HEALTH INFORMATION SCIENCE AND SYSTEMS

RESEARCH Open Access

Inferring disease transmission networks at a metapopulation level

Xiaofei Yang[1], Jiming Liu[1*], Xiao-Nong Zhou[2] and William KW Cheung[1]

Abstract

Background: To investigate transmission patterns of an infectious disease, e.g., malaria, it is desirable to use the observed surveillance data to discover the underlying (often hidden) disease transmission networks. Previous studies have provided methods for inferring information diffusion networks in which each node corresponds to an individual person. However, in the case of disease transmission, to effectively propose and implement intervention strategies, it is more realistic and reasonable for policy makers to study the diffusion patterns at a metapopulation level when the disease transmission is affected by mobile population, that is, to consider disease transmission networks in which nodes represent subpopulations, and links indicate their interrelationships.

Results: A network inference method called NetEpi (Network Epidemic) is developed and evaluated using both synthetic and real-world datasets. The experimental results show that NetEpi can not only recover most of the ground-truth disease transmission networks using only surveillance data, but also find a malaria transmission network based on a real-world dataset. The inferred malaria network can characterize the real-world observations to a certain extent. In addition, it also discloses some hidden phenomenon.

Conclusions: This research addresses the problem of inferring disease transmission networks at a metapopulation level. Such networks can be useful in several ways: (i) to investigate hidden impact factors that influence epidemic dynamics, (ii) to reveal possible sources of epidemic outbreaks, and (iii) to practically develop and/or improve strategies for controlling the spread of infectious diseases.

Keywords: Network inference, Disease transmission networks, Metapopulation, Bayesian learning

Background

Infectious diseases such as influenza and H1N1 are transmitted between individuals. This process has been widely studied by researchers in biology, statistics, epidemiology, public health, etc. for many years. Their objectives are to help front-line practitioners and policy makers to control disease outbreaks and to prevent severe morbidity and mortality. Various intervention strategies have been applied, including but not limited to vaccination, contact deduction, etc.

Another strategy, contact tracing, is also widely used to prevent disease outbreaks [1]. It is a network-based approach conducted at an individual level. Susceptible individuals are identified and monitored to minimize the chances of infection. The network-based approach not only differentiates individuals in host populations [2] but also allows for performing individual-level simulations [3]. This approach is similar to the one adopted in the research on tracing the transmission pathways of infectious diseases, e.g., malaria, particularly the disease transmission is affected by mobile population, except that here disease transmission is examined at a metapopulation level. Nodes and edges within the metapopulation-based disease transmission networks do not represent individual persons and their pairwise connections (e.g., social contacts [4]); instead, they represent patches of subpopulations (e.g., provinces, cities, and townships) and various transmission pathways among them (e.g., highways and air travel routes). Both individual-based

*Correspondence: jiming@comp.hkbu.edu.hk
[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
Full list of author information is available at the end of the article

and metapopulation-based studies of disease transmission networks are useful in the following aspects:

- Analyzing epidemic phase transition behavior [5];
- Investigating the dominant factors that underlie the spread of a disease epidemic [6];
- Providing suggestions to effectively control epidemics by cutting off transmission pathways and/or isolating certain local regions [7].

Many of the existing disease transmission studies that deal with the above two types of transmission networks share the similar limitation, that is, they assume that network structures are given in advance; for example, contact structures for influenza spreading [8,9] or airlines for the spread of H1N1 [10] and SARS [11]. In these studies, information about which person or location will be infected is given. However, in an actual epidemic, only the spatiotemporal surveillance datasets containing the infection times and locations of reported infection cases are obtained [12]. This type of data provides no knowledge of the hidden transmission pathways that denote the routes of disease propagation among geographical locations. This real-world situation poses a significant and undeniable challenge to policy makers who are responsible for applying intervention strategies at appropriate times and locations. In this regard, inferring disease transmission networks becomes an important and urgent research problem in epidemiological studies (as in [13]).

The network inference problem has been recently and widely studied in the research domain of information diffusion. Based on empirical time-series data that indicates when people become informed or infected, the static network inference problem with a homogeneous edge setting (edge weights are the same for the whole ground-truth network) can be transformed into a combinatorial optimization problem [14]. By formatting it as an MAX-k-COVER problem, Gomez-Rodriguez et al. have proven that selecting the top k edges that maximizes the likelihood of the static network structure is NP-hard. Therefore, they introduced a greedy algorithm based on the submodularity property [15] to approximate an optimal solution. A similar problem with heterogeneous edge weights was formulated into a convex optimization problem, and a maximum likelihood method was proposed to solve it [16]. In doing so, noticing that the structure of a social network is sparse, Myers and Leskovec introduced penalty functions into the objective function to improve its accuracy [16]. The same problem was further extended from inferring static network structures to inferring dynamically changing networks, and the effect of a time-varying external influence was integrated into the model [17]. Recently, to infer disease transmission networks at an individual level, Teunis and Heijne defined a pairwise kernel likelihood function to incorporate infection time difference, proximity of cases, and genetic similarity information [18].

Although information diffusion and disease transmission are to a certain extent similar, they have significant practical differences. Information diffusion networks are usually analyzed at an individual level, whereas disease transmission networks are more meaningful and practical if analyzed at a metapopulation level, for the following reasons:

- It is more appropriate to simulate disease transmission in both temporal and spatial scales [19,20].
- It is difficult to simulate complex individual human behavior and collect large amounts of personal information [6,21,22].
- Controlling disease transmission at a metapopulation level is more practical from the view point of front-line practitioners and policy makers [23].

However, the metapopulation approach leads to two additional challenges:

1. Nodes within metapopulation-based disease transmission networks connect not only with each other, but also to themselves, indicating that susceptible people may get infected by infectious people within the same subpopulation.
2. Unlike information diffusion or individual-based disease transmission networks, disease transmission at a metapopulation level does not follow Directed Acyclic Graphs, where if certain individual does not get informed or infected at the first time, he or she will never get informed or infected in the following time period. In contrast, it propagates over Directed Cyclic Graphs. That is to say, a subpopulation may repeatedly get infected as long as it contains susceptible people. In such transmission network, disease proceeds with cyclic loops rather than like a path or branches of trees.

In such a situation, inferring metapopulation-based disease transmission networks is not only desirable but also challenging. Currently, to the best of our knowledge, no such studies exist. Specifically, this research makes the following three contributions:

1. A generalized linear disease transmission model is built, which considers all the possible transmission pathways at a metapopulation level.
2. A machine learning method called NetEpi (Network Epidemic) is developed to infer hidden disease transmission networks using only the spatiotemporal surveillance data.

3. Unlike similar network inference studies which are conducted over Directed Acyclic Graphs, the proposed method addresses the problem over Directed Cyclic Graphs when analyzing real-world situations.

