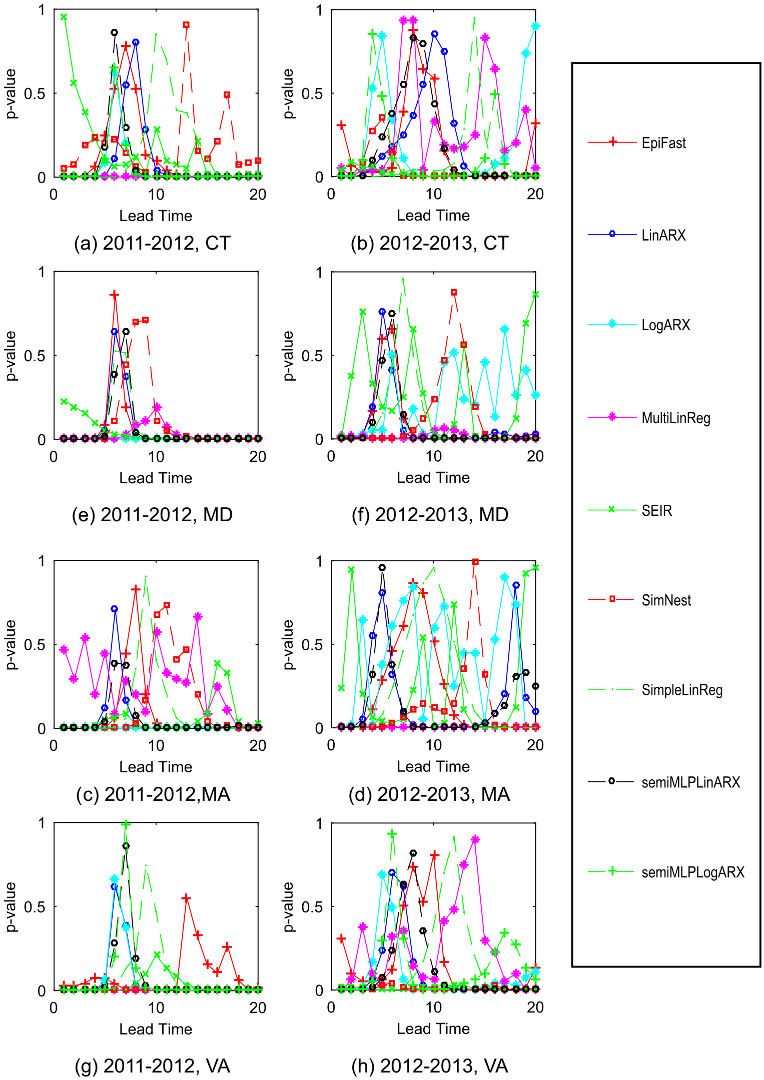
(e) 2011-2012, MD (f) 2012-2013, MD

(c) 2011-2012,MA (d) 2012-2013, MA

(g) 2011-2012, VA (h) 2012-2013, VA

Fig. 4 ILI visits percentage forecasting performance: Pearson correlation

SimNest outperformed them all by a large margin (over 0.3 in most cases) for lead times over 10 weeks. Furthermore, SimNest consistently performed far better than the computational epidemiology-based methods for lead times shorter than 10 weeks and also did better for lead times longer than 10 weeks. Overall, by combining the complementary strengths from social media-based and computational epidemiology-based methods, SimNest consistently achieved the best overall performance for both short and longer lead times. Figure 5 shows the p-values corresponding to the Pearson correlation scores in Fig. 4, which confirms that SimNest generally achieved the lowest p-values for all the different lead times. In all 4 datasets, the p-values for SimNest were generally below 0.05 for almost all the



