A Spatial-Temporal Method to Detect Global Influenza Epidemics Using Heterogeneous Data Collected from the Internet

Xichuan Zhou *Member, IEEE*, Fan Yang, Yujie Feng, Qin Li, Fang Tang, Shengdong Hu, Zhi Lin, Lei Zhang

***Abstract*—The 2009 influenza pandemic teaches us how fast the influenza virus could spread globally within a short period of time. To address the challenge of timely global influenza surveillance, this paper presents a spatial-temporal method that incorporates heterogeneous data collected from the Internet to detect influenza epidemics in real time. Specifically, the influenza morbidity data, the influenza-related Google query data and news data, and the international air transportation data are integrated in a multivariate hidden Markov model, which is designed to describe the intrinsic temporal-geographical correlation of influenza transmission for surveillance purpose. Respective models are built for 106 countries and regions in the world. Despite that the WHO morbidity data are not always available for most countries, the proposed method achieves 90.26% to 97.10% accuracy on average for real-time detection of global influenza epidemics during the period from January 2005 to December 2015. Moreover, experiment shows that, the proposed method could even predict an influenza epidemic before it occurs with 89.20% accuracy on average. Timely international surveillance results may help the authorities to prevent and control the influenza disease at the early stage of a global influenza pandemic.**

***Index Terms*—Spatial temporal method, multivariate hidden Markov model, Google trends, Internet news, international air transportation, global influenza surveillance**

I. Introduction

THE Global Influenza Surveillance Network (GISN) of the World Health Organization (WHO) is an essential foundation for monitoring influenza pandemics [1]. For now, the GISN is collecting and examining influenza virological data at the global scale, which comprises 136 national influenza centers in 106 countries and regions [2]. The GISN system has been proven to be valuable, but among the covered countries and regions, only 21 countries have national surveillance networks, which allow them to report the surveillance data in a standard form [3]. However, for most countries and regions, the surveillance data are not always available.

To improve the coverage, timeliness and sensitivity of global influenza surveillance, Internet technology has become integral to public health surveillance over the past a few years. Novel syndromic surveillance systems like the Google Flu have been developed to estimate the timing and location of influenza epidemics at the global scale [4]–[9]. The search engine based approaches generally assume that the volume of influenza-related queries is correlated with the actual influenza morbidity trend. However, recent researches indicate that, people's web search behaviors are affected by the reports of events which are not related to local infections of the influenza disease [7]. For example, at the early stage of the 2009 influenza A(H1N1) pandemic, the news reports of the epidemic outbreak in Mexico caused a global panic and a dramatic increase in influenza-related searches in the countries where few infections had occurred. As a result, the search-based surveillance systems might suffer from relatively low reliability and false epidemic alerts [10]–[12].

One way to improve the robustness of syndromic surveillance is to incorporate other influenza-related data sources. Recently, the advent of openly available news aggregators has made it possible to use the Internet news data for influenza surveillance. Since the outbreaks of influenza epidemics are usually reported online, novel systems like the HealthMap use the data extracted from news reports to produce a global view of ongoing infectious disease threats [13]. We found in our previous research that incorporating the news data source might reduce the media effect of the search-based surveillance systems [7]. Therefore, this paper examines the number of Internet news reports containing 'influenza' or 'flu' (in native languages, see Table I) as a complementary data source.

Though the Google Flu model could provide timely global surveillance results using daily-updated query data, research indicated that it suffered from inadequate sensitivity at the early stage of an influenza pandemic [10], e.g. it missed the first wave of the 2009 influenza A(H1N1) pandemic in the United States [11]. Recent research indicated that, the transmission patterns of the influenza virus across different countries were strongly correlated with the volume of international air transportation [14]. By taking advantage of the air-transportation data collected from the International Air Transport Association (IATA) website [15], our method estimated the number of passengers infected with the influenza disease, and used the estimates to improve the sensitivity for detecting potential influenza epidemics.

Since the 2009 influenza pandemic, analyzing the spatial-

X. Zhou is with the Key Laboratory of Dependable Service Computing in Cyber Physical Society Ministry of Education, College of Communication Engineering, Chongqing University, Chongqing, China, 400044. X. Zhou, F. Yang, F. Tang, S. Hu, Z. Lin and L. Zhang are with Chongqing Engineering Laboratory of High Performance Integrated Circuits, College of Communication Engineering, Chongqing University, China, 400044. Y. Feng is with the Southwest Hospital, Third Military Medical University, Chongqing, China, 400030. Q. Li is with the Chongqing Centers for Disease Control and Prevention, China, 400010.(Corresponding author: Fang Tang, e-mail: eefrank@cqu.edu.cn)

This research is partly supported by the national natural science foundation of China (contract 61103212, 61471073, 61604024), and is also supported by the Fundamental Research Funds for the Central Universities Project No. 106112016CDJZR168803.

TABLE I
KEYWORDS USED TO EXTRACT THE QUERY VOLUME DATA AND THE NEWS COUNT DATA [1]

Keywords	Countries and regions	Keywords	Countries and regions
chřipka	Czech Republics, Slovakia	流感	China
cúm	Vietnam	インフルエンザ	Japan
flunssa	Finland		
flu	Albania, Algeria, Armenia, Australia, Azerbaijan, Bangladesh, Britain, Canada, Cambodia, Egypt, Ethiopia, Fiji, Georgia, Ghana, Greece, India, Indonesia, Iran, Iraq, Ireland, Israel, Jamaica, Jordan, Kenya, Malta, Mauritius, Mongolia, Morocco, Nepal, New Zealand, Nigeria, Pakistan, Philippines, Qatar, Singapore, South Africa, Tanzania, Uganda, USA, Uzbekistan, Zambia	gripe	Angola, Argentina, Bolivia, Brazil, Chile, Costa Rica, Cuba, Ecuador, El Salvador, Dominican Republic, Guatemala, Honduras, Mexico, Mozambique, Nicaragua, Panama, Portugal, Spain, Uruguay, Venezuela
		grippe	Austria, Burkina Faso, Cameroon, Cote d'Ivorie, France, Germany, Guadeloupe, Luxembourg, Madagascar, Martinique, Senegal, Switzerland, Tunisia
gripa	Bosnia and Herzegovina, Croatia, Colombia, Latvia, Moldova, Romania, Slovenia	influenza	Denmark, Hungary, Italy
		influensa	Norway, Sweden
griep	Belgium, Netherland	тұмау	Kazakhstan
grip	Turkey	selesema	Malaysia
gripp	Estonia	грипп	Belarus, Russia, Ukraine
grypa	Poland	грип	Bulgaria, Kyrgyzstan
독감	Republic of Korea	ไข้หวัดใหญ่	Thailand

1. All keywords used are translations of influenza or flu in native languages.

temporal risk of the influenza disease for surveillance purpose has become an important goal of statistical and epidemiology researches [16]. Previous spatial-temporal methods, also known as the space-time methods, usually focus on a study area made up of smaller, non-overlapping sub-regions where cases of disease are being monitored. For global influenza surveillance, we examine totally 106 countries and regions for the purpose of detecting influenza epidemics in real time. The key variable under influenza surveillance is the binary state, i.e. one for epidemic and zero for non-epidemic, which is usually estimated using the influenza morbidity data. Statistical tests are generally applied in spatial-temporal surveillance systems to determine whether the disease incidences in a country are unusual compared to the baseline [17]–[19]. Recently, model based approaches have attracted a lot of interest because they can include other variables like syndromic indicators into the surveillance system. These model based methods have roughly three classes: linear regression based models [4], [5], [20], Bayes models [21], [22], and the models of specific space-time processes [23]–[26]. More recently, the hidden Markov models (HMM) are developed to take advantage of advanced computing power and novel data sources [7], [27], [28]. The hidden Markov model with time-correlated states can capture the temporal correlation of influenza transmission, which makes it an effective tool for detecting epidemics in a city or a country [26]; however, besides the temporal correlation, in a global pandemic, the transmission of the influenza virus is also geographically correlated among different countries and regions, which requires extra variables to incorporate the spatial correlation into the model for the purpose of influenza surveillance at the global scale.