This research is also practically meaningful as it helps to computationally predict the spread of infectious diseases and provides policy makers with new insights with potentially effective intervention strategies [20]. Partial results of this research have been reported in [24,25].

Method

Definitions

Suppose there exists an unknown directed cyclic network G over which an infectious disease transmits, the observed surveillance data can be represented in a tuple of $< id_p, it_p, loc_p >$. p is the index of a reported/confirmed case. id_p represents the unique identity. it_p is the reported infection time. loc_p is the geographical location where the reported/confirmed case p gets infected.

After aggregating infection cases based on locations and infection times, a dataset $D = \{< v_i, ic_i, t_i > \mid i = 0, 1, 2, \ldots N, t \in T\}$ is collected. i is the index of a specific node. v_i corresponds to the unique identity of a geographical location (e.g., a province, a city, a township, or an urban area). ic_i is the aggregated number of infection cases. t_i indicates a time step. T is the considered time period of disease transmission. In this research, given only the observed data D, the underlying disease transmission network G is inversely inferred. The estimated disease transmission network is referred to as G^*.

Definition 1. Disease Transmission Network: Graph $G =< V, E >$ is a directed cyclic network where $V = \{v_i \mid i = 0, 1, 2, \ldots, N\}$ is the set of nodes. The node v_0 represents the source node of the imported cases that would potentially cause local epidemics (the imported cases for a disease can be defined as the laboratory-confirmed infection cases where people have traveled to disease endemic regions or countries within days before the onset of the disease [26]). v_i $(i = 1, 2, \ldots N)$ correspond to the rest of nodes within the target region. $E = \{\boldsymbol{e}_i \mid i = 1, 2, \ldots, N\}$ denotes the set of directed edges with different weights $W = \{\boldsymbol{w}_i \mid i = 1, 2, \ldots, N\}$. $\boldsymbol{e}_i = \{e_{ji} \mid j = 0, 1, 2, \ldots, N\}$ is the set of incoming links for node i and $\boldsymbol{w}_i = \{w_{ji} \mid j = 0, 1, 2, \ldots, N\}$ is the corresponding weight vector. To be noticed, the source node v_0 does not have incoming links. The physical meanings of these edges that have non-zero weights can be understood as the generalized transmission pathways that *temporally correlate* subpopulations in terms of their infection observations.

Unlike the network structures used in previous studies, the network structures used in this research contain three types of transmission pathways (shown in Figure 1). As the data describes a real-world situation, the assumption is that infected people can infect susceptible people within the same subpopulation (shown in Figure 2). This type of transmission pathway is defined as the internal transmission component. In addition, subpopulations within metapopulation-based disease transmission networks can be affected not only by subpopulations located in adjacent geographical regions, but also by imported cases. We define them respectively as the neighborhood transmission component and the external influence component.

Definition 2. Internal Transmission Component: Within each node (subpopulation), previously infected people may correlate to newly infected people without outside disturbances. This component is disease independent.

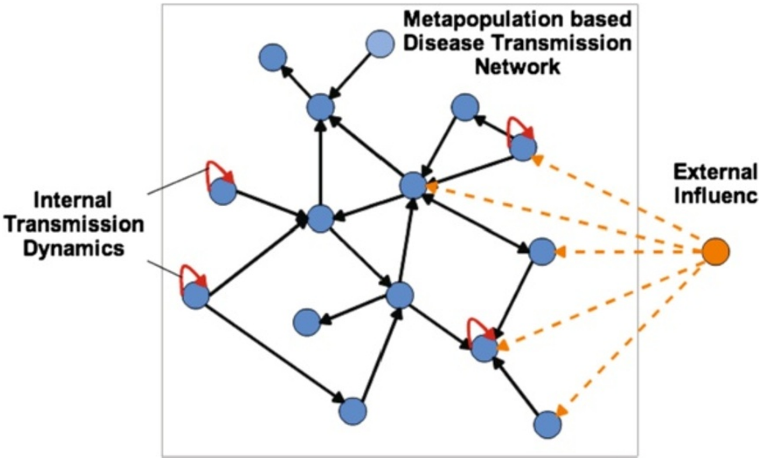


Figure 1 An illustration of three types of transmission pathways contained in our considered disease transmission networks. The internal transmission component is labeled with red solid links connecting to the nodes themselves. The neighborhood transmission component is labeled with black solid links between nodes within the metapopulation-based disease transmission network. The external influence component is introduced as dashed orange links (an external node connects to all the other nodes; for the sake of presentation, we draw only a proportion of them).

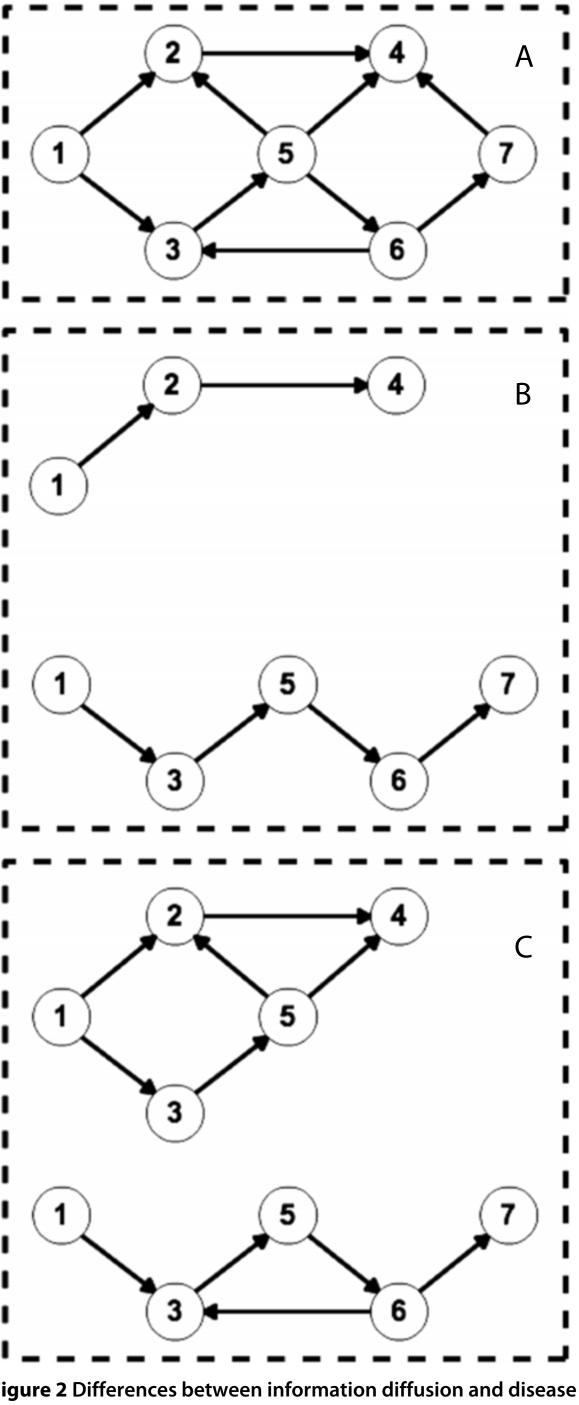


Figure 2 Differences between information diffusion and disease transmission over the same directed cyclic network. (A) shows the example of a ground-truth synthetic network. **(B)** shows two independent information cascading or individual-based disease transmission processes where no cycle exists in these processes. **(C)** shows two independent disease transmission processes at a metapopulation level.

