OPEN ACCESS Freely available online

PLOS ONE

Inferring Epidemic Network Topology from Surveillance Data

Xiang Wan[1]*, Jiming Liu[2]*, William K. Cheung[2], Tiejun Tong[3]

1 Department of Computer Science and Institute of Theoretical and Computational Study, Hong Kong Baptist University, Kowloon Tong, Hong Kong, **2** Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, **3** Department of Mathematics and Institute of Theoretical and Computational Study, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract

The transmission of infectious diseases can be affected by many or even hidden factors, making it difficult to accurately predict when and where outbreaks may emerge. One approach at the moment is to develop and deploy surveillance systems in an effort to detect outbreaks as timely as possible. This enables policy makers to modify and implement strategies for the control of the transmission. The accumulated surveillance data including temporal, spatial, clinical, and demographic information, can provide valuable information with which to infer the underlying epidemic networks. Such networks can be quite informative and insightful as they characterize how infectious diseases transmit from one location to another. The aim of this work is to develop a computational model that allows inferences to be made regarding epidemic network topology in heterogeneous populations. We apply our model on the surveillance data from the 2009 H1N1 pandemic in Hong Kong. The inferred epidemic network displays significant effect on the propagation of infectious diseases.

Citation: Wan X, Liu J, Cheung WK, Tong T (2014) Inferring Epidemic Network Topology from Surveillance Data. PLoS ONE 9(6): e100661. doi:10.1371/journal.pone.0100661

Editor: Alessandro Vespignani, Northeastern University, United States of America

Received July 1, 2013; **Accepted** May 28, 2014; **Published** June 30, 2014

Funding: This study was supported by grants (FRG1/13-14/021) from Hong Kong Baptist University, who also supported the publication costs. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: xwan@comp.hkbu.edu.hk (XW); jiming@comp.hkbu.edu.hk (JML)

Introduction

Recent outbreaks of infectious diseases have stressed the urgency of effective research on the dynamics of infectious disease spread over geographical regions and in various populations [1]. The pandemic of influenza A (H1N1) in 2009 struck more than 208 countries and territories experienced the pandemic, collectively causing at least 12,799 deaths [2]. Great benefits would be gained from the rapid formulation of appropriate control policies to contain the spread of the infectious disease and eliminate it from the population. However, the complex dynamics of infectious disease spread poses a significant challenge to the design of a realist control strategy. Computational modeling has long been an important tool for understanding spread patterns of infectious diseases, predicting outbreak severity, evaluating the efficacy of interventions, and optimizing the deployment of new control policies. The majority of disease models are based on a compartmental model called the Susceptible-Infected-Recovered (SIR) model [3–6]. It studies the spread of infectious diseases by tracking the number (S) of people susceptible to the disease, the number (I) of people infected with the disease, and the number (R) of people who have had the disease and are now recovered. Assuming the population mixes at random, three ordinary differential equations are defined for $S(t)$, $I(t)$, and $R(t)$ at time t:

$$dS/dt = -\alpha S(t)I(t) \quad (1)$$

$$dI/dt = \alpha S(t)I(t) - kI(t) \quad (2)$$

$$dR/dt = kI(t). \quad (3)$$

Here, $\alpha \geq 0$ is the effective transmission rate and $k \geq 0$ is the recovery rate. The value of α is a key indicator for the guidance of implementing control and intervention policies.

The SIR model and its variants are appropriate for modeling the temporal dynamics of epidemics in the randomly mixed population [7–9]. However, it is difficult to use such models to investigate complex social structures or mixing patterns that depend on network structure. Network epidemic models represent an alternative to compartmental models that can more easily capture the effects of social structure. An epidemic network consists of a set of nodes and a set of links that connect them, where the nodes correspond to spatial locations with reported (or

observed) disease incidences over time and the directional links indicate the probability (or likelihood) of disease transmission from one node to another over time. It can be used to characterize the temporal-spatial patterns of disease transmission. Determining an accurate epidemic network requires knowledge of every individual (or host) and every relationship between individuals. A detailed review [10] summarizes four major types of models, including patch models [11–13], distance-transmission models [14], multi-group models [15,16], and network models [17]. However, for all but the smallest population, collecting individual-level data is an impractically time-consuming task. To bypass the difficulties of collecting data, researchers started to investigate several types of computer-generated networks in the context of disease transmission in population-scale studies [18–22]. Given the mean-field theory, they have proposed to model epidemic spread in scale-free networks. However, it remains an open question whether real networks are close to scale-free, or only scale-free over a finite domain [23]. The dynamics of infectious disease spread rely strongly on the structure of the epidemic network topology.

Related topics, such as social influence through networks, the diffusion of innovations, and information propagation, have also been studied in the context of various disciplines including economics [24], public health [25], scientific publishing [26], and virus propagation [27]. However, each of these models has one or more aspects that are problematic in studying the temporal-spatial dynamics of infectious disease spread. Some do not capture the probabilistic nature of infection while others make assumptions about the types of interactions occurring between individuals that are often not valid in the context of disease transmission. How to infer the epidemic network topology remains a challenging research topic.

To accurately catch when and where outbreaks emerge at the first time, one approach at the moment is to implement surveillance systems in regional or national health and medical centers. The accumulated surveillance data including temporal, spatial, clinical, and demographic information, can provide valuable information with which to infer the underlying epidemic network of infectious disease spread. In this work, we introduce a new computational model that can discover the epidemic network of infectious disease spread from the surveillance data. In our proposed model, the dynamics modelled in the classical SIR model is described by an inhomogeneous Poisson process characterized by a piecewise rate function, and the spatial relationships are characterized by interactions of multiple inhomogeneous Poisson processes in a network. Our proposed model allows inferences to be made regarding the progression patterns of infectious diseases in heterogeneous populations. We apply our model on the surveillance data from the 2009 H1N1 pandemic in Hong Kong. The inferred epidemic network displays significant effect on the propagation of infectious diseases, and is useful to public health authorities in predicting the influence of future prevalence and the implications of control polices.