To address the challenge of real-time epidemic detection at the global scale [1], this paper presents a multivariate hidden Markov model (MHMM) to estimate the timing and location of influenza epidemics using heterogeneous data collected from the Internet. Inspired by Khan's observation [14], we assume that the present epidemic risk state of the target country is not only related to earlier epidemic states, but it is also related to the epidemic risks of other countries which are connected with the target country by international airlines. In this paper, a discrete-time Markov chain of bivariate latent states, i.e. the national-epidemic-risk states and the imported-epidemic-risk states, is built for each country and region. At each week, we assume that the present numbers of search queries and news reports containing 'influenza' or 'flu' in native languages are correlated with the current epidemic state; therefore, one could estimate the present epidemic state using these daily-updated Internet data. On the other hand, the international air transportation data are used to describe the geographical correlation of disease transmission. By taking advantage of the intrinsic spatial-temporal correlation of influenza transmission learned by the multivariate Markov process, the proposed method could sensitively detect a potential epidemic even before it occurs.

Compared with the famous Google Flu model, our method shows higher surveillance accuracy and sensitivity mainly for three reasons. (1) Instead of assessing the search data directly, our method considers the temporal correlation of disease transmission, which makes it more robust against the irrelevant searches; (2) The proposed method also considers the spatial correlation of disease transmission and incorporates the international air transportation data, which increases the sensitivity to detect an epidemic at the early stage. Furthermore, for countries lacking the virological surveillance data, the epidemic information of their airline-connected countries could provide valuable indicators of potential epidemics. (3) By incorporating the news data source, our method may relieve the media effect of the search based surveillance systems.

The rest of this paper is organized as follows. Section II briefly introduces the data used to build and evaluate the model. Section III describes the multivariate hidden Markov model and the surveillance method. Section IV presents the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2690631, IEEE/ACM Transactions on Computational Biology and Bioinformatics

experiment results. Further discussion and conclusion are given in Section V.

II. DATA SOURCES

A. Influenza Morbidity Data

The influenza morbidity data of 106 countries and regions were gathered from the WHO FluNet [3]. The morbidity data were of weekly resolution and last from January 2004 to December 2015. Fig. 1 shows an example of the influenza morbidity data for the country of Belarus, which are normalized between zero and one hundred. It is worth noting that, the WHO influenza morbidity data are not always available for most countries. As shown in Fig. 1, the influenza morbidity data were missing for Belarus during the period from the 15th week of 2009 to the 39th week of 2010. At the global scale, the average missing rate of the WHO morbidity data was over 35% for the examined 106 countries and regions. The missing rates for Asian and African countries, where most influenza epidemics were reported, were significantly higher than average, making the influenza surveillance more challenging.

For the purpose of detecting influenza epidemics, the WHO morbidity data were used to calculate the ground-truth indicators of epidemics. As previous epidemiology studies [17]–[19] did, the national epidemic risk state of each week was determined by comparing the influenza morbidity with a baseline. Fig. 1 shows an example of the national epidemic risk states for Belarus (subgraph a). Since the morbidity data were missing from time to time, the unknown national epidemic risk states were treated as hidden states in our research. Detailed method to estimate and predict the national epidemic risk states are described in the next section.

B. Air Transportation Data

The volume data of international flight itineraries for all passengers arriving at commercial airports between January 2004 and December 2015 were gathered from the International Air Transport Association (IATA) [15]. The IATA data account for more than 95% of all passenger trips worldwide via commercial airlines, they include information on the flight origins and destinations of actual passengers.

For each examined country, we aggregated the arrival flights and the number of passengers on a weekly basis. By multiplying the number of passengers and the morbidity rate of the departure countries, we estimated the weekly number of arrival passengers infected with influenza. Fig. 1 shows the estimated number of infected passengers to Belarus by airplane, which is normalized between zero and one hundred (imported infections, subgraph b). The estimates of imported infections reflect the risk that the influenza virus may transmit from other countries to the target country. Therefore, a supplementary epidemic indicator, i.e. the imported epidemic risk state, is defined to describe the geographical correlation of influenza transmission between different countries and regions. As an example, the high-lighted area in Fig 1.b shows the imported epidemic risk states calculated by comparing the estimated number of infected passengers with the baseline.

Our research assumes that the national-epidemic-risk states are correlated with the imported-epidemic-risk states. As shown in Fig. 1, the national-epidemic-risk states (subgraph a) and the imported-epidemic-risk states (subgraph b) show similar yearly pattern for Belarus, yet they may differentiate from time to time. Our method manages to improve the sensitivity and accuracy of the estimation of national-epidemic-risk states by considering the geographical correlation of disease transmission and incorporating the second variable of imported epidemic risk states in a multivariate Markov model.

C. Google Trend Data

The Google trend data of the examined countries and regions were gathered from publicly-accessible Google Trend website [29]. The daily-updated search volume of the query 'influenza' or 'flu' in native languages (Table 1) were collected from January 2004 to December 2015. The search volume was calculated by aggregating the searches submitted in a country or a region. For each query, the volume data were normalized from zero to one hundred (Fig. 1).

To reduce the noise, the search trend data were quantized as three-level data $G_t \in \{1, 2, 3\}$. Specifically, we assumed that the search volume belonged to the Gaussian distribution, and each of the three levels shared the same probability of 1/3. After fitting the Gaussian model using the search volume data, one could transform the query-volume data to the query-level data G_t, which were used to estimate contemporary national epidemic risk states.

D. Internet News Trend

Recently, the Google Trends service began to provide the news trend data associated with different keywords. Google news database is generally built using documents collected by robot programs. It selects the important and trustworthy websites as news sources, including national and local news websites, and networks for government health department, health care organizations, and traditional media companies. Given a keyword and a monitored country or region, the Google Trends service provides the count number of news documents containing the keyword. For the purpose of surveillance, the count number of news containing 'flu' or 'influenza' in native languages were examined (see Table I).

Similar to the search volume data, we also quantized the news count data with three levels of equal probability. After fitting the Gaussian model using the news count data, one could calculate the threshold to transform the news count data to the news-level data $N_t \in \{1, 2, 3\}$, which were also used to estimate contemporary national epidemic risk states.

III. METHOD

One challenge for global influenza surveillance is to catch up with the virus which disseminates worldwide quickly via international airlines in the early stage of a pandemic [30]. Since the transmission of the virus is geographically and temporally correlated, we build a multivariate discrete-state hidden Markov model (MHMM) to estimate and even predict the national epidemic risk states in real time.