(a) 2011-2012, CT

(b) 2012-2013, CT

(e) 2011-2012, MD

(f) 2012-2013, MD

(c) 2011-2012,MA

(d) 2012-2013, MA

(g) 2011-2012, VA

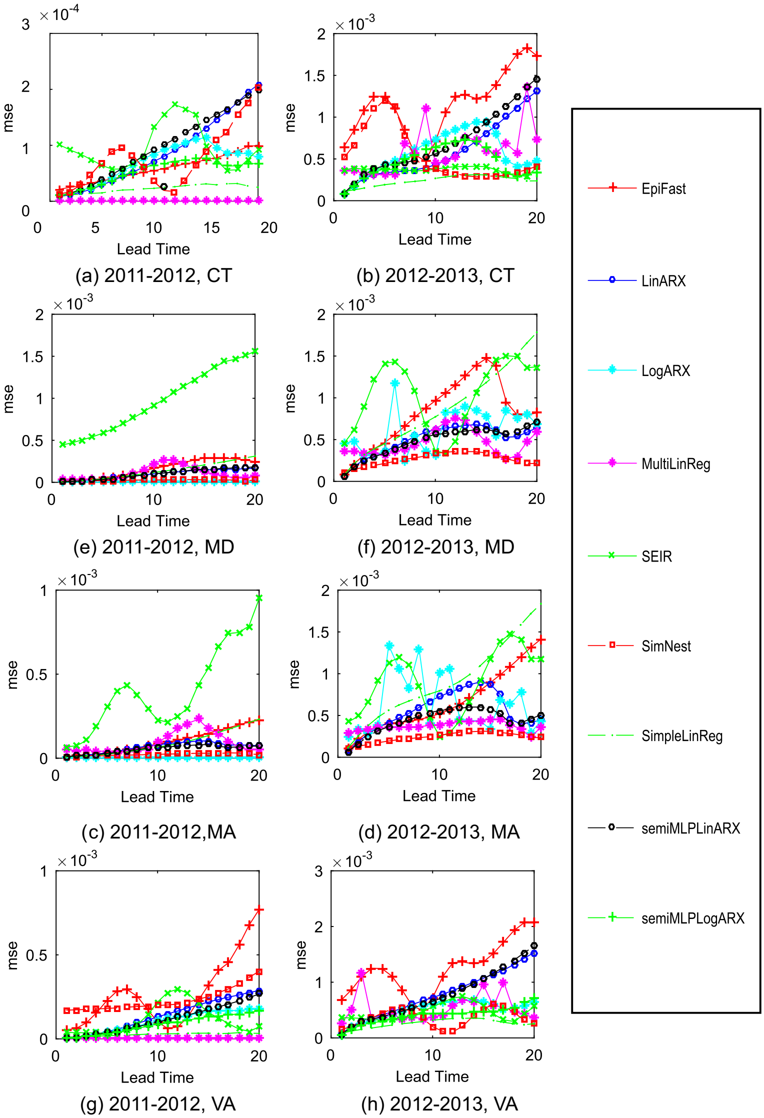
(h) 2012-2013, VA

Fig. 5 ILI visits percentage forecasting performance: p-value

different lead times. These consistently low p-values indicate the strong statistical significance of SimNest's advantageous Pearson correlation results.

7.2.2 Performance on mean squared error (MSE)

Figure 6 illustrates the performance on MSE and peak time error for all the methods. The social media-based methods again outperformed the computational epidemiology-based methods for short lead times as demonstrated by their achieving lower MSEs, which was also reflected by the Pearson correlations shown in Fig. 4. As the lead times increased, however, the MSEs for the social media-based methods typically increased by 5-10 times.



(a) 2011-2012, CT

(b) 2012-2013, CT

(e) 2011-2012, MD

(f) 2012-2013, MD

(c) 2011-2012,MA

(d) 2012-2013, MA

(g) 2011-2012, VA

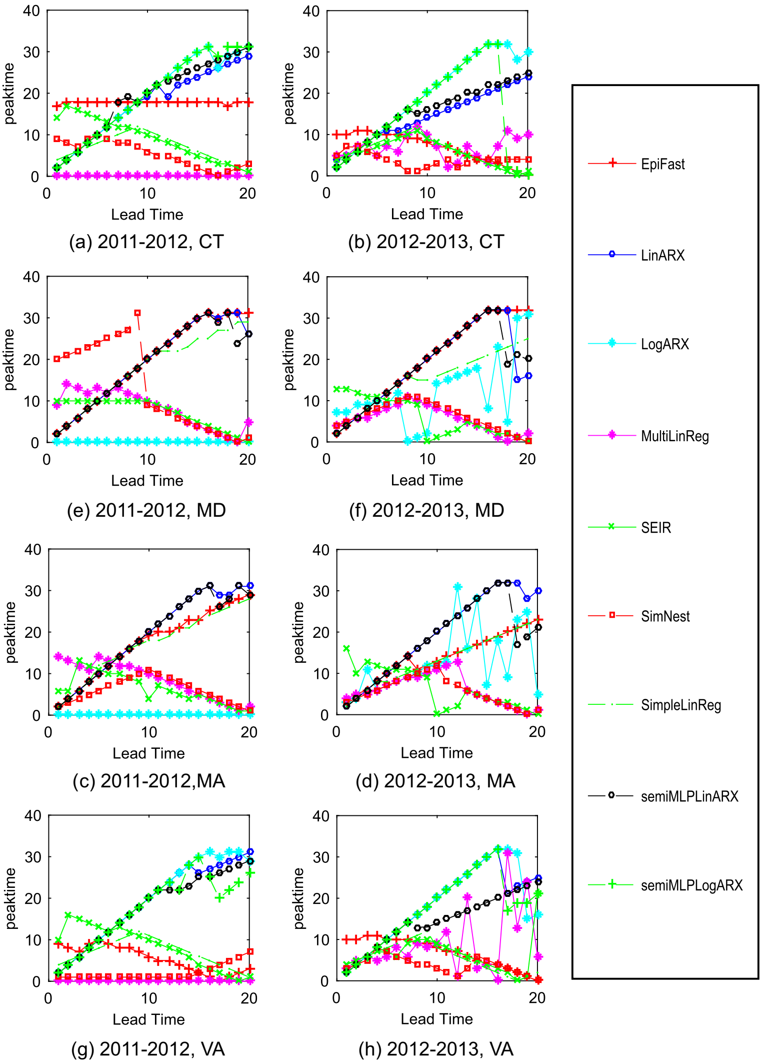
(h) 2012-2013, VA

Fig. 6 ILI visits percentage forecasting performance: mean squared error (MSE)

Compared to these methods, the MSEs for the computational epidemiology-based methods such as SEIR generally started with larger MSEs but these did not consistently increase as the lead times became larger. Among all the methods, SimNest generally achieved the smallest MSEs of less than 5×10^{-4} in both the training and testing sets of all the datasets. Specifically, SimNest started with low MSEs and exhibited no obvious increase in MSEs as the lead times became longer. When the lead time was short, it obtained better performances than the computational epidemiology-based methods and as the lead time grew it became increasingly advantageous compared to the social media-based methods. These results again demonstrate the advantage of combining the complementary strengths from social media-based and computational epidemiology-based methods.