Materials and Methods

Classic modeling of infectious diseases assumes that the population is well-mixed. However, this assumption is unrealistic for many diseases with spatial spread patterns. Here we first describe the dynamics of classical SIR models through an inhomogeneous Poisson process and then formulate a new stochastic network model that explicitly considers the geographical structure to capture the temporal-spatial dynamics of infectious disease spread in heterogeneous populations.

Poisson process for modeling the dynamics of classical SIR models

There is no analytic solution to solve SIR-type dynamics without making approximations. To model the dynamics in continuous time, the discrete-time models are often used with a given time interval Δt. Let S_t, I_t, and R_t be discrete random variables for the number of susceptible, infected, and recovered individuals at time t. Using the Euler method, the SIR model for the sub-population l can be rewritten as three equations:

$$S_{t+\Delta t} = S_t - \Delta t \alpha S_t I_t \tag{4}$$

$$I_{t+\Delta t} = I_t + \Delta t \alpha S_t I_t - \Delta t k I_t \tag{5}$$

$$R_{t+\Delta t} = R_t + \Delta t k I_t. \tag{6}$$

In [28,29], the progression of disease spread is characterized by tracking the number of S_t with a chain binomial model. The number of susceptible members $S_{t+\Delta t}$ (Δt represents the infectious period of the disease and is always chosen to be $1/k$) at time $t+\Delta t$ is a binomial random variable that depends on S_t and $I_t\alpha$, $S_{t+\Delta t} \sim Bin(S_t, 1-I_t\alpha)$, which provides a recursive relationship between $S_{t+\Delta t}$ and S_t and produces a formal stochastic process. We use an alternative approach to model the dynamics of infectious disease spread. In an epidemic outbreak, the number of new infections during a time interval is of major concern. Let $i_{t+\Delta t} = S_t - S_{t+\Delta t}$ be the number of new infections between time t and time $t+\Delta t$. Let p be the probability that a contact between a susceptible and an infected individual results in a new infection and c is the average number of susceptible members to whom an infected individual may spread the disease at time t. The infectious individuals I_t are assumed to infect susceptible members S_t only at time t. After that time, they are no longer infectious. This is reasonable because patients, once confirmed as infected, will have much less possibilities to spread the disease since they may start the treatment, take rest at home, adopt some measure to prevent the disease spread (such as wearing a face mask outside), or be quarantined. With this assumption, we have the following proposition.

Proposition 1.

$$i_{t+\Delta t} \sim Poisson(\lambda(t)), \lambda(t) = \psi i_t, \psi = cp. \tag{7}$$

Proof: Given one infected person and one person in the population, the probability they meet each other is c/N. Then the probability that the contact results in an infection is $\frac{cp}{N}$. Given the infected person, the number of new infected people in $S(t)$ is $Bin(S(t), \frac{cp}{N})$. Since $S(t)$ is large, $\frac{cp}{N}$ is very small, and $S(t)\frac{cp}{N}$ is finite, we can approximate $Bin(S(t), \frac{cp}{N})$ with $Poisson(\frac{cpS(t)}{N})$. Because $\frac{S(t)}{N} \approx 1$, we get that given one infected person, the

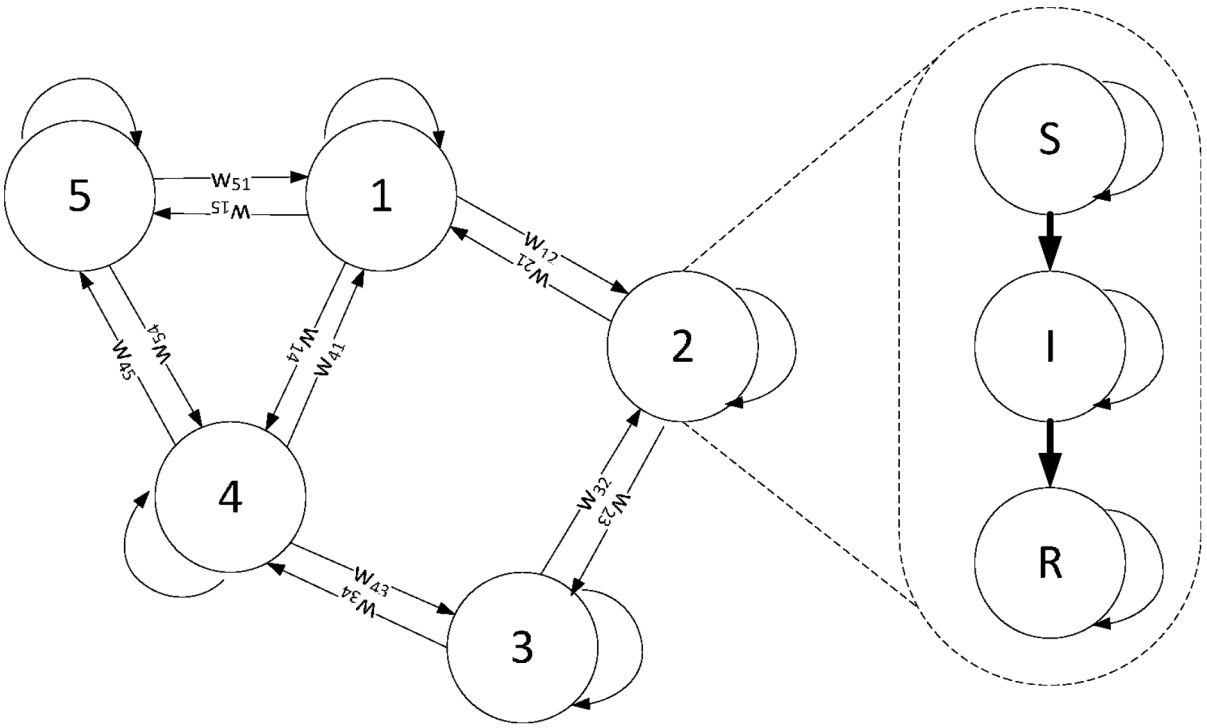


Figure 1. A toy example of infectious disease spread in a five node network. Each node represents a physical location associated with an inhomogeneous Poisson process shown as an example for node 2. Each edge is associated with w_{ij} measuring the spreading trend from node i to node j.
doi:10.1371/journal.pone.0100661.g001