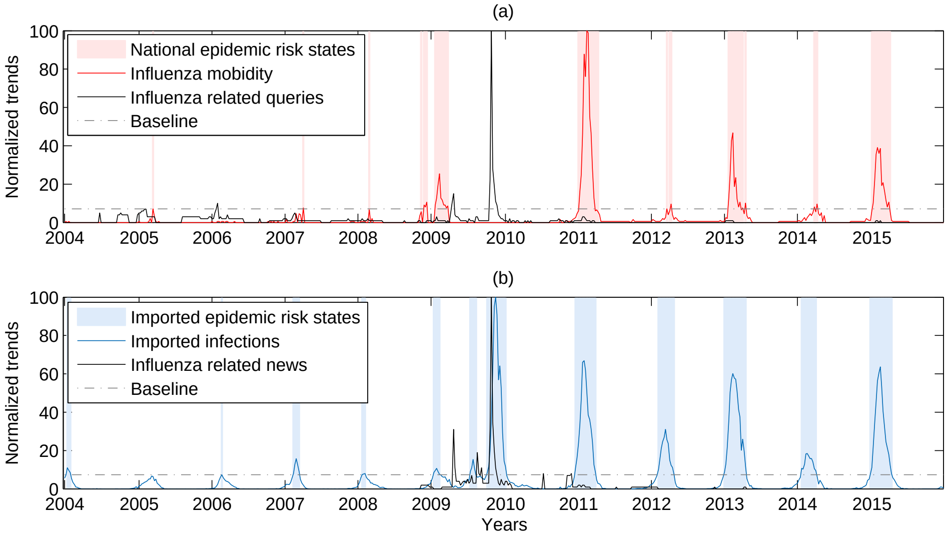


Fig. 1. Data examined for monitoring influenza epidemics in Belarus during the period from year 2004 to 2015. Four types of heterogeneous data are examined, including the influenza morbidity data, the Google search trend of ‘flu’ (in Russian), the number of news containing ‘flu’ (in Russian) and the international air transportation data. The highlighted area in subgraph (a) indicates the ground truth of national-epidemic-risk states calculated based on the morbidity data. The highlighted area in subgraph (b) indicates the imported-epidemic-risk states calculated based on the estimated volume of arrival aeroplane passengers infected with influenza. The goal of this paper is to predict the national epidemic risk states whose ground truth are not always available for most countries. As an example, the influenza morbidity data for Belarus are mostly missing for year 2009 and 2010.

A. *Influenza Epidemic Risk States*

At the *t*th week, the proposed MHMM model has two correlated latent variables, i.e. the national epidemic risk state $S_{t,1}$ and the imported epidemic risk state $S_{t,2}$. Ideally, the ground-truth value of $S_{t,1}$ should be determined by contemporary morbidity data m_t. Suppose the influenza epidemics occur with a relatively small probability σ in a country or region, then a baseline b_1 can be calculated to determine the ground-truth value of national epidemic risk states as

$$S_{t,1} = \begin{cases} 0 & \text{when } m_t \leq b_1 \\ 1 & \text{when } m_t > b_1 \end{cases}$$

where zero stands for non-epidemic and one stands for epidemic. Since the morbidity data published by the WHO are often missing, the unknown national epidemic risk state $S_{t,1}$ is treated as a hidden variable in the MHMM model, which attempts to estimate and predict the weekly-updated $S_{t,1}$ using data collected from the Internet.

To take advantage of the geographical correlation of influenza transmission, a second variable $S_{t,2}$ is defined based on the number of arrival aeroplane passengers at the tth week. Intuitively, the latent variable $S_{t,2}$, i.e. the imported epidemic risk state, reflects the risk that an influenza epidemic could transmit to the target country via infected aeroplane passengers. Given a target country, suppose c_{it} and m_{it} are the passenger number and the influenza morbidity of its ith origin country. Suppose the ith origin country's population is p_i, and γ_{it} is the morbidity rate. Then the number of arrival passengers infected with influenza can be estimated as

$$d_t = \sum\nolimits_i c_{it}\gamma_{it} \text{ and } \gamma_{it} = \frac{m_{it}}{p_i}$$

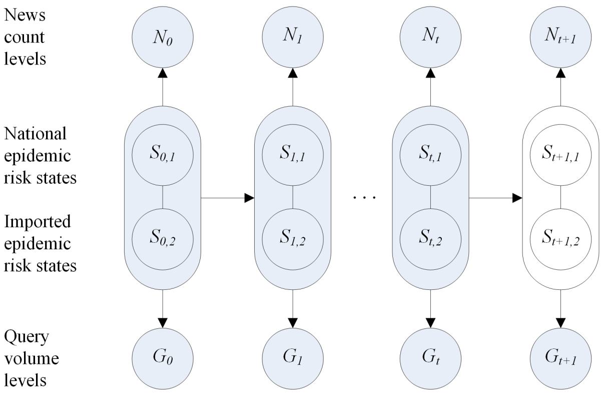


Fig. 2. The multivariate hidden Markov model built for surveillance purpose. The latent variable $S_{t,1}$ is the national epidemic risk state defined to detect epidemics, and the $S_{t,2}$ is the imported epidemic risk state which reflects the risk to import an epidemic from airline-connected countries. At the week $t+1$, the proposed model calculates the real-time epidemic indicator $S_{t+1,1}$ based on the query volume level G_{t+1}, the news count level N_{t+1} and the previous epidemic risk states ($S_{t,1}$, $S_{t,2}$).

Similar to the definition of $S_{t,1}$, we assume that an imported epidemic occurs with a small probability σ, then the baseline value b_2 can be calculated for each country to determine the ground-truth value of $S_{t,2}$ as

$$S_{t,2} = \begin{cases} 0 & \text{when } d_t \leq b_2 \\ 1 & \text{when } d_t > b_2 \end{cases}$$

Since the definition of $S_{t,2}$ depends on the morbidity data which are missing from time to time, it is also treated as a hidden variable in the MHMM model.

In our research, the latent national epidemic risk state $S_{t,1}$ is defined as the indicator of the epidemic occurs at the tth week.

To incorporate the geographical correlation in our model, we assume that the epidemic indicator $S_{t,1}$ is correlated with the imported epidemic risk variable $S_{t,2}$. This assumption is coherent with the finding that, the actual influenza morbidity m_t and the number of imported infections d_t are generally correlated, with positive average correlation among the examined 106 countries and regions (Pearson correlation coefficient r = 0.41).

B. Multivariate Hidden Markov Model

Fig. 2 illustrates the spatial-temporal MHMM model built for influenza surveillance, where the connection between the two variables $S_{t,1}$ and $S_{t,2}$ reflects the geographically correlation between airline-connected countries, and the connection between the epidemic risk states of continuous time steps reflects the temporal correlation of disease transmission. Formally, in the multivariate HMM process, the temporal evolution of epidemic risk states are driven by a latent Markov chain, which can be conveniently described as a multinomial process in discrete time. Accordingly, we introduce a sequence $\mathbf{S}_{0:T} = (\mathbf{S}_t,\ t = 0, ..., T)$ of multinomial variables $\mathbf{S}_t = (S_{t,1}$, $S_{t,2})$, whose binary components are the national epidemic risk state and the imported epidemic risk state. The state entered at each step depends on two initial probabilities and four state transition probabilities as $\pi_1 = P(S_{0,1} = 1), \pi_2 = P(S_{0,2} = 1)$ and $\pi_{h,k} = P(S_{t,k} = 1|S_{t-1,h} = 1), h, k = 1, 2$. Similar to the standard Markov model, we assume the sequence of epidemic risk states $\mathbf{S}_{0:T}$ occurs with a probability

$$p(\mathbf{S}_{0:T}; \boldsymbol{\pi}) = \pi_1^{S_{0,1}} \pi_2^{S_{0,2}}$$

$$* \prod\nolimits_{t=1}^{T} (\pi_{1,1}^{S_{t-1,1}S_{t,1}} \pi_{1,2}^{S_{t-1,1}S_{t,2}} \pi_{2,1}^{S_{t-1,2}S_{t,1}} \pi_{2,2}^{S_{t-1,2}S_{t,2}})$$

The standard MHMM usually assumes that the multivariate variables are independent; however, in our research, the national epidemic risk of a country is correlated with the epidemic risks of surrounding countries. As a second difference from the standard MHMM model, our method does not assume constant initial probabilities for the risk states, instead, the values of π_1, π_2, $\pi_{n,k}$ are estimated using the actual epidemic states calculated based on the WHO morbidity data. The proposed method can do that because, different from the standard MHMM method, the hidden states of the epidemic risks are partially available in our application.