Air-borne diseases such as influenza, vector-borne diseases such as malaria, and other infectious diseases all have this property. It is represented as an edge linking to itself with weight w_{ii} for each node i in the disease transmission network G.

Definition 3. Neighborhood Transmission Component: Among groups of nodes (subpopulations), the temporal correlations among infected people could be caused by physically connected highways, air travels, adjacent borders, etc. This component signifies the interactions happening between infected people in different subpopulations. In G, it is represented as a directed link e_{ij} from nodes i to j with weight w_{ij}, indicating the correlations between infected people in both i and j.

Definition 4. External Influence Component: In disease transmission, the imported cases from foreign or distant endemic countries and regions are another major factor that can cause local epidemics [27]. Thus, we consider this factor in the disease transmission network as an external node connected to all the other nodes. In G, this is denoted as an edge e_{0i} from external node to node i with weight w_{0i}.

Linear transmission model

To characterize a disease transmission process over G, we integrate both of the internal transmission component and the external influence component with the neighborhood transmission component. The internal transmission component characterizes the possible transmission relationships between previously infected people and current infected people within each subpopulation. The assumption in [19], that "individuals do not change disease states during movements" is retained. Thus the neighborhood transmission component describes the temporal correlations between infected people in different subpopulations. The external influence component depicts the introduction of the imported cases from external sources. The above three types of transmission pathways are defined in mathematical forms, respectively, as follows:

$$\begin{cases} itc_i{}^t = w_{ii} \times ic_i{}^{t-1} \\ ntc_i{}^t = \sum_{j}^{N_i} w_{ji} \times ic_j{}^{t-1} \\ eic_i{}^t = w_{0i} \times ic_0{}^{t-1}, \end{cases} \tag{1}$$

where $itc_i{}^t$, $ntc_i{}^t$, and $eic_i{}^t$ refer to the number of infection cases from the internal transmission, neighborhood transmission, and external influence components of node i ($i \neq 0$) at time step t, respectively. N_i is the number of the neighbors of node i. w_{ii}, w_{ji}, and w_{0i} are the corresponding edge weights. ic_i is the total number of infection cases in node i, which can be written as a linear combination of the above three components plus an error term ε. ε is used to capture unpredicted biases. The assumption is that the infection number for each node follows a zero-mean normal distribution, $\varepsilon \sim N(0, \beta)$:

$$\begin{aligned} {ic_i}^t &= {itc_i}^t + {ntc_i}^t + {eic_i}^t + \varepsilon \\ &= w_{ii} \times {ic_i}^{t-1} \\ &\quad + \sum_{j}^{N_i} w_{ji} \times {ic_j}^{t-1} \\ &\quad + w_{0i} \times {ic_0}^{t-1} \\ &\quad + \varepsilon. \end{aligned} \tag{2}$$

Equations 1 and 2 characterize the temporal dynamics of the infection cases at each location. Note that in the real world, once a reported/confirmed case is diagnosed, the physicians or hospitals would take necessary treatment and intervention measures, for example, medication or isolation. Thus, in the above linear transmission model, the infection cases at the current time step would be set to be isolated in the subsequent time steps.

Network inference problem

The network inference problem to be solved here is how to inversely infer the existence of the edges within the hidden disease transmission network G and their corresponding weights $W = \{\boldsymbol{w}_i \mid i = 0, 1, 2, \ldots, N\}$, given an observed surveillance dataset $D = \{< v_i, ic_i, t_i > \mid i = 0, 1, 2, \ldots, N, t \in T\}$. Since the disease transmission process at the metapopulation level does not follow the Directed Acyclic Graphs pattern (Figure 2), it would be inaccurate to infer disease transmission networks following the cascading process in the information diffusion [14].

To recover the network structure G, it is necessary to first write the likelihood function for a specific node i based on Eq. 2:

$$\mathcal{L}(\boldsymbol{w}_i, \beta | ic_i) = \prod_{t=1}^{T} \frac{1}{(2\pi\beta)^{(1/2)}} e^{-\frac{1}{2\beta}\varepsilon^*}, \tag{3}$$

where all the parameters are the same as those in Eq. 2, except we use $\varepsilon^* = \left(ic_i^t - w_{ii} \times ic_i^{t-1} - \sum_j^{N_i^*} w_{ji} \times {ic_j}^{t-1} - w_{0i} \times {ic_0}^{t-1} \right)$, and N_i^* rather than N_i, to indicate the number of estimated neighbors of node i within the inferred network G^*. β is the variance of the normal distribution for the error term ε. Based on this equation, we transform the network inference problem into an optimization problem, which is to find the optimal combination of neighbors with accurate weights for a specific node i.

Then for the entire estimated network G^*, the objective is to maximize the likelihood function:

$$\mathcal{L}\left(W, \beta | D\right) = \prod_{i=1}^{N} \mathcal{L}(\boldsymbol{w}_i, \beta | ic_i)\,, \tag{4}$$

To evaluate the estimated network G^*, we will use precision-recall measures. Specifically, we will compare both the existences of edges and their corresponding weights in the synthetic network G and the estimated network G^*.

Partial correlation network construction

Because there could be many combinations for a node to form its neighborhood, the solution space for the above problem is huge. At the first step, we plan to reduce this space in order to improve both accuracy and performance for further tuning.

When using the Pearson correlation to analyze the correlation between two selected nodes i and j, a problem arises in the analysis of disease transmission networks. As shown in Figure 3(A), disease transmission may follow a path from node i to k, then to j. Take nodes i and j as our analysis targets. Although they are not directly connected, and the overall time-series surveillance data exhibits time delay, they may still be correlated. Therefore, in the approximate network structure G^p, they may be connected. The same problem exists in the case of Figure 3(B), where both nodes i and j are the children of node k in the disease transmission process. The correlation between nodes i and j is still strong even though the weights w_{ki} and w_{kj} are very different. To solve the biases produced by the intermediate node and the sharing of the same parent node, a first-order partial correlation analysis is carried out.