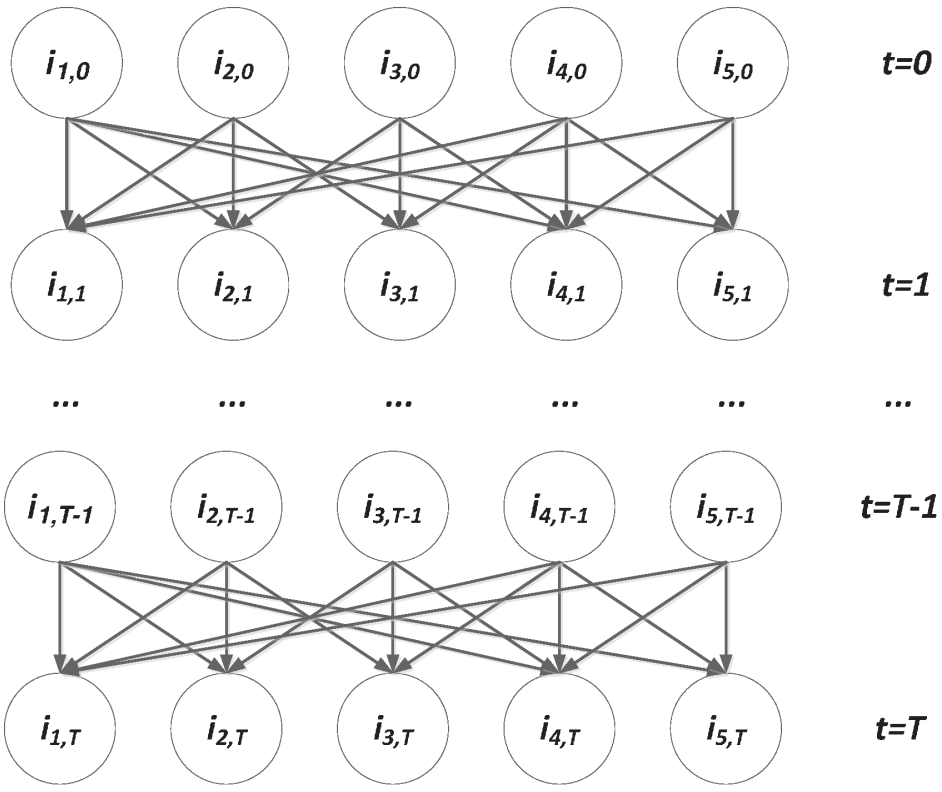


Figure 2. The Markov network, reduced from the spreading network in Figure 1 with the independence assumption. The state of each node $i_{l,t}$ is the number of new infections at location l (node l in Figure 1) at time t. The states of nodes at time t are independent and only dependent on the states of nodes at time $t-\Delta t$.
doi:10.1371/journal.pone.0100661.g002

Table 1. Genetic Algorithm.

1. Randomly generate an initial population M(0) of the network structure and for each $W_i \in M(0)$, estimate Ψ_i and Φ_i using Eq.(9) and Eq.(10) and use the computed maximum likelihood as the fitness of W_i.
2. Copy the top 10 percent of $M(t)$ into $M(t+1)$.
3. Randomly choose four network structures from $M(t)$ as parents and use crossover operator to generate two child network structures.
4. Conduct the mutation for both generated child network structures.
5. Compute the fitness of two generated child network structures and save the better one in $M(t+1)$.
6. Repeat Step 3–5 until the capacity of $M(t+1)$ is full.
7. $t=t+1$.
8. Repeat Step 2 until the new generated population does not improve the fitness value.
9. Output the adjacency matrix with the best value of the fitness function in the last generation.

doi:10.1371/journal.pone.0100661.t001

number of new infected people follows $Poisson(cp)$. Then given i_t infections at time t, the number of new infections $i_{t+\Delta t} \sim Poisson(cpi_t)$ (assume the social contacts of infected people are independent of each other).

The above equation (Eq.(7)) defines an inhomogeneous Poisson process - a stochastic counting process characterized by an intensity function [30]. It has an advantage over the chain binomial models in the dynamic modeling of infectious disease spread. In the chain binomial models, varying the selection of Δt can distort the dynamic patterns [31]. In contrast, the Poisson distribution can be freely adjusted with respect to Δt without affecting the stochastic process due to the Poisson property. Motivated by this advantage, we propose a new stochastic network model of infectious disease spread.

Network modeling of infectious disease spread

Consider the surveillance data $\mathcal{I} \in \mathbb{R}_+^{L \times T}$ from L locations. Each element $i_{l,t}$ of $\mathcal{I}_l = [i_{l,1}, \cdots, i_{l,T}]$ corresponds to the number of new infections between time $t-\Delta t$ and time t at location l. Let $\mathcal{I}^\tau$ denote the transpose of $\mathcal{I}$ and then $\mathcal{I}_t^\tau = [i_{1,t}, \cdots, i_{L,t}]$ correspond to the new infections at all locations between time $t-\Delta t$ and time t. Ignoring the effects of network structure, the dynamics of $\mathcal{I}_l$ can be modeled as the inhomogeneous Poisson process defined in Eq.(7). In the network modeling, we need to capture both the dynamics of the Poisson process at every location and the dependency among multiple Poisson processes at different locations linked in a geographic network.