As a challenge for global influenza surveillance, for most countries, the morbidity-based epidemic risk states may not always be available; however, the influenza-related Google query-volume level G_t and news-count level N_t can be obtained as syndromic indictors. Suppose $\alpha_k^{S_{t,h}}$ and $\beta_k^{S_{t,h}}$ are the conditional probability of $G_t = k$ and $N_t = k$ respectively, where $k \in \{1,\ 2,\ 3\}$ is the level of syndromic indicator in week t. As in the standard hidden Markov model, the specification of the MHMM is completed by assuming that, the conditional distribution of the syndromic indicator process, given the sequence of latent epidemic states, takes the form of a product density as

$$p(\mathbf{G}_{0:T}|\mathbf{S}_{0:T}, \boldsymbol{\alpha}) = \prod\nolimits_{t=0}^{T} \prod\nolimits_{k=1}^{3} \prod\nolimits_{h=1}^{2} \alpha_k^{S_{t,h}}$$

$$p(\mathbf{N}_{0:T}|\mathbf{S}_{0:T}, \boldsymbol{\beta}) = \prod\nolimits_{t=0}^{T} \prod\nolimits_{k=1}^{3} \prod\nolimits_{h=1}^{2} \beta_k^{S_{t,h}}$$

C. Parameter Estimation for Real-time Epidemic Detection

The proposed method is designed for real-time detection of influenza epidemics using heterogeneous data collected from the Internet. To overcome the GISN's reporting lag, our spatial-temporal surveillance is performed in an online fashion. Specifically, at the tth week, we use the morbidity data, the query volume levels, the news count levels and the international air transportation data published before the $(t-1)$th week as training data. The training process updates the model parameter set $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ by maximizing the following likelihood function as

$$\arg\max_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}|\mathbf{S}_{0:t-1}, \mathbf{G}_{0:t-1}, \mathbf{N}_{0:t-1}) =$$

$$\arg\max_{\boldsymbol{\pi},\boldsymbol{\alpha},\boldsymbol{\beta}} p(\mathbf{S}_{0:t-1}; \boldsymbol{\pi}) p(\mathbf{G}_{0:t-1}|\mathbf{S}_{0:t-1}, \boldsymbol{\alpha}) p(\mathbf{N}_{0:t-1}|\mathbf{S}_{0:t-1}, \boldsymbol{\beta})$$

After updating the parameters, we can use the week-t query data and news data to estimate the current national epidemic risk state $S_{t,1}$, which is achieved by choosing the risk state with the maximum probability to generate the current G_t and N_t as

$$\arg\max_{S_{t,1}} p(S_{t,1}|S_{t-1,1}, S_{t-1,2}, G_t, N_t, \boldsymbol{\lambda})$$

$$= \arg\max_{S_{t,1}} \pi_{1,1}^{S_{t-1,1},S_{t,1}} \pi_{1,2}^{S_{t-1,1}S_{t,2}} \pi_{2,1}^{S_{t-1,2}S_{t,1}} \pi_{22}^{S_{t-1,2}S_{t,2}}$$

$$* \prod\nolimits_{h=1}^{2} \alpha_{G_t}^{S_{t,h}} * \prod\nolimits_{h=1}^{2} \beta_{N_t}^{S_{t,h}}$$

where $S_{t,1}, S_{t,2} \in \{0, 1\}$ is the feasible set of the epidemic risk states. Since the Google search volume and the news count data are updated on a daily basis, the proposed method can provide timely estimates of national influenza epidemic states, which is generally one week ahead of the GISN's reports.

Furthermore, since the Markov process could capture the temporal relation of influenza transmission, it could predict future epidemic risk states. Specifically, at tth week, one could predict the national epidemic risk state $S_{t+1,1}$ by using the estimated risk states of the tth week ($S^*_{t,1}$, $S^*_{t,2}$) as ground-truth, and calculate the most probable epidemic risk state of the $(t+1)$th week as

$$\arg\max_{S_{t+1,1}} p(S_{t+1,1}|S_{t,1}, S_{t,2}, \boldsymbol{\lambda})$$

$$= \arg\max_{S_{t+1,1}} \pi_{1,1}^{S^*_{t,1}S_{t+1,1}} \pi_{1,2}^{S^*_{t,1}S_{t+1,2}} \pi_{2,1}^{S^*_{t,2}S_{t+1,1}} \pi_{2,2}^{S^*_{t,2}S_{t+1,2}}$$

D. Predicting Unmonitored Epidemics

One important challenge of global influenza surveillance is the temporal gaps in the surveillance data [31]. In fact, among all the examined countries and regions, the average missing rate of the WHO morbidity data is over 35% for the examined period of eleven years. To address this challenge, our method takes advantage of the Internet syndromic data to provide timely indicators for unmonitored epidemics.

Suppose the morbidity data of the target country are not available from the tth week to the $(t+n)$th week. Given