The first-order partial correlation is a measurement of the dependence between two variables X and Y, after removing or fixing a third variable Z. In our case, to compute it between nodes i and j, the effect of another node k, where $k = 0, 1, 2, \ldots, N$, and $k \neq i, j$ is sequentially removed or fixed. From the results, only those coefficients that indicate strong correlations with significant p-values are chosen. It should be mentioned that a partial correlation analysis usually does not provide edge direction information [28,29]. Therefore, to infer a directed

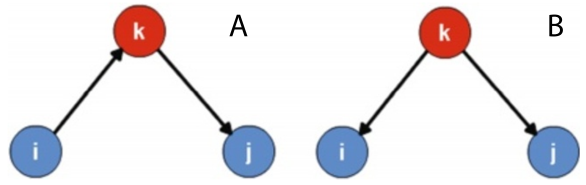


Figure 3 The possible transmission relationships among three nodes [28]. The blue ones are the target nodes for which we aim to identify their relationships. The red nodes are the intermediate nodes. **(A)** shows no directed edge between nodes *i* and *j*. The disease transmission follows a path from node *i* to the intermediate node *k*, then to the target node *j*. **(B)** shows that node *k* transmits to nodes *i* and *j*, simultaneously and independently.

relationship, in this research, we analyze the partial correlation with a time lag. The physical meaning of the time lag is a time step during the disease transmission process (e.g., one day, one week, or one month). Here, we use a time lag of one unit as example, but the time lag is not limited to one unit, other options are also allowed. The direction is defined as from the node using the previous-time-step time-series data to the node using the current-time-step time-series data. Defining the partial correlation coefficient between nodes i and j after fixing the variable of node k as $\rho_{ij.k}$, it can be computed as follows:

$$\rho_{ij.k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}}, \tag{5}$$

where ρ_{ij}, ρ_{ik} and ρ_{jk} are the covariances between each pair of node i, j and k respectively.

Back-tracking Bayesian learning

Given the partial correlation network G^p, an approximate disease transmission network structure is obtained that contains possible neighbors for each node. However, some edges in G^p still do not exist in the synthetic network G. A possible solution is to set the weights of these false positive edges within G^p as zero during the inference process. This is similar to the procedure of removing irrelevant basis components, which is the basis for dimension reduction [30]. In the proposed inference method, the Bayesian learning is based on the Sparse Bayesian Learning (SBL) framework [31]. Related work has been widely and well reported in signal processing studies [30]. To be noticed, if two components are similar, SBL only chooses one of them in order to compress the relevant information. However, in our case, even two nodes are similar, we aim to find both of them.

For a specific node i, the preprocessed surveillance dataset D is divided into two subsets: an $M \times 1$ vector of $\boldsymbol{y} = \{< v_i, ic_i, t_i > \mid t_i = 2, 3, \ldots, M+1, M \in T\}$ and a $M \times |N^p|$ matrix of $\boldsymbol{x} = \{< v_j, ic_j, t_j > \mid j \in N^p, t_j = 1, 2, \ldots, M, M \in T-1\}$. M is the size of output variable $\boldsymbol{y}$ and input variable $\boldsymbol{x}$. N^p represents the indices of the possible neighbors that node i has based on G^p. $T-1$ is the previously considered time period of disease transmission. For the sake of presentation, in the following, we omit the index i for $\boldsymbol{y}$, $\boldsymbol{x}$, and other parameters. If not specifically stated, all the parameters are formulated for node i. Here, we use a time lag of 1 between $\boldsymbol{y}$ and $\boldsymbol{x}$.The relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ can be formulated based on the generalized linear transmission model introduced earlier as follows:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{w} + \varepsilon, \tag{6}$$

where $\boldsymbol{w} = \{w_j \mid j \in N^p\}$ is a vector indicating the weights of all possible incoming links estimated based on G^p. ε is an error term. As mentioned earlier, the solution space is huge. Thus we hope to limit $\boldsymbol{w}$ within a smooth range. Here we follow the framework of SBL, and let both $\boldsymbol{w}$ and ε follow a zero-mean Gaussian distribution with variances of $\boldsymbol{\alpha}$ and β, respectively [31]. They are defined as:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{N^p} N\left(w_i|0, \alpha_j^{-1}\right), \tag{7}$$

$$p(\varepsilon) = N(0, \beta). \tag{8}$$

Because there is no prior knowledge of $\boldsymbol{w}$ and ε, it is reasonable to set them with non-informative prior distributions, such as a Gamma distribution. Here, $\boldsymbol{\alpha}$ and β are assumed to have the same hyperparameters for all nodes.

Given the observation data $\boldsymbol{y}$ and the prior distribution $\boldsymbol{\alpha}$ and β, the posterior distribution of $\boldsymbol{w}$ is:

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \beta) &= \frac{likelihood \times prior}{normalizefactor} \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\alpha}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha}, \beta)}{p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)} \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)}, \end{aligned} \tag{9}$$

which is a Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \beta^{-1}\boldsymbol{\Sigma}\boldsymbol{x}^T\boldsymbol{y} \tag{10}$$

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \beta^{-1}\boldsymbol{x}^T\boldsymbol{x}\right)^{-1}, \tag{11}$$

where $\boldsymbol{\Lambda} = diag(\alpha_1, \alpha_2, \ldots, \alpha_{N^p})$. "Type-II maximization likelihood" maximization combined with a maximum a posteriori probability (MAP) estimate [31] transforms the whole problem into the following marginal likelihood function:

$$p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta) = \int p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w}. \tag{12}$$

Writing Eq. 12 into a logarithm form $\mathcal{L}(\boldsymbol{\alpha})$, we have:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}) &= \log p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta) = \log \int p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} \\ &= -\frac{1}{2}\left[M \log 2\pi + \log|\boldsymbol{C}| + \boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}\right] \end{aligned} \tag{13}$$

with

$$\boldsymbol{C} = \beta I + \boldsymbol{x}\Lambda^{-1}\boldsymbol{x}^T. \tag{14}$$

The derivatives of Eq. 13 with respect to α_j and β are [32]:

$$\frac{\partial\mathcal{L}(\boldsymbol{\alpha})}{\partial \log \alpha_j} = \frac{1}{2}\left(1 - \alpha_j\Sigma_{jj} - \alpha_j\mu_j^2\right) \tag{15}$$

$$\frac{\partial\mathcal{L}(\boldsymbol{\alpha})}{\partial \log \beta} = \frac{1}{2}\left[\frac{M}{\beta} - \|\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\mu}\|^2 - trace\left(\boldsymbol{\Sigma}\boldsymbol{x}^T\boldsymbol{x}\right)\right]. \tag{16}$$