7.2.3 Performance on peak time error

Figure 7 illustrates the performance in terms of the peak time error for all the methods. In general, the computational epidemiology-based methods achieved better performances than the social media-based methods. This is reasonable because the prediction of the peak time of epidemic outbreaks typically requires strong prior knowledge. For example, the seasonal flu A outbreaks typically occur between December and February in the US. Computational epidemiology-based methods model the mechanism of disease and intervention and thus



(a) 2011-2012, CT

(b) 2012-2013, CT

(e) 2011-2012, MD

(f) 2012-2013, MD

(c) 2011-2012,MA

(d) 2012-2013, MA

(g) 2011-2012, VA

(h) 2012-2013, VA

Fig. 7 ILI visits percentage forecasting performance: peak time error

can take into account this prior knowledge. For lead times shorter than 5 weeks, the social media-based methods are very competitive, which is similar to the results shown in Figs. 4 and 6. However, for lead times longer than 5 weeks, the computational epidemiology-based methods start to dominate the social media-based methods when the performance is assessed in terms of the peak time error. For example, EpiFast achieved a peak time error of less than 3 weeks when the lead time was over 10 weeks for the Massachusetts, Maryland, and Connecticut datasets. In contrast, LinARX exhibited 5-10 weeks peak time errors for lead times of below 10 weeks, and around 15 weeks when the lead time was over 10 weeks. SimNest had a clear advantage over computational epidemiology-based methods, not only achieving the peak time errors as small as the computational epidemiology-based methods for lead times of over 10 weeks, but also outperforming them by a significant margin for lead times below 10 weeks.

7.2.4 Efficiency

As shown in Table 2, the total runtime for our SimNest on four datasets ranges from 2 - 4 hours in order to get the simulation results for the future epidemic situations. Notice that such runtime is achieved on each dataset containing large networks with at least millions of nodes and almost billions of edges, which is superior to most of the individual-based simulation methods, thanks to the high efficiency of EpiFast [10]. The efficiency of our simulation-data-driven hybrid system is definitely lower than simplest machine learning models such as autoregressive and linear regression, but its advantages in accuracy, especially in longer-term forecasting is very important to epidemics modeling and intervention. In addition, we observe that the runtime increases almost linearly with the size of the network, which indicates good scalability. Finally, our method could be able to be further accelerated by adding more CPU cores for the parallel computing in our simulation model component based on EpiFast framework, according to the scalability analysis of EpiFast method [10].

7.3 Inner-state performance evaluation of influenza epidemics modeling

Because of the way it models the demographics, disease mechanism, and disease contact networks, the new model proposed here, SimNest can perform individual-level epidemic modeling and forecasting. Traditional social media-based methods are unable to achieve acceptable individual-level results when the surveillance data is coarse-grained in spatial resolution. For example, the ILI visits percentage surveillance data reported by CDC is state-wide and covers a week at a time, which seriously limits its utility for individual-level research.

In this section, a detailed performance evaluation of influenza epidemic modeling is presented and discussed. The dataset for Connecticut is used here as an example because

Table 2 Runtime of SimNest on different datasets

	VA	CT	MD	MA
#Nodes	7,882,590	3,518,288	5,699,478	6,593,587
#Edges	407,976,012	175,866,264	285,159,648	332,194,314
Runtime (sec)	14,566	6,950	12,391	13,260

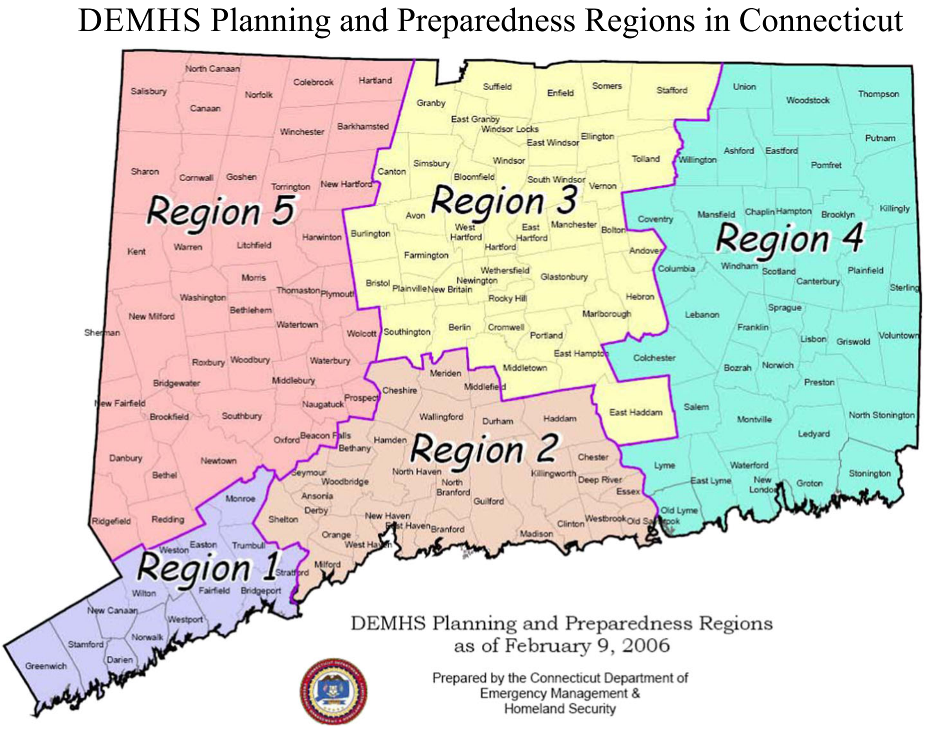


Fig. 8 DEMHS regions of the state Connecticut

Connecticut provides within-state flu activity reports that we can use to validate the performance of our model against. The performance of our proposed SimNest method is compared against that of the EpiFast method [9]. After conducting a quantitative evaluation of the influenza outbreak forecasting performance for spatial sub-regions within a state, an epidemic's diffusion across a real world disease contact network is illustrated and discussed. Not only can the predicted distributions of age, gender, infection distance, and flu duration be shown and analyzed, but individual-level influenza epidemic simulation results can also be provided. Finally, the spatial and temporal mining of influenza vaccination based on social media streams are considered (Fig. 8).