We use a directed graph $G(V,W)$ to represent a geographic network, where V is the set of nodes and $W = [w_{ij}]_{L \times L}$ $(L = |V|)$ is the adjacency matrix of the graph. Each node indicates a sub-population in one location. Each $w_{ij} \in \{0,1\}$ indicates the existence

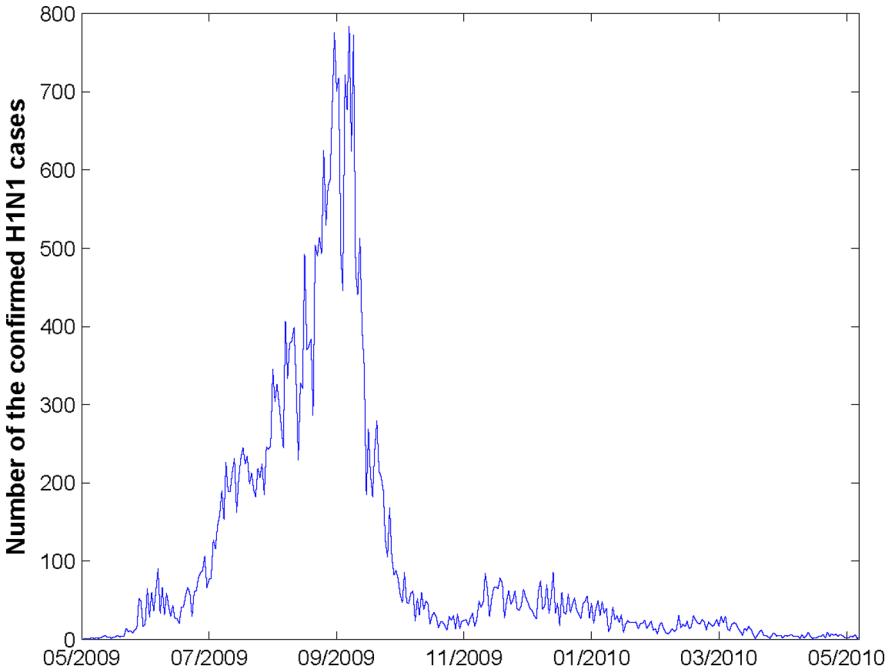


Figure 3. Daily H1N1 epidemic curve in Hong Kong from May 1, 2009 to May 23, 2010. The epidemic curve of confirmed H1N1 cases reaches its peak at the end of September, 2010.
doi:10.1371/journal.pone.0100661.g003

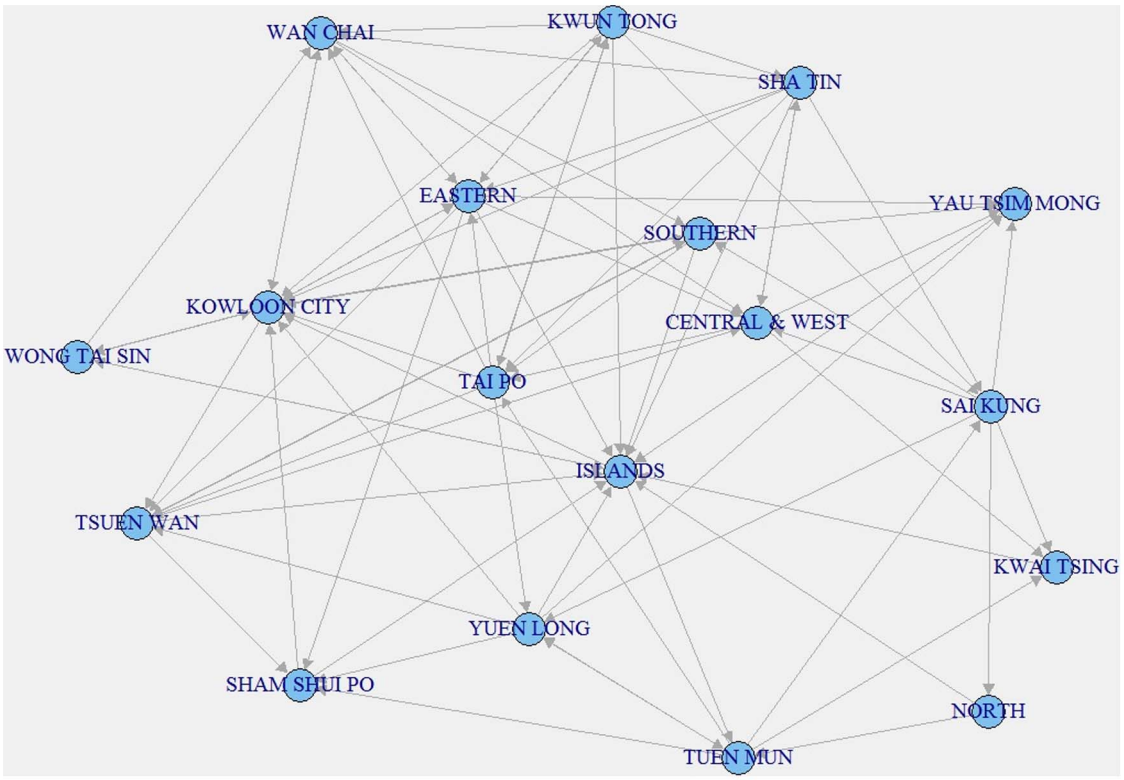


Figure 4. Computed spreading network of 2009 H1N1 in Hong Kong.
doi:10.1371/journal.pone.0100661.g004

of infection spread from node i to node j. Figure 1 provides a toy example of a five node network. Each node l is associated with an inhomogeneous Poisson process characterized by an intensity function $\lambda_l(t)$ and its own $\psi_l = c_l p$ (p is a constant for a specific infectious disease). Let $n(l) = \{m_1, \cdots, m_k\}$, where $w_{m \in n(l),l} = 1$, represent the adjacent nodes of node l in the disease spread network. The spatial interactions will change the intensity function of each node. Thus $\lambda_l(t)$ will not only depend on $(i_{l,t}, \psi_l)$ but also be associated with $(i_{m \in n(l),t}, \psi_m)$.