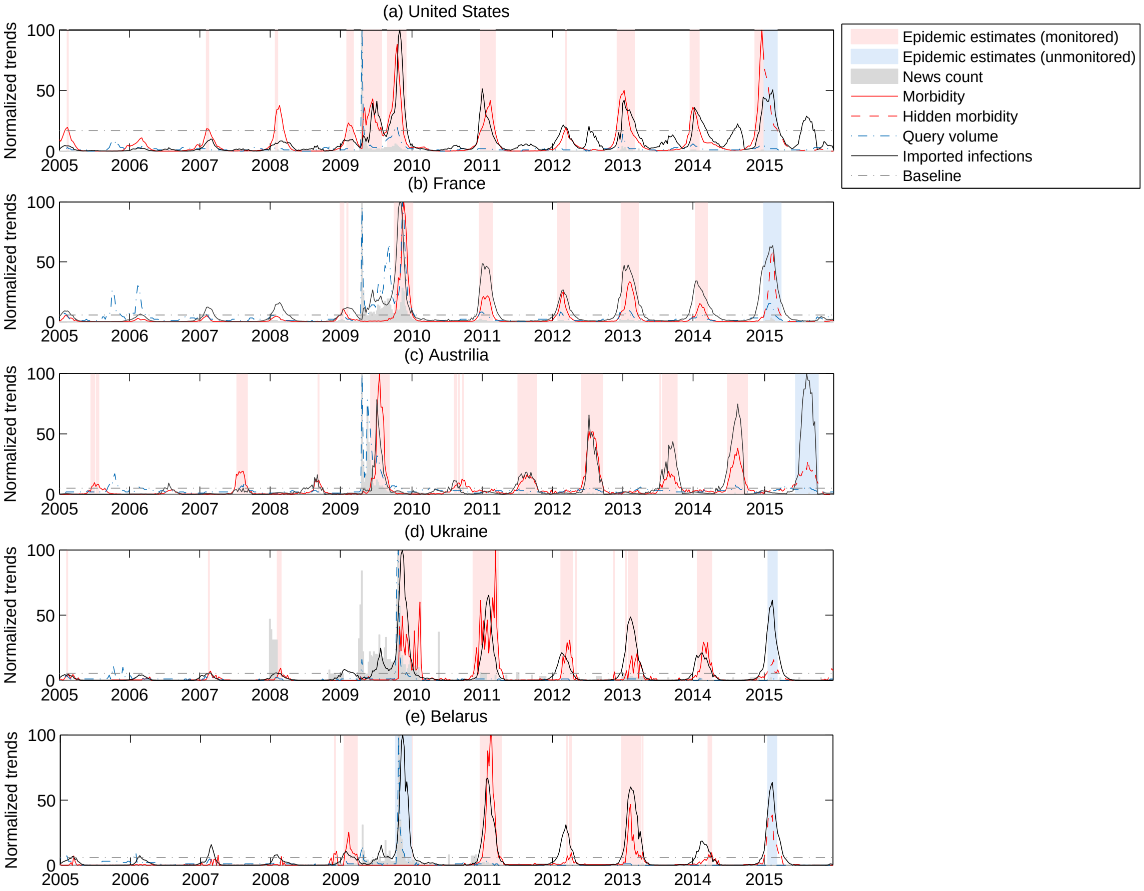


Fig. 3. Surveillance results of the proposed method for five representative countries. The whole examined period has two stages, i.e. the real-time surveillance stage (year 2005 to 2014) and the blind surveillance stage (year 2015). Note that the unmonitored epidemic in Belarus around 2010 is detected by the MHMM model due to large volume of imported infections from Ukraine.

contemporary sequence of query volume levels $\mathbf{G}_{t:t+n}$, the sequence of news count levels $\mathbf{N}_{t:t+n}$, and the sequence of imported epidemic risk states $\mathbf{S}_{t:t+n,2}$ which is calculated based on the existing morbidity data of airline-connected countries, then the Markov model could predict the latent sequence of national epidemic risk states $\mathbf{S}_{t:t+n,1} = (S_{t,1}, S_{t+1,1}, \ldots, S_{t+n,1})$ by maximizing the probability

$$\arg\max_{\mathbf{S}_{t:t+n,1}} p(\mathbf{S}_{t:t+n,1}, \mathbf{S}_{t:t+n,2}, \mathbf{G}_{t:t+n}, \mathbf{N}_{t:t+n}; \boldsymbol{\lambda})$$

which approximates the most probable sequence of national epidemic risk states that leads to the observed sequence of influenza-related query volume levels and news count levels. In practice, the above maximization can be computed efficiently using the Viterbi algorithm, which is a widely adopted algorithm for approximating latent states in the hidden Markov model [32].

IV. EXPERIMENTS

This section evaluates the effectiveness of the proposed spatial-temporal method for detecting influenza epidemics at the global scale in real time. Respective models were built for each examined country and region. The examined period were divided into two stages, i.e. (1) the real-time surveillance stage and (2) the blind surveillance stage. The real-time surveillance stage covered the period from January 2005 to December 2014 using the multivariate Hidden Markov model whose parameters were incrementally updated on a weekly basis. The blind surveillance stage was designed to evaluate the proposed method for its ability to detect influenza epidemics in unmonitored period. The rest of this section gives more detailed information of the experiment in each stage.

A. Real-time Epidemic Surveillance Results

We first examined the MHMM method for real-time detection of influenza epidemics at the global scale. Generally speaking, despite that disease transmission was affected by many factors, the multivariate hidden Markov model captured the spatial-temporal pattern of influenza epidemics in all examined 106 countries. The average accuracy of the real-time estimates of national-epidemic-risk states was remarkably high (93.34 %) compared with the ground truth. The yearly average accuracies over the examined countries ranged from 92.20% (2013) to 97.50% (2006). Moreover, the estimation accuracy was consistently high for both pandemic (year 2009, 94.79%) and non-pandemic (94.41%) periods. Since the WHO surveillance data are based on virological testing, its surveillance reports usually have one-week lag, while the Google query data and the news count data are updated on a daily

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2690631, IEEE/ACM Transactions on Computational Biology and Bioinformatics

basis, which enables us to detect national influenza epidemics of different countries in nearly real time.

Fig. 3 shows the real-time estimates of the national epidemic risk states for five representative countries. For each country, the morbidity data published by the WHO is normalized between zero and one hundred, and a baseline line is calculated to threshold the epidemic period with the largest 15% morbidity. As one can see, the volume of the examined search queries has multiple peaks which are irrelevant to local influenza morbidity. For example, there was a peak in the amount of searches for ‘grippe’ (meaning influenza) in France early in 2006 when official media reported the discovery of multiple dead birds infected with H5N1 virus [33]. On the other hand, the heavy media reports of the influenza A(H1N1) epidemic in Mexico caused dramatic increase in the amount of influenza-related searches at the early stage of the pandemic in 2009, which were irrelevant to contemporary morbidity for most countries. These irrelevant searches may result in false alerts for the Google Flu method; however, thanks to the heterogeneous data sources incorporated, the MHMM method can provide accurate surveillance results which are robust against irrelevant peaks in search volume data.

Though the estimated number of aeroplane-imported infections are strongly correlated with the actual influenza morbidity for many countries like France (Pearson correlation coefficient $r = 0.75$) and Australia ($r = 0.71$), the airline data also suffer from irrelevant peaks. For example, multiple peaks of imported infections could be observed in the United States in 2012 and 2014, which were not correlated with the actual morbidity data (Fig. 3). It seems that, to improve the robustness of syndromic surveillance, it is preferable to incorporate multiple data sources.

Besides real-time detection of influenza epidemics, at each week, the multivariate hidden Markov model could also be used to predict the national epidemic risk state of the next week. After updating the parameters using the latest morbidity data published by the WHO at each week, the MHMM achieved over 89.10% average accuracy among all the examined countries for predicting the next-week epidemics in the period from 2005 to 2014. More detailed results can be found in Table II.

The real-time surveillance results of each continent were also calculated, and the average accuracies ranged from 91.80% (Africa) to 93.90% (Oceania). It seemed that, the accuracy of the real-time surveillance results were affected by the missing rate of the morbidity data. Specifically, we found positive correlation existed ($r = 0.36$) between the missing rate and the error rate of epidemic detection among the examined countries. Fortunately, despite the high missing rate of many developing countries, the lowest accuracy achieved by the multivariate hidden Markov model is still relatively high for real-time epidemic detection.

B. Detecting Unmonitored Epidemics

The Global Influenza Surveillance Network has been proven essential for preventing a global influenza pandemic; however, until the present day, many developing countries still lack the resources to monitor the influenza morbidity constantly. For example, as illustrated in Fig. 3, there was a server influenza epidemic in Belarus later in 2009 [34]; however, no official surveillance data were available at that time. The proposed method detected the influenza epidemic by the significant increase in the number of imported infections via international air transportation.

To fully evaluate the proposed method for its ability to detect unmonitored influenza epidemics, in the second stage, we trained respective models for different countries and regions using the data collected before the 52th week of year 2014. By presuming that the morbidity data of the target country were not available in 2015, we frozen the model parameters and estimated the latent national-epidemic-risk states at each week in 2015. By comparing with the ground-truth epidemic states, one could calculate the average accuracy for blind prediction of influenza epidemics among different countries. Experiment showed that, even for the extreme case of missing the morbidity data for a whole year, the multivariate hidden Markov model could still provide early alarms of influenza epidemics with relatively high accuracy (89.20%).

C. Comparing with other Methods

To further evaluate the effectiveness of the proposed method, we compared it with the widely adopted statistical thresholding method, the linear regression method and the Naive Bayes method. For the statistical thresholding method, we simplied thresholded each examined data and treated the period associated with the largest $\sigma \times 100\%$ values as the epidemic periods; Both the Naive Bayes method and the linear regression method used the Google query data, the news count data and the volume of infected airplane passengers as input. At each week, the Naive Bayes method treated the estimation of the national epidemic risk state as a binary classification problem. Similar to the Google Flu method, the linear regression model was fitted to approximate contemporary influenza morbidity, and the morbidity estimates were further classified as pandemic and non-pandemic using the probability threshold σ.