7.3.1 Influenza epidemics outbreaks forecasting performance for inner-state subregions

The models used to leverage individual-based network epidemiology techniques can be used to generate individual-level information even when the surveillance data is highly spatially coarse-grained. The spatial distribution of the individual outbreaks for a predicted influenza epidemic is shown in Fig. 9, where each green point denotes an infected person. To evaluate the accuracy of it, the counts of infected person first were aggregated into the five DEMHS[3] regions shown in Fig. 8, divided by the respective population bases of these regions, and

[3] DEMHS regions are defined by the Division of Emergency Management and Homeland Security.

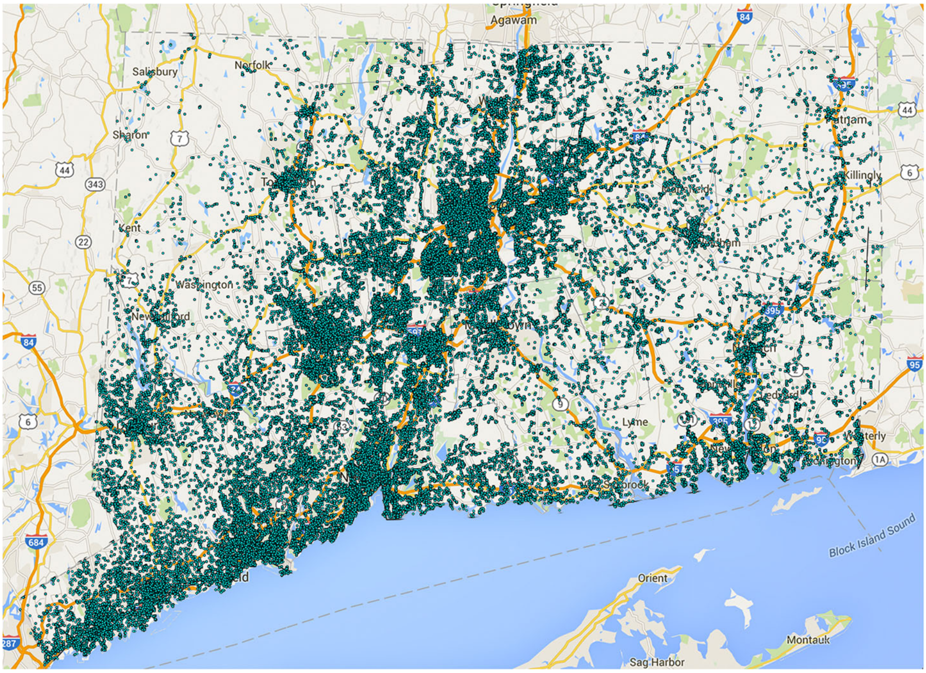


Fig. 9 Individual-level influenza infections simulated by SimNest based on demographics of Connecticut

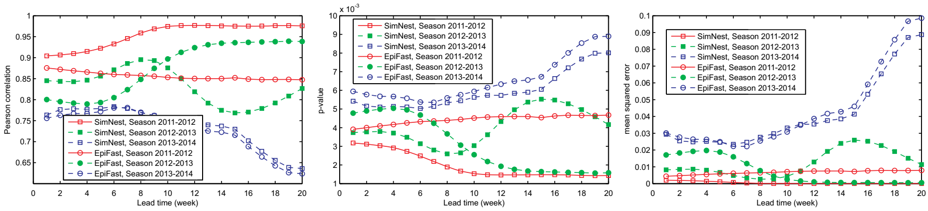
then validated against the corresponding infected ratios reported by the Department of Public Health of Connecticut.[4] Figure 10 depicts the average forecasting performance of the infected ratios for these DEMHS regions.

According to Fig. 10a and b, the SimNest model outperformed EpiFast in the Pearson correlation for Season 2011-2012, Season 2013-2014, and half of Season 2012-2013. The p-values for both methods were less than 0.01 for all three seasons, showing a statistically significant result for their Pearson correlations. Finally, our new SimNest model again outperformed EpiFast in terms of MSE for Season 2011-2012, Season 2013-2014, and half of Season 2012-2013. This is because SimNest is utilizing the social media as a source of individual-level surveillance data, effectively allowing it to monitor people's flu infection status in real time.

7.3.2 Qualitative evaluation of epidemics diffusion across disease contact networks

SimNest can not only predict influenza outbreaks in fine-grained spatial subregions, but also model the epidemic's diffusion, along with the types of diffusion paths. Figure 11 shows the epidemic's spatial diffusion as predicted by SimNest within a subregion of Connecticut that contains two major towns, Torrington and Winchester. The red nodes denote the infectors while the black nodes denote the infectees. The position of these nodes represent their residence locations. The lines between the infectors and infectees denote the "infection" relationship between them; the lines' colors represent the different types of infection routes.

[4] Influenza report for Connecticut: http://www.ct.gov/dph/cwp/view.asp?a=3136&q=410788. Accessed Apr 2016.