We consider a generalized linear model (GLM) for the intensity function $\lambda_l(t)$ with respect to $(i_{l,t}, \psi_l)$ and $(i_{m \in n(l),t}, \psi_m)$. The rate of the Poisson process defined in Eq.(7) is rewritten as

$$\lambda_l(t) = \psi_l i_t + \phi_l \sum w_{ml} i_{m,t}. \tag{8}$$

The values of w_{ml} give rise to a transmission network of infectious disease across different locations. For a specific location l, the value of ψ_l measures the speed of disease spread caused by internal infections and the value of ϕ_l measures the speed of disease spread caused by external infections (or imported infections). Our goal is to estimate w_{ml}, ψ_l, and ϕ_l using the surveillance data $\mathcal{I}$.

Parameter estimation

In principle, it is intractable to infer the parameters in Eq.(8) on an arbitrary network because the conditional distribution $P(\mathcal{I}|W,\Psi,\Phi)$ is computationally too expensive to obtain. To make the inference tractable, the $P(\mathcal{I}|W,\Psi,\Phi)$ is factorized with an independence assumption, which is that the states of nodes at time t are independent and only dependent on the states of nodes at time $t-\Delta t$. This assumption has been widely applied in the area of machine learning to factorize an exact joint probability distribution into a multiplication of many marginal probability distributions [32]. Then the dependency graph shown in Figure 1 is reduced into the Markov network shown in Figure 2. Consequently, the $P(\mathcal{I}|W,\Psi,\Phi)$ is defined as

$$\begin{aligned} p(\mathcal{I}_0^\tau, \mathcal{I}_1^\tau, \cdots, \mathcal{I}_T^\tau | W,\Psi,\Phi) &= p(\mathcal{I}_0^\tau) \prod_{t=1}^{T} p(\mathcal{I}_t^\tau | \mathcal{I}_{t-\Delta t}^\tau, W,\Psi,\Phi) \\ &= p(\mathcal{I}_0^\tau) \prod_{t=1}^{T} \prod_{l=1}^{L} \frac{\lambda_l(t)^{i_{l,t}} e^{-\lambda_l(t)}}{i_{l,t}!}. \end{aligned} \tag{9}$$

Our target is to find the following maximum likelihood estimators:

$$[\hat{W}, \hat{\Psi}, \hat{\Phi}] = \arg\max_{W,\Psi,\Phi} \mathbf{P}(\mathcal{I}|W,\Psi,\Phi). \tag{10}$$

It can be easily proven that the negative $\log P(I|W,\Psi,\Phi)$ is a biconvex function of W, Ψ, and Φ, which means that the negative $\log P(I|W,\Psi,\Phi)$ is a convex function of W if Ψ and Φ is fixed, and vice versa. Given W, it is straightforward to estimate Ψ and Φ. However, it is still challenging to infer W for the given Ψ and Φ. Although there are many popular optimization approaches for bi-convex problems, they can not be applied in our work because our model is non-continuous and non-differential while most available approaches are gradient based methods. Here we use the genetic algorithm to solve this issue.

Table 2. Results of the analysis of 2009 H1N1 epidemic in Hong Kong.

	AIC	BIC	Log − likelihood	Compared model	Benefit (P − value)
M1	5191.36	5205.39	−2592.58		
M2	5222.34	5315.83	−2591.17	M1	1.00
M3	5177.47	5196.17	−2584.73	M1	0.005
M4	5150.36	5284.53	−2554.18	M2	1.19×10^{-9}
M5*	4821.7	4916.86	−2389.85	M3	≈ 0.0
M6	4841.49	5019.12	−2382.74	M4	≈ 0.0

The P − value is computed by performing the likelihood ratio test between two models in comparison. M5 is the best model that fits the data.
doi:10.1371/journal.pone.0100661.t002

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection. They have many advantages to solve problems where candidate solutions can be described with the chromosome encoding. Finding the network topology is one such problem where one network topology can be viewed as the chromosome of one individual in a generation. The basic concept of GAs is to simulate processes of survival of the fittest. They represent an intelligent exploitation of a random search within a defined search space to solve a problem. Experiments show that many such problems, which prove difficult for traditional methods, are ideal for GAs [33]. We design a genetic algorithm (please see Table 1) to infer the network structure W. The only input to the GA algorithm is the surveillance data (the number of new incidences in a subpopulation during a time interval). The GA algorithm starts from the first generation - a pool of randomly generated adjacency matrix. Based on the evolutionary theory, individuals in subsequent generations can be generated using the typical GA operators (crossover and mutation) and be selected in a way that resembles the natural selection. The crossover (also called recombination) operator is to produce a child individual of the next generation from two parent individuals of the current generation. The mutation operator is to randomly change some parts of the new generated individual. Please check [33] for more details about genetic algorithms. The output is the adjacency matrix with the best value of the fitness function in the last generation.

Results and Discussion

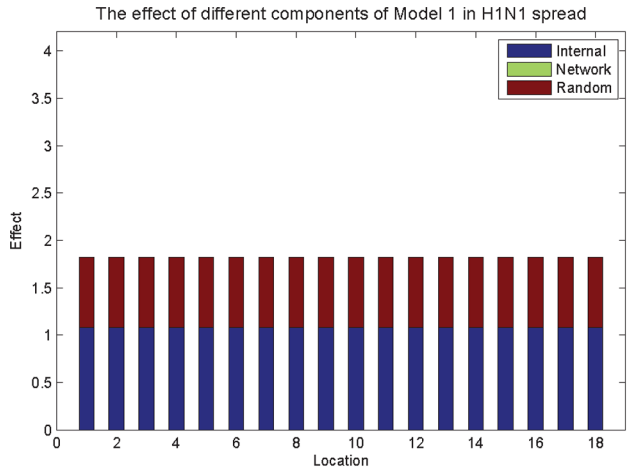
Case study

In the case study, we apply our model on the surveillance data from the 2009 H1N1 pandemic in Hong Kong. We have acquired the time series data of daily number of confirmed H1N1 cases with symptom onset from May 1, 2009 to May 23, 2010. The database includes 36,547 confirmed cases with demographic information on location, age, and sex along with the laboratory-confirmation dates. The epidemic curve of confirmed H1N1 cases (see Figure 3) reaches its peak at the end of September, 2009, after which the intervention procedure comes into effect and the curve goes down. We use the data up to Sept 30, 2009 including 27,898 cases (more than 2/3 of all cases).The infectious period Δt of H1N1 is set 3 days.