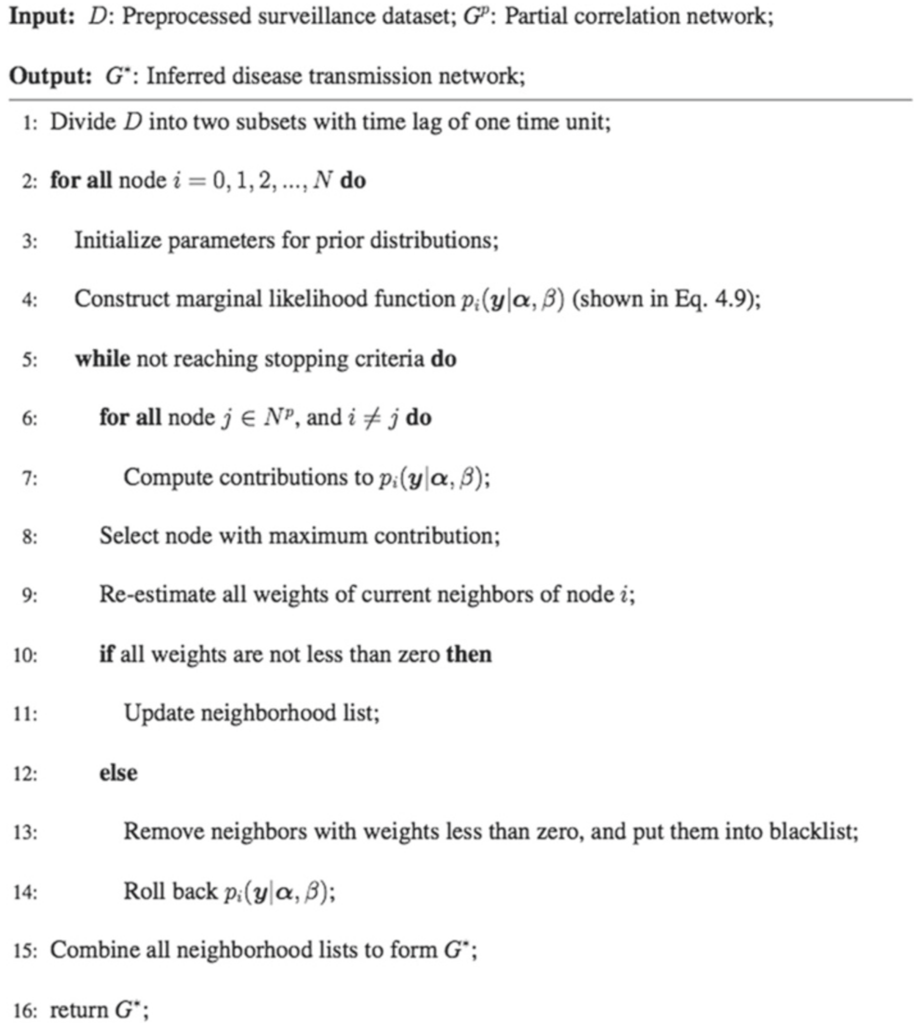
Setting Eqs. 15 and 16 to zero, the estimations of α_j and β become:

$$\alpha_j^{new} = \frac{1 - \alpha_j \Sigma_{jj}}{\mu_j^2} \tag{17}$$

$$\beta^{new} = \frac{M - \sum_{j=1}(1 - \alpha_j \Sigma_{jj})}{\left\| \boldsymbol{y} - \boldsymbol{x\mu} \right\|^2} \tag{18}$$

The above iterative estimation procedure can be solved by using the Expectation-Maximization. In each iteration, the contributions to the marginal likelihood function are estimated for all the nodes in G^p. The one with the maximum contribution is selected as the candidate neighbor. Its corresponding weight is then computed.

In the disease transmission network G, only positive links indicating the existence of transmission pathways exist. However, the prior distribution shown in Eq. 7 may cause $\boldsymbol{w}$ to be negative. To avoid this, a constraint limiting $\boldsymbol{w}$ to a positive value is introduced. To incorporate this constraint into the framework of the above Bayesian learning, a back-tracking technique is used. During the EM learning procedure, the marginal likelihood function and other parameters are updated sequentially. Consequently, each time $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, α_j, and β are updated, any α_j that fail the constraint are selected out, and their corresponding indices are put into a blacklist. The learning procedure is then rolled back, including the marginal likelihood value, to the previous step, and proceeds by selecting only nodes that do not appear in the blacklist, while at the same time maximizing the marginal likelihood function. The algorithm for the Back-Tracking Bayesian Learning is shown in Figure 4.



```
Input: $D$: Preprocessed surveillance dataset; $G^p$: Partial correlation network;
Output: $G^*$: Inferred disease transmission network;
Divide $D$ into two subsets with time lag of one time unit;
for all node $i = 0, 1, 2, ..., N$ do
    Initialize parameters for prior distributions;
    Construct marginal likelihood function $p_i(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)$ (shown in Eq. 4.9);
    while not reaching stopping criteria do
        for all node $j \in N^p$, and $i \neq j$ do
            Compute contributions to $p_i(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)$;
        Select node with maximum contribution;
        Re-estimate all weights of current neighbors of node $i$;
        if all weights are not less than zero then
            Update neighborhood list;
        else
            Remove neighbors with weights less than zero, and put them into blacklist;
            Roll back $p_i(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)$;
Combine all neighborhood lists to form $G^*$;
return $G^*$;
```

Figure 4 Algorithm for the Back-Tracking Bayesian Learning.

Results

Experiments based on synthetic data

Three types of Kronecker Graphs [33] are constructed: (i) core-periphery networks, which have a cluster of nodes in the core of the network and other nodes with less connections distributed in the periphery area, (ii) hierarchical community networks in which nodes form several small communities that are connected to form one large cluster, and (iii) random graphs, which have no obvious pattern. For each type of network structure, different scale parameters are set to generate different ground-truth networks: (i) 64 nodes with 100 edges and 150 edges, (ii) 128 nodes with 180 edges and 200 edges, (iii) 256 nodes with 350 edges and 400 edges, and (iv) 512 nodes with 720 edges and 800 edges. The external links and self-connected edges are generated independently for each ground-truth network. For each synthetic network, disease transmission model (Eq. 2) is run ten times to generate independent synthetic datasets. For a single dataset, the transmission process is made to cover all the edges of *G*. In total, there are three types of network topologies × 8 different sizes × 10 independent transmission processes = 240 datasets.

To our best knowledge, there has not been much prior work on inferring network structures over Directed Cyclic Graphs. Therefore, we utilize a probability based baseline method. At two adjacent time steps $t = n$ and $t = n + 1$, all the nodes that have infection cases at $t = n$ will have directed connections to those nodes that have infection cases at $n + 1$ (shown in Figure 5). The edge weight is affected by both the number of infection cases and the number of infected nodes in the previous time step. The top *k* edges with the highest weights are selected, and the estimated disease transmission network G^* is constructed accordingly. The mathematical formula to compute the baseline edge weight is as follows:

$$w_{ij} = \frac{{ic_i}^t {ic_j}^{t+1}}{\sum_{i=1}^{N} {ic_i}^t}. \tag{19}$$

To evaluate the inference results, the precision-recall curves are computed as shown in Figure 6. Similar to the definitions in [14], the precision is defined as “what fraction of edges in G^* is also present in *G*”, and the recall is defined as “what fraction of edges of *G* appears in G^*”. For two nodes *i* and *j*, if both the ground-truth edge e_{ij} and the inferred edge e^*_{ij} exist, and the difference in their corresponding weights $|w_{ij} - w^*_{ij}|$ is less than a predefined threshold, we say the inferred edge is accurate. In our experiments, NetEpi outperforms the baseline method in all 240 datasets.