(b) Pearson Correlation v.s. Lead Time (a) P-value v.s. Lead Time (c) MSE v.s. Lead Time

Fig. 10 Forecasting performance for 5 DEMHS regions in Connecticut for three flu seasons

For example, the majority of the lines consist of “school” , “home”, and “work”, while the minority of the lines fall under the “shop” type. The lengths of the lines under the types of “school” and “work” can extend several miles, which indicates that people who do not live in close proximity to one another or attend the same school can still infect each other. The typical lengths of the “shop” lines are noticeably shorter. This could be because people prefer to use shops that are not too far away from their residences. The “home”-type infections typically happen within the residence, hence there is almost no distance between the infectors and infectees; a green circle is therefore utilized to signify an infector and an infectee within the same location. Interestingly, those residences with multiple infection lines to other locations typically also have a “home” infection (denoted as green circles). This is because the influenza is often spread across different family members who are involved in

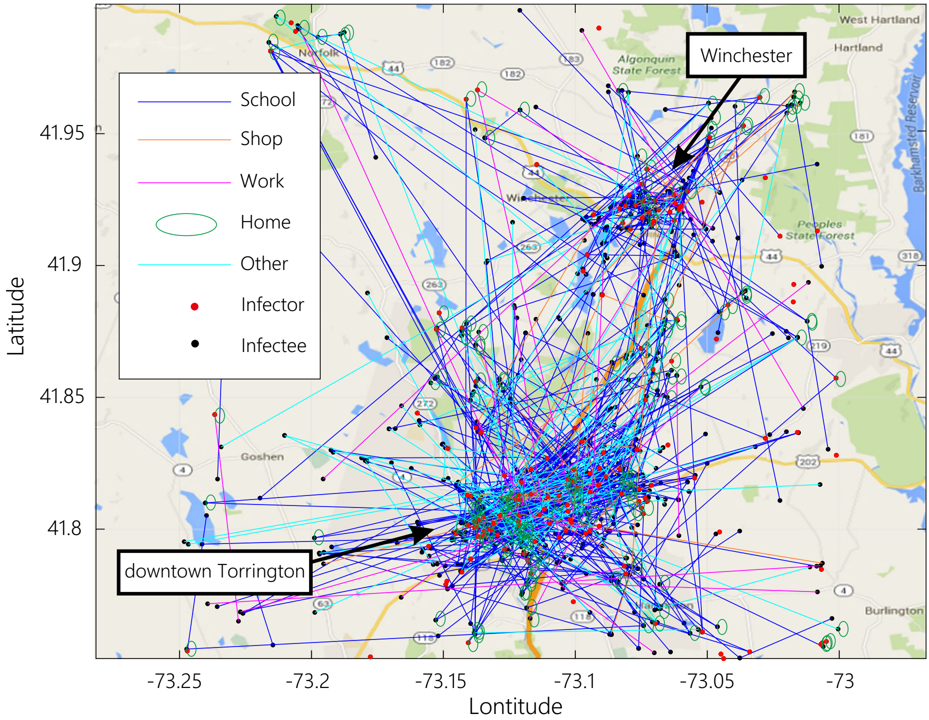


Fig. 11 Epidemic diffusion across a disease contact network

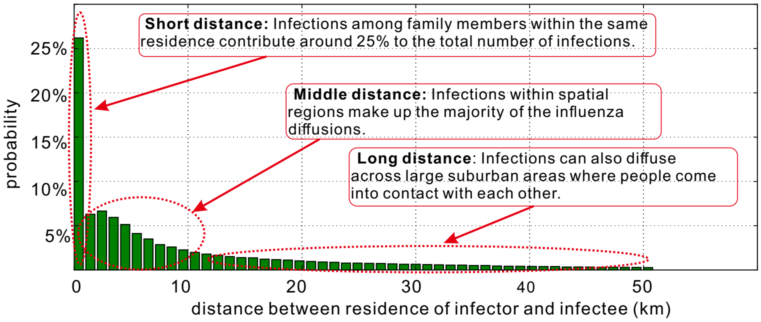


Fig. 12 Distribution of distance between infectors for the simulated influenza population

the infection process within their respective workplaces or schools in the downtowns. The infection type of “other” includes all the infection ways other than the above four, such as disease contacts that occur when eating, entertaining, and traveling.

Figure 12 shows the statistics for the residence distance between infectors and infectees during the influenza diffusion process depicted in Fig. 11. The data in both figures demonstrate that infections that occur within the same residence represents one of the major pathways by which influenza infections spread.

7.3.3 Qualitative evaluation of epidemic modeling in the demographics

SimNest leverages the demographics model and mines the influenza infectious status of individuals from social media data. This capability enables SimNest to forecast not only

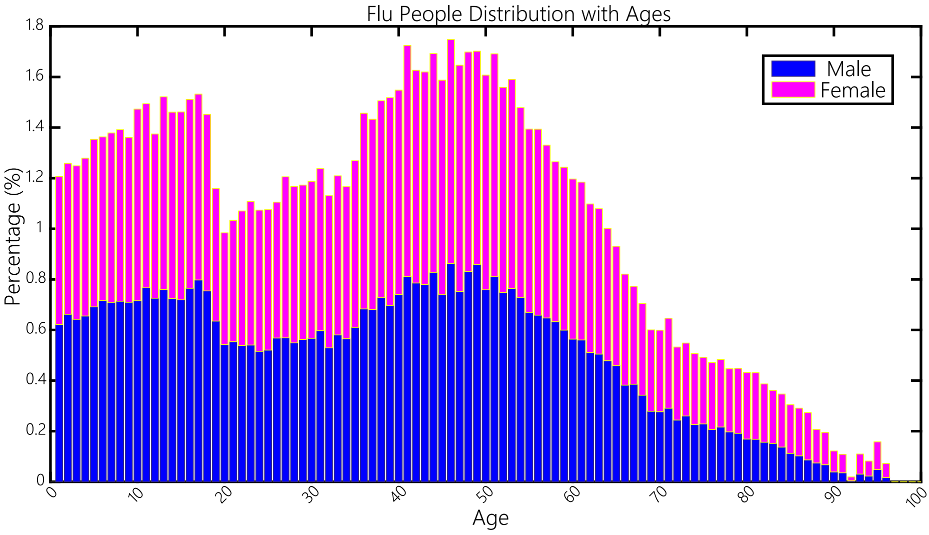


Fig. 13 Age and gender distribution of infectors in the simulated influenza population