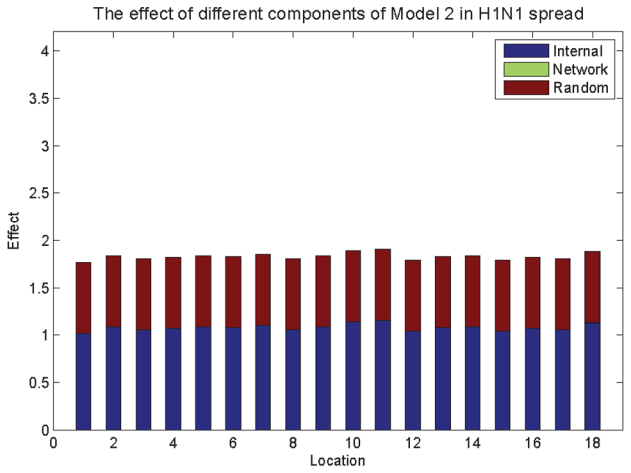
Hong Kong is geographically divided by 18 political areas (districts). Each district is considered as one node in the epidemic network. The learned epidemic network in Figure 4 show how the H1N1 spreads in the geographical network of Hong Kong. To examine the effect of epidemic network topology in the spread of H1N1, we compare the following models:

- M1: $\lambda_l(t)=\psi i_l^t+\varepsilon$.
  M1 is an independent homogeneous model where the infectious disease spreads independently and at the same internal growth rate in different locations.
- M2: $\lambda_l(t)=\psi_l i_l^t+\varepsilon$.
  M2 is an independent heterogeneous model where the infectious disease spreads independently but at the different internal growth rates in different locations.
- M3: $\lambda_l(t)=\psi i_l^t+\phi\sum w_{ml}i_m^t+\varepsilon$.
  M3 is a dependent homogeneous model with uniform network effect where the infectious disease spreads dependently with the same external effect and at the same internal growth rate in different locations.
- M4: $\lambda_l(t)=\psi_l i_l^t+\phi\sum w_{ml}i_m^t+\varepsilon$.
  M4 is a dependent heterogeneous model with uniform network effect where the infectious disease spreads dependently with the same external effect but at the different internal growth rates in different locations.
- M5: $\lambda_l(t)=\psi i_l^t+\phi_l\sum w_{ml}i_m^t+\varepsilon$.
  M5 is a dependent homogeneous model with non-uniform network effect where the infectious disease spreads dependently with the different external effects but at the same internal growth rate in different locations.
- M6: $\lambda_l(t)=\psi_l i_l^t+\phi_l\sum w_{ml}i_m^t+\varepsilon$.
  M6 is a dependent heterogeneous model with non-uniform network effect where the infectious disease spreads dependently with the different external effects and at the different internal growth rates in different locations.

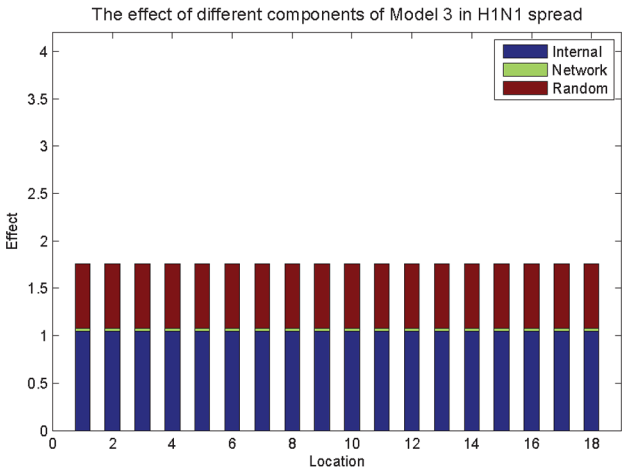
Table 2 summarizes the results for different model formulations. In this paper, the model selection is conducted mainly based on the likelihood ratio test, which is often used to compare the fits of two nested models, one of which (a reduced model) is a special case of the other (the full model). The likelihood ratio of two models can be used to compute a p-value which is then compared to a critical value to decide whether to reject the reduced model in favour of the full model. However, for two models which are not nested (for instances, M4 and M5), we have to use other assessments. AIC [34] and BIC [35] are two popular choices. AIC denotes Akaike Information Criterion that deals with the trade-off between the goodness of fit of the model and the complexity of the model. BIC denotes Bayesian Information Criterion (BIC) that is closely related to the AIC but penalizes the complexity of the model (the number of free parameters in the model) more strongly. Both assessments in the model selection have advantages and disadvantages. Therefore, we report both