For a specific node in the disease transmission network, NetEpi treats all the other nodes homogeneously and independently. That is to say, the connections between two nodes *i* and *j* are only affected and estimated by the time-series surveillance data of these two nodes. This exactly satisfies the real-world requirements discussed above. The underlying network topology is not taken into account during the inference procedure. For networks that have same sizes but different topologies, NetEpi performs best on the core-periphery networks.

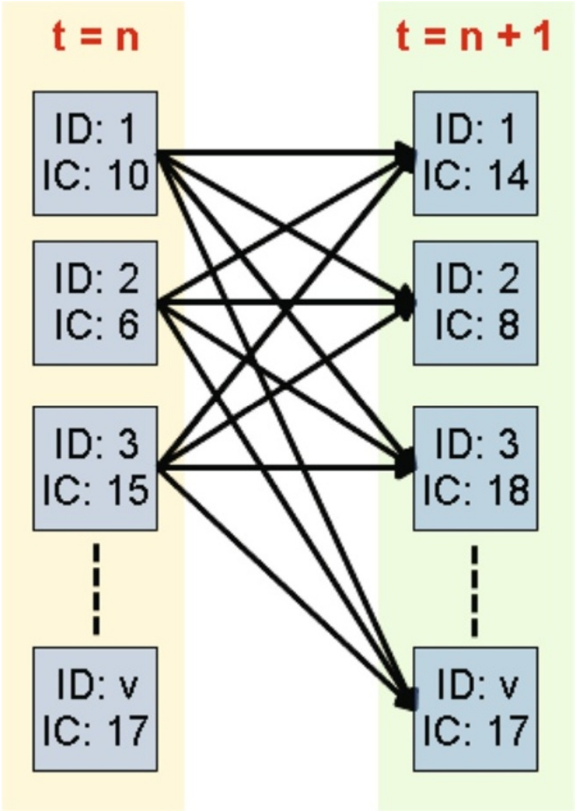


Figure 5 An overview of the baseline method. Nodes (represented as squares) have infection cases at time step $t = n$, and have probabilities of infecting nodes that have infection cases at time step $t = n + 1$. This is shown as the dashed lines in the figure. The ID notation represents the unique identity number of each node. The IC notation represents the number of infection cases at the current time step.

In core-periphery networks, nodes are located in the core region. These nodes have more connections than those distributed in the periphery region. Therefore, to achieve an optimal solution, core-located nodes will have higher probabilities of possessing many neighborhood combinations. In other words, the probability of finding a globally optimal solution for such nodes will decrease as the number of their incoming edges increases. The accuracy of NetEpi over networks with core-periphery topology is consequently biased by the tradeoff between core-located nodes and periphery-located nodes. In comparison, networks with a hierarchical communities

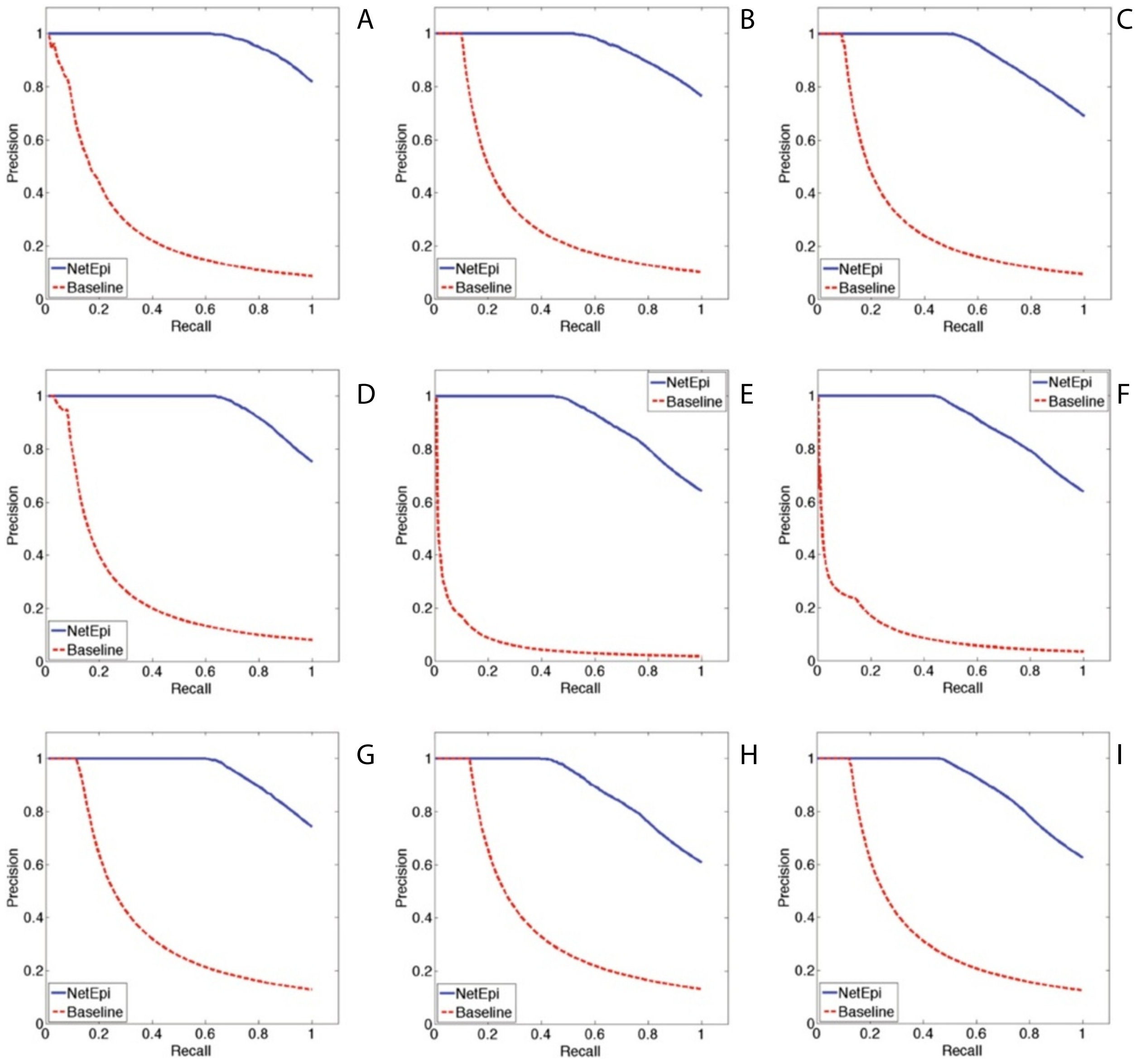


Figure 6 The precision-recall curves for synthetic networks. We test NetEpi over 24 different networks, and we select nine of them here for illustrations. **(A) - (C)** show core-periphery networks with size of 64 nodes with 100 edges, 128 nodes with 180 edges, and 256 nodes with 350 edges. **(D) - (F)** show hierarchical community networks with size of 64 nodes with 100 edges, 128 nodes with 180 edges, and 256 nodes with 350 edges. **(G) - (I)** show random graphs with size of 64 nodes with 100 edges, 128 nodes with 180 edges, and 256 nodes with 350 edges. We compare the performances of NetEpi with the baseline method. To be noticed, NetEpi outperforms the baseline method in all datasets.