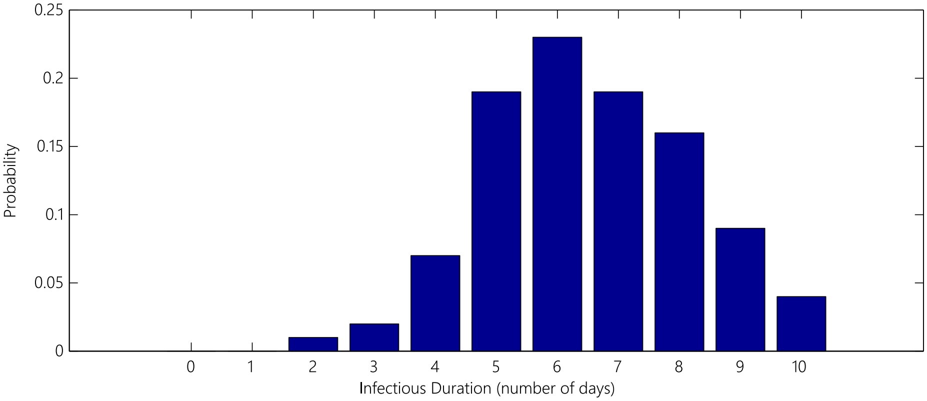


Fig. 14 Distribution of flu duration of the simulated influenza population

the age and gender distribution of the infected population, but also monitor the infectious duration in real time.

As Fig. 13 shows, the majority of influenza infectors consist of 1) children and adolescents (< 20 years old) and 2) middle-aged and older people (> 35 years old). The percentages of male and female infectors are roughly the same.

Figure 14 illustrates the distribution of the influenza infectious duration predicted by SimNest. Influenza infectious duration is the time span from the initial date of infection to the date of recovery (or death). As shown in Fig. 14, the mean infectious duration is around

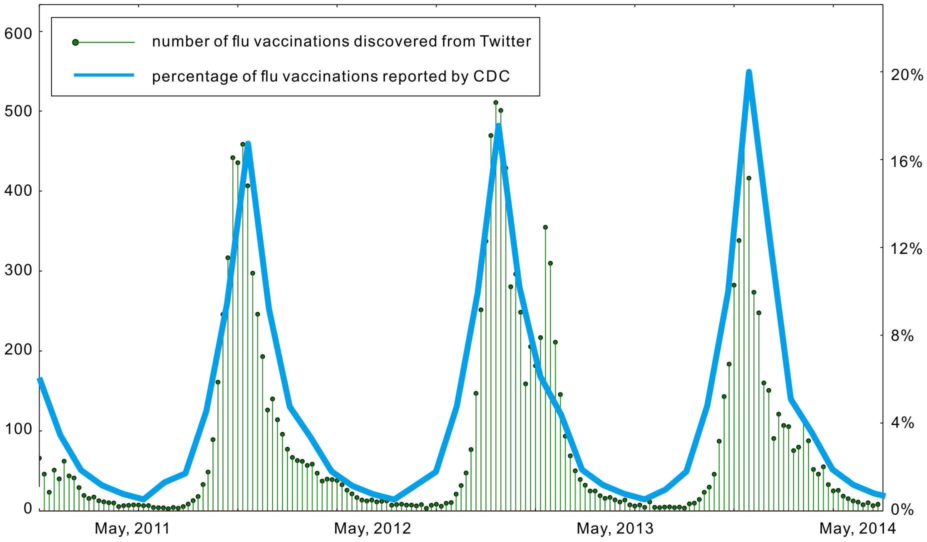


Fig. 15 Flu vaccination temporal patterns detected by SimNest

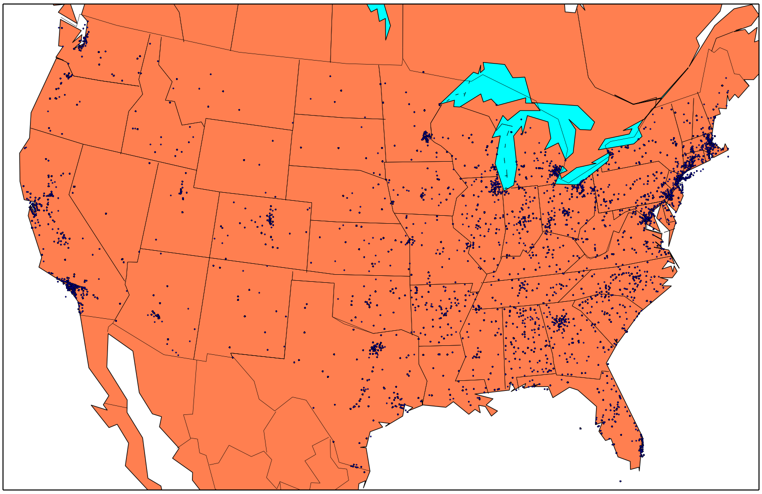


Fig. 16 Spatial patterns for flu vaccination detected by SimNest

6 days. The majority of the cases last between 5 and 7 days, which matches the CDC's Flu Symptoms & Severity report,[5] which was updated in August 2015.