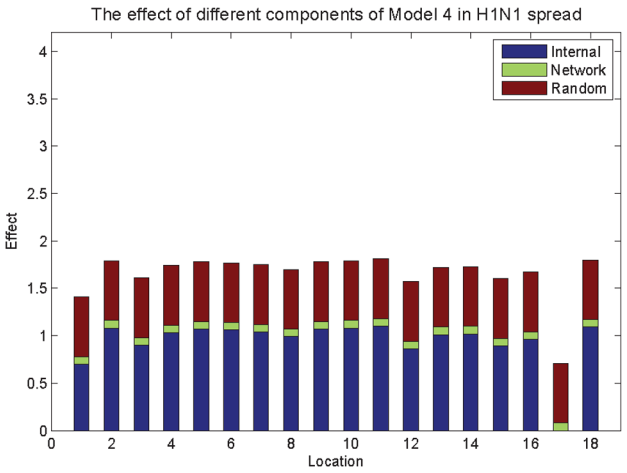
(a) Model 1



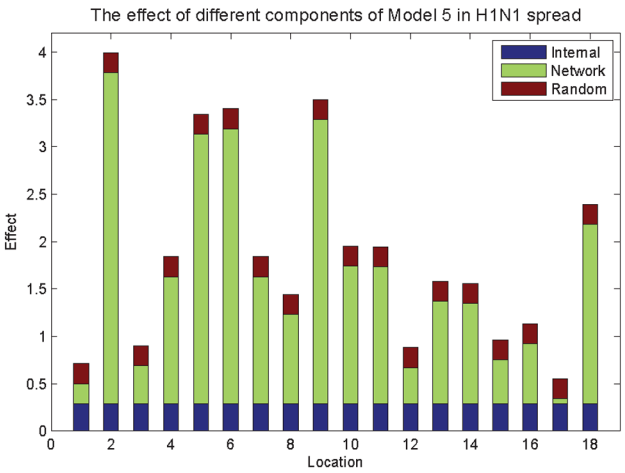
(b) Model 2



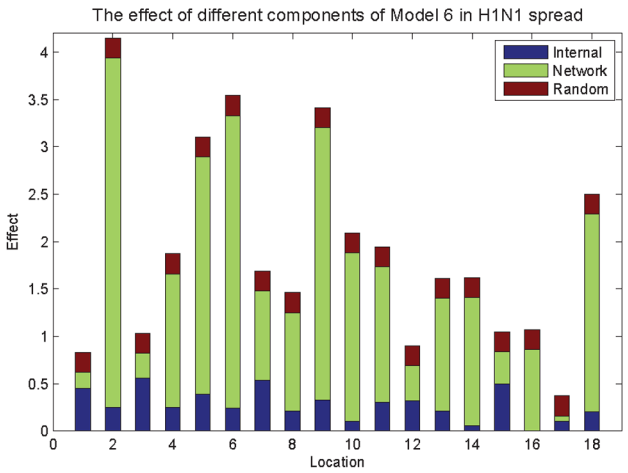
(c) Model 3



(d) Model 4



(e) Model 5



(f) Model 6

Figure 5. The comparison of different effects in 2009 H1N1 spread in Hong Kong. The internal effect is the exponential value of ψ or ψ_I. The network effect is the exponential value of ϕ or ϕ_I. The random effect is the exponential value of ε.
doi:10.1371/journal.pone.0100661.g005

AIC and BIC in Table 2. In most of the times, two assessments agree on the preferred model.

The comparison between M2 and M1 indicates that if the network effect is ignored, the disease spreads at the same growth rates for different locations in Hong Kong. Hong Kong is one of the most densely populated places in the world. Although population varies in different districts, the concentration of people is high in all districts due to the fact that more than 75 percent of Hong Kong area comprises no-built-up areas. Therefore, if each district is examined individually, the density of population in the living space will be the main factor to affect the spread rate of infectious disease. However, once the network topology is taken into consideration in disease spread, different locations show different spreading patterns. Both M4 and M3 provide a better fitness over M2 and M1. The significant benefit of M5 over M3 and M6 over M4 indicates that the network effect, which measures the imported infection, varies between locations. This is mainly due to people's daily travels. Hong Kong possesses a heavy heterogeneous traffic pattern. Therefore, the imported infections vary significantly among different locations. Figure 5 illustrates the effect of different components of all models. We can see that if the network effect is considered for each location, the models M5 and M6 can explain the data very well. The random effect only accounts for a small portion in the explanation of disease propagation within the different locations, which indicates that our approach is an empirically feasible solution in the analysis of future epidemic in Hong Kong. There is no benefit of choosing M6 over M5 ($P - \text{value} = 0.982$). We can see in Figure 5 that in comparison with the network effects, the internal effects play a very small role in explaining the data in both M5 and M6. In Hong Kong, there are intensive transits between districts and as a result, the network effects dominate the epidemic and already explains most variations in the disease spread. Therefore, there is little benefit gained by looking detailed into the internal differences.

How to verify the inferred network topology remains an unresolved issue because the true epidemic network topology is unknown. To our knowledge, the best way to do so is to use the contact data among some infected patients to verify the results. However, such data is not always available and sometimes difficult to collect due to many issues (such as privacy). In our work, we infer the epidemic network topology based on model selection and make the decision from the statistical point of view. Researchers have shown that the spatial spread of infectious diseases has a high correlation with the human mobility both on a large and short scale [36–39]. The inferred epidemic network topology in our work displays such correlations. Some locations that have high connectivity in Figure 4, such as Kowloon City, Central&West, and Eastern, are transit centers in the public transportation network of Hong Kong.

Conclusion

In this paper, we have developed and demonstrated a computational model that extracts the epidemic network topology from the surveillance data of infectious diseases. Especially for disease spread in non-random mixing populations, heterogeneity is very likely exist and should be accounted for. This is done by including region-specific spreading patterns in a stochastic network model. The proposed model distinguishes itself from previous studies in fundamental ways:

- The dynamics of the classical infectious disease model are described by an inhomogeneous Poisson process characterized by a piecewise rate function.
- The spatial dynamics among multiple locations are characterized by interactions of multiple inhomogeneous Poisson processes in a network.
- With one reasonable assumption, the dynamic network is approximated with a Markov network so that the parameters describing the temporal and spatial dependence can be estimated in a tractable computational complexity.
- An efficient genetic algorithm is designed to infer the epidemic topology.

We apply our model on the surveillance data from the 2009 H1N1 pandemic in Hong Kong. It is generally very difficult to verify the inferred network topology from real data because the true epidemic network topology is unknown and it may vary for different types of infectious diseases for the same population. In this work, we propose a new method based on model selection. Our intuition is that if the epidemic network plays an important role in the disease spread, then the heterogenous network model will describe the data better than the homogenous model without considering the epidemic network. Furthermore, the more similar is the network topology to the true one, the better does the model fit the data. Both inferred epidemic networks display significant effects on the propagation of infectious diseases. Therefore, our findings may help policy makers reduce the risk of future epidemics. Besides the study of epidemics, the model developed in this project can be extended to study a wide range of propagation patterns in other complex systems such as the Internet and World Wide Web (WWW), where individuals form multiple communities through which information can propagate in a similar way as the infectious disease does. We believe our work can contribute theoretically and empirically to both computing science and epidemiology.