topology do not have single cores. The single core is divided into several sub-cores that individually form sub-communities. This structure increases the average number of combinations for each node and directly affects the inference accuracy. As for the random graphs, no matter where the nodes are located, their number of connections does not have a fixed pattern. Consequently, NetEpi achieves oscillating results, which means the precision-recall results for random graphs are sometimes the best, and sometimes the worst.

Here, the out-degree is used to illustrate the accuracy differences between networks with different topologies. It is defined as follows:

$$d_{avg} = \frac{\sum_{i=1}^{N} d_i}{N}, \tag{20}$$

where d_i is the out-degree for node i and d_{avg} is the average out-degree for the whole network. The out-degree statistics for all the 24 synthetic networks are listed in Table 1.

Table 1 Out-degrees for synthetic networks

Size	Core-periphery network	Hierarchical community network	Random graph
64 nodes, 100 edges	1.4154	1.5385	1.6923
64 nodes, 150 edges	1.7385	1.8308	1.8923
128 nodes, 180 edges	1.3876	1.4496	1.4651
128 nodes, 200 edges	1.5504	1.6357	1.5969
256 nodes, 350 edges	1.3619	1.5175	1.5097
256 nodes, 400 edges	1.5525	1.6537	1.6615
512 nodes, 720 edges	1.4016	1.5107	1.5029
512 nodes, 800 edges	1.5439	1.6199	1.6686

For networks with the same topologies but a different number of nodes, NetEpi achieves better results when inferring smaller networks, as shown in Figure 6. At the beginning of the inference process, no edge information is given. Therefore, a ground-truth network is treated as a complete network. Even given its approximate structure G^p, the complexity quadratically increases as the number of nodes increases. Meanwhile, as the edge number increases, the number of neighborhood combinations needed for each node to achieve an optimal solution also increases, which directly interferes the inference results, as shown in Figure 7.

However, inferring disease transmission networks at a metapopulation level is different from inferring individual-based information diffusion networks from the perspective of network size. Network size is usually small when calculated at the administrative level (e.g., province and township levels). For example, for a global epidemic disease, WHO publishes statistical reports at the country level (e.g., dengue and malaria [34] (two types of neglected tropical disease that transmit between human and mosquitoes)); for an infectious disease such as SARS, China reports at a province level, on a daily basis. One possible method for inferring large-size networks that cross several levels is to perform hierarchical clustering. NetEpi begins the analysis at the highest level, where each node represents a cluster of lower-level nodes. Then, within each higher-level node, NetEpi can be performed again to infer the lower-level transmission networks. This whole process can be repeatedly and sequentially conduced to get a whole picture of large-size networks.

Experiments show that all the predicted epidemic trends that occur in the ground-truth networks are captured by the inferred networks, no matter how large the networks are. Because of the space limitation, here we show some examples in Figure 8. This confirms that NetEpi converges to a optimal solution, although it may not be the global one.

Experiments based on real-world data

The real-world dataset was provided by the Chinese Center for Disease Control and Prevention. It contains the reported malaria infection cases in Yunnan province, China. Two types of cases, infected by two distinct types of malaria parasites, (*Plasmodium falciparum, and Plasmodium vivax*), are mixed together. Here, the focus is on *Plasmodium vivax*, which is the dominant type in the Yunnan region. There were 2928 cases reported in 51 townships in 2005. These townships are distributed along the border between China and Myanmar (a high malaria-endemic country). The data are preprocessed by merging those cases reported in the same townships and filtering out those infected with another type of malaria parasite that is not the focus in this research. These townships

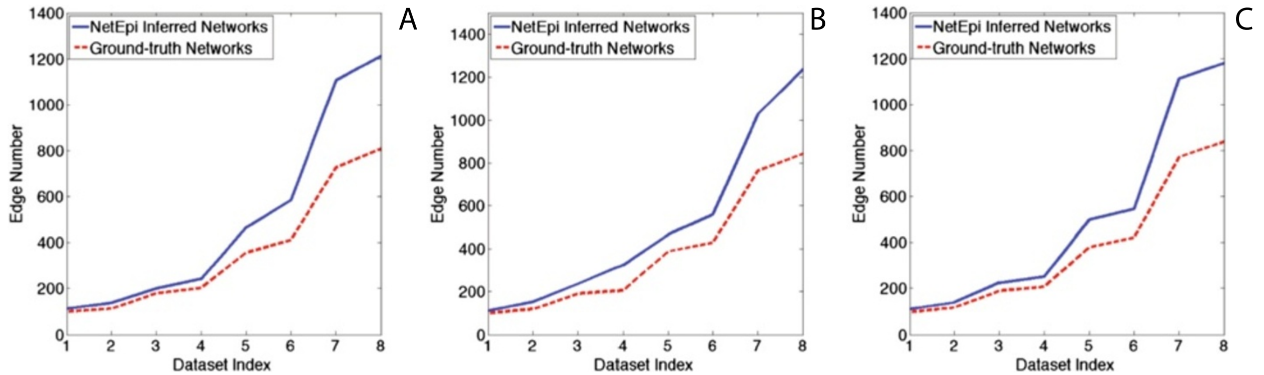


Figure 7 Differences between the edge number in the inferred networks and the ground-truth networks. For each dataset index, we take the average of all the networks with different topologies but same size. The network size increases as the index increases. **(A) - (C)** show the results of core-periphery networks, hierarchical community networks, and random graphs respectively. It is obvious that as the ground-truth network size increases, the accuracy of NetEpi decreases. The number of false edges increases as well. This results from the increased number of possible combinations of neighbors for each node to achieve its global optimal solution.