As mentioned in the second section, the proposed surveillance system uses the probability threshold σ to calculate the ground-truth epidemic states. It seems that the value of σ controls the tradeoff between sensitivity and accuracy of the surveillance system. Experiment results in Table III show that, all the compared approaches achieve the highest surveillance accuracy with the smallest probability value of $\sigma = 5\%$; however, larger probability tends to increase the sensitivity of the surveillance system.

Our experiment also compared the MHMM with the standard hidden Markov model (HMM), which used the query volume levels and news count levels to estimate the national epidemic risk states. The difference between the MHMM and HMM was that, the HMM only considered the temporal correlation of disease transmission. Experiment results showed that, by incorporating the geographical correlation between countries connected by international airlines, the MHMM method achieved notably higher accuracy. Moreover,

TABLE II
ESTIMATION AND PREDICTION OF INFLUENZA EPIDEMICS IN DIFFERENT COUNTRIES AND REGIONS

Country and region	**Estimation (Stage 1)**	**Prediction (Stage 1)**	**Prediction (Stage 2)**	**Morbidity missing**	**Country and region**	**Estimation (Stage 1)**	**Prediction (Stage 1)**	**Prediction (Stage 2)**	**Morbidity missing**
Algeria	94.32±0.74%	91.22±1.51%	95.65%	46.41%	Croatia	97.18±3.17%	87.91±1.20%	89.47%	50.88%
Angola	88.95±2.22%	75.43±8.10%	-	77.67%	Denmark	97.82±1.35%	95.63±4.91%	88.46%	21.53%
Burkina Faso	95.41±3.14%	94.22±4.42%	100.00%	71.93%	Estonia	96.34±1.42%	90.24±3.06%	88.46%	41.63%
Cameroon	94.39±3.91%	93.02±4.67%	80.39%	34.29%	Finland	86.88±1.11%	80.11±0.52%	76.92%	36.52%
Cote d'Ivoire	83.09±5.97%	72.56±4.30%	88.46%	47.37%	France	97.71±1.08%	96.50±5.94%	100.00%	19.62%
Egypt	97.63±1.38%	96.02±6.07%	70.00%	31.58%	Georgia	94.32±0.89%	89.07±0.18%	92.59%	52.31%
Ethiopia	94.25±4.06%	91.57±4.39%	100.00%	65.23%	Germany	94.01±0.20%	89.83±0.15%	92.59%	18.66%
Ghana	90.68±4.38%	88.00±5.53%	82.35%	31.26%	Greece	95.53±0.77%	93.04±2.44%	81.82%	30.46%
Kenya	91.42±2.50%	83.17±0.88%	-	29.67%	Hungary	90.01±2.03%	86.59±1.61%	83.33%	63.48%
Madagascar	95.21±4.45%	94.01±4.69%	78.85%	2.23%	Ireland	95.57±0.89%	92.76±2.17%	72.00%	38.28%
Mauritius	92.29±3.67%	91.47±17.7%	100.00%	72.73%	Italy	94.47±0.10%	88.76±1.31%	100.00%	47.21%
Morocco	92.97±1.08%	88.87±0.30%	92.86%	41.31%	Latvia	96.50±1.09%	90.28±3.21%	86.96%	32.85%
Mozambique	97.75±1.93%	94.32±5.75%	82.69%	68.74%	Lithuania	97.30±1.24%	86.96±0.08%	81.48%	56.46%
Nigeria	77.49±2.60%	75.63±3.00%	97.67%	51.83%	Luxembourg	93.29±2.01%	90.72±3.66%	89.47%	45.14%
South Africa	93.02±0.76%	89.06±0.06%	86.54%	10.05%	Malta	86.21±3.19%	78.94±10.5%	88.46%	65.07%
Tanzania	85.27±0.40%	80.39±4.70%	96.15%	47.69%	Moldova	94.42±0.57%	83.48±4.50%	81.48%	54.70%
Tunisia	94.77±2.56%	92.31±9.79%	79.41%	44.18%	Netherlands	92.20±1.98%	88.22±0.71%	96.30%	43.06%
Uganda	93.90±6.68%	87.90±5.46%	84.78%	33.81%	Norway	98.36±1.18%	96.53±2.62%	88.46%	12.12%
Zambia	86.72±1.40%	84.67±1.53%	86.54%	52.15%	Poland	92.83±0.36%	93.71±2.66%	88.24%	8.29%
Bangladesh	92.80±2.46%	87.29±0.50%	78.00%	50.88%	Portugal	93.22±1.10%	90.39±0.11%	90.91%	17.22%
Cambodia	96.20±2.37%	93.82±4.00%	96.15%	21.85%	Romania	93.72±1.56%	90.37±1.42%	94.12%	14.19%
China	97.89±2.54%	96.77±3.86%	84.78%	1.12%	Russian	97.47±1.59%	96.95±6.19%	90.38%	17.22%
India	95.95±2.26%	92.36±2.67%	92.16%	36.84%	Slovakia	88.58±0.70%	86.09±0.95%	81.48%	48.01%
Indonesia	93.51±2.71%	90.81±1.65%	76.92%	44.02%	Slovenia	94.79±0.91%	93.02±1.43%	96.30%	11.00%
Iran	97.69±2.68%	92.77±2.18%	69.23%	21.85%	Spain	96.57±0.73%	94.20±3.83%	100.00%	18.50%
Iraq	93.94±0.74%	91.00±0.41%	74.00%	41.95%	Sweden	96.07±0.67%	93.42±3.17%	92.00%	24.88%
Israel	96.26±2.30%	90.56±1.53%	85.71%	54.07%	Switzerland	95.99±1.51%	93.32±2.94%	73.33%	26.00%
Japan	95.10±1.41%	91.06±0.85%	100.00%	0.00%	Ukraine	96.61±3.18%	93.12±5.04%	80.00%	36.36%
Jordan	88.07±0.97%	87.89±0.29%	74.51%	55.50%	Canada	94.90±1.89%	93.55±2.61%	98.08%	23.29%
Kazakhstan	96.68±2.19%	93.31±2.89%	96.00%	53.75%	Costa Rica	94.62±2.85%	89.86±1.11%	100.00%	34.77%
Korea	93.46±1.89%	88.04±3.87%	100.00%	17.07%	Cuba	86.77±8.89%	83.32±1.38%	87.88%	44.82%
Kyrgyzstan	95.65±2.30%	90.94±4.45%	-	55.50%	Dominican	96.84±6.29%	96.19±6.34%	96.97%	15.31%
Malaysia	84.47±5.42%	78.02±9.61%	100.00%	7.81%	El Salvador	92.79±3.96%	90.08±4.79%	100.00%	31.90%
Mongolia	92.47±1.54%	90.23±1.03%	84.21%	15.79%	Guatemala	88.42±1.42%	75.53±5.21%	100.00%	51.36%
Nepal	95.72±2.95%	88.20±0.32%	79.07%	65.87%	Honduras	94.90±3.30%	93.34±3.74%	100.00%	29.82%
Pakistan	94.14±4.61%	90.94±5.07%	73.17%	34.45%	Jamaica	92.70±3.58%	89.55±2.80%	81.25%	36.84%
Philippines	91.20±5.14%	88.17±4.95%	100.00%	1.91%	Mexico	98.45±1.84%	97.70±3.04%	81.82%	3.19%
Qatar	98.77±3.61%	82.60±6.70%	72.55%	58.85%	Nicaragua	91.80±0.66%	89.36±0.09%	100.00%	49.44%
Singapore	94.27±2.93%	92.11±3.43%	94.23%	26.32%	Panama	95.20±1.69%	91.98±1.69%	100.00%	34.45%
Thailand	88.48±3.32%	86.64±4.31%	100.00%	3.35%	USA	93.66±0.50%	92.31±1.88%	100.00%	4.15%
Turkey	98.20±1.31%	94.45±1.30%	80.77%	43.06%	Argentina	96.22±0.79%	92.94±2.11%	79.59%	2.07%
Uzbekistan	72.68±0.10%	65.75±14.0%	100.00%	87.88%	Bolivia	94.47±2.79%	92.11±3.50%	100.00%	53.11%
Viet Nam	92.46±3.24%	86.84±5.10%	100.00%	17.22%	Brazil	96.78±3.38%	95.29±4.63%	90.20%	0.16%
Albania	79.99±4.43%	76.25±6.12%	87.50%	79.43%	Chile	96.48±0.58%	93.26±0.88%	90.91%	3.51%
Armenia	86.99±0.70%	83.84±4.70%	-	70.33%	Colombia	96.57±4.04%	95.23±5.51%	87.50%	5.42%
Austria	96.69±1.57%	92.94±0.70%	85.71%	56.46%	Ecuador	91.89±1.31%	86.93±2.90%	100.00%	48.48%
Azerbaijan	95.98±6.93%	86.94±1.41%	100.00%	66.35%	Guadeloupe	87.43±2.75%	83.87±0.01%	90.38%	49.44%
Belarus	93.03±2.51%	71.79±1.30%	85.19%	58.37%	Martinique	89.25±3.10%	85.94±12.1%	76.92%	53.11%
Belgium	87.98±1.04%	84.68±1.01%	92.00%	33.33%	Uruguay	89.51±0.12%	82.98±0.70%	100.00%	39.55%
Bosnia	90.42±0.14%	86.04±1.74%	-	70.02%	Venezuela	95.77±5.83%	87.54±0.07%	100.00%	54.07%
Britain	97.24±1.76%	97.05±4.54%	88.89%	9.25%	Australia	96.77±1.14%	95.58±1.79%	97.73%	2.23%
Bulgaria	94.83±1.09%	90.33±1.84%	85.19%	40.67%	Fiji	86.35±1.01%	84.74±0.82%	78.95%	49.92%
Czech	96.83±2.95%	92.75±7.34%	83.33%	53.43%	New Zealand	98.65±1.32%	97.30±6.73%	81.82%	53.27%