7.3.4 Analysis of disease vaccination surveillance from social media

Based on the online text classification in social media streams, SimNest can identify the postings whose authors have just been vaccinated as they happen. The identified postings are then leveraged to estimate the spatiotemporal flu vaccination rate in near real time. The temporal and spatial patterns for the flu vaccination in the US are shown in Figs. 15 and 16, respectively.

As the green bars in Fig. 15 reveal, the temporal pattern of flu vaccination identified from social media follows a yearly periodicity, with people typically being vaccinated in the months from September to December in each year. These patterns were verified by comparing the results with the official reports on flu vaccination coverage provided by CDC, shown by the blue line in Fig. 15. Note that these CDC reports only become available at the end of each flu season, while the flu vaccination coverage identified from social media is in real time.

SimNest not only discovers the temporal patterns for flu vaccination, but also its spatiotemporal distribution in real time by leveraging the location of the social media postings. Figure 16 illustrates the spatial distribution for flu vaccinations in the US during the flu season 2013-2014. As the figure shows, the flu vaccinations are typically concentrated in the larger cities, with more being administered. The volume of flu vaccination in the eastern part of the country than the western part.

[5] CDC Flu Symptoms & Severity: http://www.cdc.gov/flu/professionals/acip/clinical.htm

8 Conclusions

To achieve timely and accurate epidemic diffusion modeling, the computational epidemiology and social media mining communities have achieved important progress in recent years, although both still suffer from a number of different drawbacks. This paper seeks to combine the advantages of both in the new model proposed here, SimNest, which is a novel bispace co-evolving framework that integrates the complementary strengths of computational epidemiology and social media mining. The new model is capable of learning social media users' health states and behaviors in real time using both an MLP classifier and unsupervised pattern constraints based on the underlying disease model and contact network. The knowledge learned from social media can be fed back into the computational epidemic model to improve the efficiency and accuracy of the disease diffusion modeling. By utilizing our new online optimization algorithm, the above interactive learning process iteratively achieves a consistent stage between these two spaces. Extensive experiments based on the data for multiple states in the US and over several flu seasons demonstrated the advantages of integrating the respective strengths of computational epidemiology and social media mining. The detailed geographical subregion outbreak forecasting performance was also improved by using social media that provides individual-level surveillance data. Although this paper focuses on influenza epidemics, SimNest has good potential of being extended and utilized for modeling and predicting other epidemic disease such as Ebola, and we leave this as future work.

Acknowledgments This work was supported by the National Science Foundation grants: 1755850 and 1841520. We are also grateful of the support of computational resources partially supported by NVIDA company. This research was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337. Additionally, this work has been partially supported by DTRA CNIMS Contract HDTRA1-11-D-0016-0001.

References

1. Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B (2011) Predicting flu trends using Twitter data. In: INFOCOM WKSHPS, pp 702–707
2. Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B (2013) Online social networks flu trend tracker: a novel sensory approach to predict flu trends. In: Biomedical engineering systems and technologies. Springer, pp 353–368
3. Anderson RM, May RM (1979) Population biology of infectious diseases part i. Nature 280:361–7
4. Barrett C, Beckman R, Khan M, Kumar V, Marathe M, Stretz P, Dutta T, Lewis B (2009) Generation and analysis of large synthetic social contact networks. In: WSC, pp 1003–1014
5. Barrett C, Bisset K, Eubank S, Feng X, Marathe M (2008) Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In: ICS, pp 1–12
6. Barrett C, Beckman R, Khan M, Anil Kumar V, Marathe M, Stretz PE, Dutta T, Lewis B (2009) Generation and analysis of large synthetic social contact networks. In: Winter simulation conference, pp 1003–1014. Winter simulation conference
7. Bhatele A, Yeom J.-S., Jain N, Kuhlman CJ, Livnat Y, Bisset KR, Kale LV, Marathe MV (2017) Massively parallel simulations of spread of infectious diseases over realistic social networks. In: Proceedings of the 17th IEEE/ACM international symposium on cluster, cloud and grid computing. IEEE Press, pp 689–694
8. Bishop CM et al (2006) Pattern recognition and machine learning, vol 4. Springer, New York
9. Bisset K, Chen J, Feng X, Kumar VSA, Marathe M (2009) Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: ICS, pp 430–439
10. Bisset KR, Chen J, Feng X, Kumar V, Marathe M (2009) Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: ICS. ACM, pp 430–439