There are some limitations in our proposed network model. First, our model only focuses on the disease spread within the network and does not consider the imported cases. The parameters may be over-estimated if the number of imported cases is large. One possible solution is to create a pseudo node in our stochastic network model, which imports some infected cases to the network from time to time. Second, the network structure in our model is static - the connections remain constant over time. This may be a problem for the long-term disease spread because the behaviour of a sub-population may change markedly as a consequence of an outbreak. The possible solution is to design a new online optimization algorithm that progressively estimates the network topology over time. Third, our SIR-based model is only suitable for the situation where the susceptible population maintains a relatively constant size and structure. To model malaria transmission where asymptomatic infection plays a central role, the SIR-based model is not a good candidate. We will investigate them in our future work.

Acknowledgments

We thank the editor and two anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

Author Contributions

Conceived and designed the experiments: XW JML WKC. Performed the experiments: XW. Analyzed the data: XW TJT. Contributed reagents/materials/analysis tools: XW. Wrote the paper: XW JML WKC TJT.

References

1. Cohen J, Enserink M (2009) As swine flu circles globe, scientists grapple with basic questions. Science 324: 572–573.
2. Organization WH Pandemic (H1N1) 2009. http://www.who.int/csr/don/2010_01_08/en/index.html. Accessed 2010 Jan 8.
3. Bailey N (1975) The mathematical theory of infectious diseases and its applications. Charles Griffin and Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.
4. Kermack W, McKendrick A (1932) Contributions to the mathematical theory of epidemics. II. the problem of endemicity. Proceedings of the Royal society of London Series A 138: 55–83.
5. Anderson R, May R (1979) Population biology of infectious diseases: Part I. Nature. 280: 361.
6. May R, Anderson R (1979) Population biology of infectious diseases: Part II. Nature 280: 455.
7. Li M, Muldowney J (1995) Global stability for the SEIR model in epidemiology. Mathematical Biosciences 125: 155–164.
8. Kuznetsov Y, Piccardi C (1994) Bifurcation analysis of periodic SEIR and SIR epidemic models. Journal of Mathematical Biology 32: 109–121.
9. Hethcote H (1976) Qualitative analyses of communicable disease models. Mathematical Biosciences 28: 335–356.
10. Riley S (2007) Large-scale spatial-transmission models of infectious disease. Science 316: 1298–1301.
11. Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ (2006) Delaying the international spread of pandemic influenza. PloS Medicine 3: e212.
12. Hufnagel L, Brockmann D, Geisel T (2004) Forecast and control of epidemics in a globalized world. Proceedings of the National Academy of Sciences of the United States of America 101: 15124–15129.
13. Hollingsworth TD, Ferguson NM, Anderson RM (2006) Will travel restrictions control the international spread of pandemic influenza? Nature medicine 12: 497–499.
14. Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, et al. (2001) Dynamics of the 2001 uk foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 294: 813–817.
15. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, et al. (2005) Strategies for containing an emerging influenza pandemic in southeast asia. Nature 437: 209–214.
16. Longini IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, et al. (2005) Containing pandemic influenza at the source. Science 309: 1083–1087.
17. Riley S, Ferguson NM (2006) Smallpox transmission and control: spatial dynamics in great britain. Proceedings of the National Academy of Sciences 103: 12637–12642.
18. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. Physical Review E 63: 066117.
19. Pastor-Satorras R, Vespignani A (2002) Epidemics and immunization in scale-free networks. arXiv preprint cond-mat/0205260.
20. Kuperman M, Abramson G (2001) Small world effect in an epidemiological model. Physical Review Letters 86: 2909–2912.
21. Newman M, Jensen I, Ziff R (2002) Percolation and epidemics in a two-dimensional small world. Physical Review E 65: 021904.
22. Boguná M, Pastor-Satorras R, Vespignani A (2003) Absence of epidemic threshold in scale-free networks with degree correlations. Physical Review Letters 90: 28701.
23. Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. Physical Review E 65: 035108.
24. Rogers EM (2010) Diffusion of innovations. Simon and Schuster.
25. Luke D, Harris J (2007) Network analysis in public health: history, methods, and applications. Annu Rev Public Health 28: 69–93.
26. West JD, Bergstrom TC, Bergstrom CT (2010) The eigenfactor metricstm: A network approach to assessing scholarly journals. College & Research Libraries 71: 236–244.
27. Chen Z, Ji C (2005) Spatial-temporal modeling of malware propagation in networks. IEEE Transactions on Neural Networks 16: 1291–1303.
28. Allen L (2008) An introduction to stochastic epidemic models. Mathematical Epidemiology: 81–130.
29. Becker N, Britton T (2002) Statistical studies of infectious disease incidence. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61: 287–307.
30. Papoulis A (1991) Probability, Random Variables, and Stochastic Processes, 3rd ed. McGraw-Hill, New York.
31. Gustafsson L, Sternad M (2007) Bringing consistency to simulation of population models-poisson simulation as a bridge between micro and macro simulation. Mathematical Biosciences 209: 361–385.
32. Opper M, Saad D (2001) Advanced mean field methods: Theory and practice. MIT press.
33. Mitchell M, Forrest S (1994) Genetic algorithms and artificial life. Artificial Life 1: 267–289.
34. Akaike H (1974) A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19: 716–723.
35. Schwarz G (1978) Estimating the dimension of a model. The annals of statistics 6: 461–464.
36. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. science 312: 447–451.
37. Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, et al. (2012) Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. BMC medicine 10: 165.
38. Colizza V, Barrat A, Barthélemy M, Vespignani A (2007) Predictability and epidemic pathways in global outbreaks of infectious diseases: the sars case study. BMC medicine 5: 34.
39. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. (2012) Quantifying the impact of human mobility on malaria. Science 338: 267–270.