we found that the MHMM method was more sensitive and accurate for epidemic surveillance during the pandemic period. Specifically, the HMM method had lower detection accuracy during the year 2009 (86.77%, σ=15%), while the MHMM method showed 94.79% (σ=15%) average accuracy among the examined countries and regions in 2009.

V. DISCUSSION AND CONCLUSION

The timing of influenza epidemic outbreaks varies from country to country, which makes it a challenge to provide timely, sensitive and reliable surveillance results at the global scale [31]. To address this challenge, this paper presents a spatial-temporal method to detect influenza epidemics using heterogeneous data collected from the Internet. By taking advantage of the geographical and temporal correlation of influenza transmission, the proposed multivariate hidden Markov model could detect potential influenza epidemics in 106 countries and regions with over 90% accuracy on average.

Traditional influenza surveillance networks built on virological testing data suffer from one to two weeks' lag of reporting

TABLE III
ACCURACY OF REAL-TIME DETECTION OF INFLUENZA EPIDEMICS

Epidemic probability(σ)	**Statistical thresholding method (%)**[1]			**Linear regression (%)**	**Naive Bayes (%)**	**HMM (%)**	**MHMM (%)**
	News count	**Query volume**	**Imported infections**				
5%	85.23	93.43	91.99	80.57±2.30	95.47±2.86	89.92±1.68	97.10±1.01
10%	70.88	86.70	85.06	72.87±2.48	91.52±1.62	89.56±1.34	94.99±1.02
15%	58.01	78.39	79.61	65.44±2.20	88.03±2.19	87.18±1.13	93.34±1.11
20%	47.06	73.38	75.36	61.49±1.88	84.12±2.86	87.86±0.57	91.82±1.12
25%	40.89	67.72	72.30	57.30±2.56	80.78±2.52	86.31±0.23	90.26±1.11

[1] The statistical thresholding method binarizes each type of data using probability and treats the period with the largest $\sigma \times 100\%$ values as epidemics.

time; However, thanks to the daily-updated influenza-related query data and news data, the proposed Internet based syndromic surveillance method could provide epidemic surveillance results in nearly real time. Furthermore, the weekly updated Markov model could even provide epidemic alerts ahead of the outbreaks of influenza epidemics with over 89% accuracy. Timely surveillance results may enable public health officials and health professionals to better respond to epidemic outbreaks. For example, if a particular country or region is predicted to experience an influenza epidemic, it may be possible to focus additional resources to identify the source of the outbreak, and provide extra drug capacity as necessary.

The Internet based surveillance may become an important defense line against influenza epidemic, yet there are some restrictions. First, the WHO morbidity data reported by different countries may be missing or inaccurate. The lack of reliable surveillance data makes it a challenge to fit more sophisticated models with more parameters, e.g. the random Markov field, for epidemic simulation at the global scale. Secondly, our method used the morbidity of the departure location to estimate the number of infections in airline passengers. Given the large volume of airline passengers registered by the IATA (up to 2.85 million each week) and our long examined period of eleven years, the morbidity rate of departure location might be a proper estimation of the passenger morbidity rate. However, since the passengers are not necessarily uniformly sampled from the population, more precise models are planed to be studied in the future if more comprehensive and reliable data are available.

VI. ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their constructive comments on this paper. We would also like to thank Shengli Li for his contribution in preparing the data.

REFERENCES

[1] D. Lavanchy, The importance of global surveillance of influenza, *Vaccine*, 17, S24-S25, Aug. 1999.

[2] Introduction of the WHO Global Influenza Surveillance Network is available at http://www.influenzacentre.org/centre_GISN.htm, accessed at Jun 10, 2015.

[3] The influenza morbidity data are available at the FluNet http://www.who.int/influenza/gisrs_laboratory/flunet/en/, accessed at Jan 20, 2016.

[4] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012-1014, 19 February 2009.

[5] P. Polgreen, Y. Chen, and D. Pennock. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11), December 2008.

[6] K. Wilson and J. Brownstein. Early detection of disease outbreaks using the internet. *Canadian Medical Association journal*, 180 (8), April 14, 2009.

[7] X. Zhou and H. Shen, Notifiable infectious disease surveillance with data collected by search engine, *Journal of Zhejiang University Science C: Computer and Electornics*, 11(4), Apr. 2010.

[8] X. Zhou, Q. Li, Z. Zhu, H. Zhao, H. Tang and Y. Feng, Monitoring Epidemic Alert Levels by Analyzing Internet Search Volume, *IEEE Transactions on Biomedical Engineering*, vol.60, no.2, pp.446,452, Feb. 2013.

[9] A. Valdivia, J. Lpez-Alcalde, M. Vicente, M. Pichiule, M. Ruiz, M. Ordobas, Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks - results for 2009-10. *Eurosurveillance: bulletin europeen sur les maladies transmissibles*, European communicable disease bulletin 15.29(2010):2-7.

[10] S. Cook, C. Conrad, A. L. Fowlkes, M. H. Mohebbi. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *Plos One*, 6.8(2011):61-61.

[11] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, L. Simonsen. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *Plos Computational Biology*, 9.10(2013).