11. Brennan S, Sadilek A, Kautz H (2013) Towards understanding global spread of disease from everyday interpersonal interactions. In: IJCAI. AAAI Press, pp 2783–2789
12. Centers for Disease Control and Prevention (CDC) (2015) CDC fluview interactive. Accessed May 31, 2015. http://www.cdc.gov/flu/weekly/fluviewinteractive.htm
13. Chen L, Hossain KT, Butler P, Ramakrishnan N, Prakash BA (2014) Flu gone viral: syndromic surveillance of flu on Twitter using temporal topic models. In: ICDM. IEEE, pp 2783–2789
14. Choisy M, Guégan J-F, Rohani P (2007) Mathematical modeling of infectious diseases dynamics. Encyclopedia of infectious diseases: modern methodologies, pp 379–404
15. Collier N, Son NT, Nguyen NM (2011) Omg u got flu? analysis of shared health messages for biosurveillance. J Biomedical Semantics 2(S-5):S9
16. Craft ME, Volz E, Packer C, Meyers LA (2011) Disease transmission in territorial populations: the small-world network of serengeti lions. J R Soc Interface 8(59):776–786
17. Culotta A (2010) Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the First Workshop on Social Media Analytics. ACM, pp 115–122
18. Dredze M, Paul MJ, Bergsma S, Tran H (2013) Carmen: a Twitter geolocation system with applications to public health. In: AAAI workshop on expanding the boundaries of HIAI. Citeseer, pp 20–24
19. Gao Y, Zhao L (2018) Incomplete label multi-task ordinal regression for spatial event scale forecasting. In: AAAI conference on artificial intelligence
20. Gough K (1977) The estimation of latent and infectious periods. Biometrika 64(3):559–565
21. Groendyke C, Welch D, Hunter DR (2012) A network-based analysis of the 1861 hagelloch measles data. Biometrics 68(3):755–765
22. Hirose H, Wang L (2012) Prediction of infectious disease spread using Twitter: a case of influenza. In: PAAP. IEEE, pp 100–105
23. Krieck M, Dreesman J, Otrusina L, Denecke K (2011) A new age of public health: Identifying disease outbreaks by analyzing tweets. In: Websci
24. Lamb A, Paul MJ, Dredze M (2013) Separating fact from fear: tracking flu infections on Twitter. In: HLT-NAACL, pp 789–795
25. Murray JD (2002) Mathematical biology i: an introduction, vol 17 of interdisciplinary applied mathematics
26. Pan American Health Organization (PAHO) (2015) PAHO interactive. Accessed May 31, 2015. www.paho.org/hq/
27. Paul MJ, Dredze M (2012) A model for mining public health topics from Twitter. Health 11:16–6
28. Presanis AM, De Angelis D, Hagy A, Reed C, Riley S, Cooper BS, Finelli L, Biedrzycki P, Lipsitch M, et al. (2009) The severity of pandemic H1N1 influenza in the united states, from april to July 2009: a Bayesian analysis. PLoS Med 6(12):e1000207
29. Vynnycky E, White RG (2010) An introduction to infectious disease modelling. Oxford University Press, Oxford
30. Wang J, Zhao L (2018) Multi-instance domain adaptation for vaccine adverse event detection. In: Proceedings of the 2018 World Wide Web conference on World Wide Web, pp 97–106. International World Wide Web conferences steering committee
31. Wang J, Zhao L, Ye Y, Zhang Y (2018) Adverse event detection by integrating twitter data and vaers. Journal of Biomedical Semantics 9(1):19
32. World Health Organization (WHO) (2015) WHO ebola data and statistics. Accessed May 29, 2015. http://apps.who.int/gho/data/view.ebola-sitrep.e{bola-summary-latest}
33. World Health Organization (WHO) (2015) WHO influenza (season) fact sheet. Accessed May 15, 2015. http://www.who.int/mediacentre/factsheets/fs211/en/
34. Yu H, Ho C, Juan Y, Lin C (2013) Libshorttext: a library for short-text classification and analysis. Technical report. http://www.csie.ntu.edu.tw/~cjlin/papers/libshorttext.pdf
35. Zhao L, Chen F, Lu C-T, Ramakrishnan N (2015) Spatiotemporal event forecasting in social media. In: SDM, vol 15. SIAM, pp 963–971
36. Zhao L, Chen F, Lu C-T, Ramakrishnan N (2016) Multi-resolution spatial event forecasting in social media. In: 2016 IEEE 16Th international conference on data mining (ICDM). IEEE, pp 689–698
37. Zhao L, Ye J, Chen F, Lu C-T, Ramakrishnan N (2016) Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 2085–2094

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Liang Zhao is an Assistant Professor at Information Science and Technology Department of George Mason University. He received the Ph.D. degree from Virginia Tech, USA. His research interests include natural language processing, text mining, machine learning, and robotics. In recent years, he has worked primarily on applications to social media, civil unrests, and public health informatics.

Jiangzhuo Chen is a Research Scientist in the Network Dynamics and Simulation Science Laboratory of Biocomplexity Institute of Virginia Tech. He received his B.A. in Economics from Nanjing University, M.A. in Economics from Boston College, and Ph.D. in Computer Science from Northeastern University.

Feng Chen received the B.S. degree from Hunan University, Changsha, China, in 2001; the M.S. degree from Beihang University, Beijing, China, in 2004; and the Ph.D. degree from Virginia Tech USA, in 2012, all in computer science. He is an Assistant Professor with the University at Albany, SUNY.

Fang Jin is an assistant professor at Computer Science Department of Texas Tech University. She has a broad interest in data mining, information propagation of social networks, misinformation propagation detection, finance market prediction with social media, and using big data to discover more insights for social good.

Wei Wang is a staff researcher at Microsoft Research. His research interest spans on number of areas in data mining and machine learning: event encoding, anomaly detection and predictive modeling.

Chang-Tien Lu received the M.S. degree in computer science from the Georgia Institute of Technology in 1996 and the Ph.D. degree in computer science from the University of Minnesota in 2001. He is an Associate Professor with the Department of Computer Science, Virginia Tech. His research interests include spatial databases, data mining, geographic information systems, and intelligent transportation systems.

Naren Ramakrishnan received the PhD degree in computer sciences from Purdue University, West Lafayette, IN, in 1997. He is currently the Thomas L. Phillips Professor of Engineering in the Department of Computer Science at Virginia Tech, Blacksburg, VA. His research has been supported by NSF, DHS, NIH, NEH, DARPA, IARPA, ONR, General Motors, HP Labs, NEC Labs, and Advance Auto Parts.