[12] D. Butler, When Google got flu wrong. *Nature*, 494.7436(2013):155-6.

[13] C. Freifeld, K. Mandl, B. Reis, J. Brownstein. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association* 15.2(2008):150-157.

[14] K. Khan, J. Arino, W. Hu, P. Raposo, J. Sears, F. Calderon, C. Heidebrecht, M. Macdonald, J. Liauw, A. Chan, M. Gardam. Spread of a novel influenza A (H1N1) virus via global airline transportation. *New England Journal of Medicine* 361.2(2009):212-4.

[15] The IATA webstite, The airline trasportation data is available at http://www.iata.org/services/statistics/pages/index.aspx, accessed at 1 Jan. 2016.

[16] C. Robertson, T. Nelson, M. Ying, A. Lawson. Review of methods for spacetime disease surveillance. *Spatial and spatio-temporal epidemiology* 1.2-3(2010):105-16.

[17] J. Besag, J. Newell. The detection of clusters in rare diseases. *J R Stat Soc Ser A* 1991;154:14355.

[18] M. Kulldorff, U. Hjalmars. The Knox method and other tests for spacetime interaction. *Biometrics* 1999;55:54452.

[19] M. Kulldorff, R. Heffernan, J. Hartman, R. Assuncao, F. Mostashari. A space time permutation scan statistic for the early detection of disease outbreaks. *PLoS Med* 2005;2:21624.

[20] K. Kleinman, R. Lazarus, R. Platt. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Am J Epidemiol* 2004;159(3):21724.

[21] N. Best, S. Richardson and A. Thomson, A comparison of Bayesian spatial models for disease mapping, *Stat Methods Med Res*, 14(1):35C59, Feb. 2005.

[22] A. Lawson. Bayesian disease mapping; hierarchical modeling for spatial epidemiology. *New York: CRC Press*, 2009.

[23] B. Reis, C. Kirby, et al. AEGIS: a robust and scalable real-time public health surveillance system. *J AHIMA*, 2007;14(5):5818.

[24] W. Wong, A. Moore. Classical time-series methods for biosurveillance. Handbook of biosurveillance. London: Elsevier Academic Press, 2006. p. 21734.

[25] Y. Kim, M. O'Kelly. A bootstrap based spacetime surveillance model with an application to crime occurrences. *J Geogr Syst* 2008;10(2):14165.

[26] X. Zhou, J. Ye and Y. Feng, Tuberculosis Surveillance by Analyzing Google Trends, *IEEE Transactions on Biomedical Engineering*, Vol. 58, No. 8, Aug. 2011.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2690631, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[27] R. Watkins, S. Eagleson, B. Veenendaal, G. Wright, A. Plant. Disease surveillance using a hidden Markov model. *BMC Med Inform Decis Mak* 2009;9(1):39.

[28] W. Sun, T. Cai. Large-scale multiple testing under dependence. *J R Stat Soc Series B Stat Methodol* 2009;71(2):393424.

[29] The Google Trend Service. available at: http://www.google.com/insights/search/, accessed at 10 Jun. 2016.

[30] S. Briand, A. Mounts and M. Chamberland, Challenges of global surveillance during an influenza pandemic, *WHO report*, available at http://www.who.int/influenza/surveillancemonitoring/Challenges_global_surveillance.pdf, accessed at 10 Jun. 2016.

[31] H. Oshitani, T. Kamigaki, A. Suzuki, Major issues and challenges of influenza pandemic preparedness in developing countries, *Emerging Infectious Diseases* 14.6(2008):875-80.

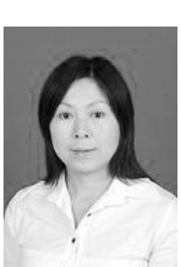
[32] D. Forney. The viterbi algorithm. *Proceedings of the IEEE* 61.5(2015):268-278.

[33] Local news about the H5N1 infected birds in France, http://www.futura-sciences.com/magazines/sante/infos/actu/d/medecine-grippe-aviaire-premier-cas-virus-h5n1-confirme-france-8282/, accessed at 10 Jun. 2016.

[34] News reports of the influenza epidemic outbreaks in 2009, Belarus closes schools to contain swine flu, available at http://data.minsk.by/belarusnews/112009/70.html, accessed at 10 Jun. 2016.

Xichuan ZHOU (S'06-M'13) received the B.S. and Ph.D degrees from Zhejiang University, China, in 2005 and 2010 respectively. Since the end of 2010, he is an Associate Professor and the Assistant Dean of the College of Communication Engineering, Chongqing University (CQU), China. He received the Young and Middle Age Backbone Faculty Award Program of Chongqing province in 2016, the Outstanding Young Faculty Award of Chongqing University in 2012. Dr. Zhou's research interests include medical image processing, machine learning and parallel computing circuits and systems. He has published over 20 papers in the area of electronic engineering and information science on peer-reviewed journals including IEEE Transactions on Biomedical Engineering, IEEE Transactions on Circuits and Systems and IEEE Transactions on Electron Devices.

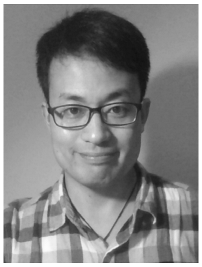
Fan YANG received the B.S. degree from Chongqing University, China, in 2013. He is now a Graduate Student in the College of Communication Engineering, Chongqing University, China. His research interests include machine learning and medical image processing.



Qin LI received her B.S. degree and master degree from Chongqing Medical University and Chinese CDC, China. She is the director of Infectious Disease Institution of the Chongqing Centers for Disease Control and Prevention, China. She is senior epidemiology expert and a member of Chinese Ministry of Health's Committee of Infectious Disease Standard.

Yujie FENG received the B.S. degree from Shanxi Medical University, China, in 2005. She received the master degree from Chongqing Medical University, China, in 2009. She is now working as a resident physician at the Southwest hospital of the Third Military Medical University in Chongqing, which is one of the best hospitals in west China. Her research interests include medical image processing and computer aided diagnosis.

Fang TANG (S'07-M'14) received the B.S. degree from Beijing Jiaotong University, China, in 2006. He received M.Phil and Ph.D degrees from Hong Kong university of science and technology in 2009 and 2013, respectively, where he worked as research associate. Since the end of 2013, he is a Distinguished Research Fellow and Tenure-Track Assistant Professor in the College of Communication Engineering, Chongqing University (CQU), China. His research interests include electronic engineering for biomedical applications.



Shengdong HU received the M.S. degree in material science and engineering in 2005 from Sichuan University, Sichuan, China, and the Ph.D. degree in microelectronics in 2010 from University of Electronic Science and Technology of China (UESTC). From 2010, he has worked in the Chongqing University, Chongqing, China, where he has engaged in research on electronic engineering.

Zhi LIN received the B.S. and Ph.D. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2009 and 2015, respectively. From 2016, he has worked in the Chongqing University, Chongqing, China, where he has engaged in research on electronic engineering.

Lei ZHANG Lei Zhang (M14) received the Ph.D. degree in circuits and systems from the College of Communication Engineering, Chongqing University, He was selected as a Hong Kong Scholar in China in 2013. He has authored more than 50 scientific papers in top journals. His current research interests include machine learning, pattern recognition, computer vision, electronic olfaction and intelligent systems. Dr. Zhang was a recipient of Outstanding Reviewer Award of Sensor Review Journal in 2016, Hong Kong Scholar Award in 2014, Academy Award for Youth Innovation of Chongqing University in 2013 and the New Academic Researcher Award for Doctoral Candidates from the Ministry of Education, China, in 2012.