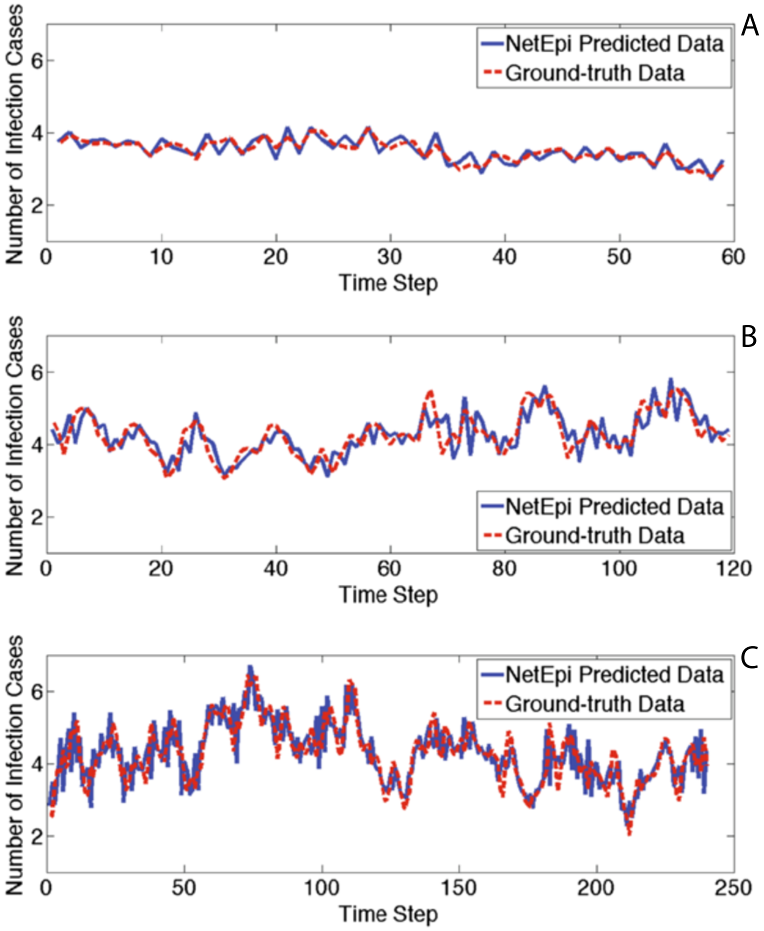


Figure 8 NetEpi accurately captures the disease transmission trend. Three synthetic networks with different topologies and sizes are selected for illustration: **(A)** shows a core-periphery network with 64 nodes and 100 edges. **(B)** shows a hierarchical community network with 128 nodes and 180 edges. **(C)** shows a random graph with 256 nodes and 350 edges.

are further classified into five categories based on their disease severities. They correspond to different numbers of infection cases during the year and are labeled with different colors: $(200, +\infty)$ (red node), $(150, 200]$ (purple nodes), $(100, 150]$ (green nodes), $(50, 100]$ (yellow nodes), and $(0, 50]$ (blue nodes).

The dataset is sparse, with missing data. Moreover, there is no complete labels indicating the imported cases or information about the sources that introduce the imported cases in the original surveillance dataset. Thus, a fixed external node could not be set up during the inference procedure. Like the periodical pattern of the Internal Transmission Component, the External Influence Component also presents regular pattern because of the frequent human mobility motivated by cross-border trade and business. We consequently merge EIC with ITC, and represent either of them, or their combination, by self-connected edges. This is reasonable because it has been recorded that most of these imported cases were due to working, trading, and/or visiting in/with Myanmar regularly. Therefore, self-connected edges are able to capture these regular patterns and identify the imported cases. It is expected that there are many cases imported from neighboring countries, especially Myanmar; therefore, the inferred malaria transmission network contain many self-connected edges. It has been widely reported that the incubation time for *Plasmodium vivax* is $12 \sim 17$ days [35]. However, studies have also reported that the incubation time can be longer, from several months to several years [35,36]. Therefore, in this research, 21 days has been selected as the time window for inferring the hidden malaria transmission network.

In the inferred malaria transmission network, the self-connected edges are labeled with dashed red lines, and edges between neighboring nodes are linked with solid black lines. The width of the edges indicates the strength of the transmission pathways. There are basically two classes of nodes. Some of them connect to themselves, as expected, whereas others form two small communities. In the following, the two types are interpreted separately.

Small Communities: Figures 9 and 10 show that there are two communities in the whole malaria transmission network. The larger one (Figure 9) includes the nodes with the most severe epidemic situations. The severest township, 6, has connections to all the other second-level severity townships (green nodes), indicating that their disease transmission interactions may be the dominant reason for the local malaria endemics in the region. It is obvious that most nodes are connected by highways (e.g., S231, S233, S317 and S318) and rivers. The highways allow infectious patients to move among subpopulations, thus increasing the exposure risk of susceptible populations. The river usually plays a significant role in malaria endemics. It provides a suitable environment for the vector of malaria to reproduce and its flow moves the larva of vector downstream. Therefore, it is possible that the endemics within townships are affected by internal malaria transmission dynamics.

It can readily be noted from Figure 9 that some inferred edges are thicker than others, denoting higher transmission influences (larger edge weights). e_{18-6} (the dash in the index is used for separation) is much thicker than the others, for example, e_{14-6}, e_{4-6}, and e_{28-6}. We interpret this based on Figure 11(A)-(F) in which reported cases are aggregated on an eight-day basis for clear presentation. As shown, although township 18 (Figure 11(E)) has fewer reported cases than other example townships and contains many zero-case intervals, its temporal trend does not significantly violate the trend of township 6 (Figure 11(B)). In comparison, the "mountain-valley-mountain" pattern of township 6 can only be partially matched with other townships (e.g., townships 4 (Figure 11(A)), 14 (Figure 11(D)) and 28 (Figure 11(F))). The influence from township 6 to 4 is much less than that from the reverse direction. This is because the second highest peak appearing between time step 20 to 30 in the trend of township 6 cannot contribute to the valley appeared at the same time interval in the trend of township 4. However, the reverse contribution is reasonable. Intuitively, the pair of townships 4 and 8 (Figure 11(C)), and the pair of townships 14 and 28 have similar trends respectively, but NetEpi only finds edges between townships 14 and 28. This is due to that, for townships 4 and 8, their trends before time step 20 seem to be similar, but those after step 20 present a time lag of around 8*8 days.

As for the small community, it contains townships 1, 41, 49 and 50. The distance between townships 1 and 49 is much longer compared with that between townships 50 and 49. In addition, townships 49 and 50 share the same river. However, the relationships between 49 and 50 are much weaker than those between 1 and 49. It is speculated that townships 1 (Figure 11(G)) and 49 (Figure 11(H)) have the same source of imported cases.

Self-Connected Nodes: As mentioned previously, the external influence component is merged with the internal transmission component. Therefore, these inferred self-connected edges may represent either of these two components, or their combination. Here, we take one group of nodes as an illustrative example. For townships 42, 45, 46, 47, 48, and 51, it is obvious that the endemic disease cases are most likely to be caused by imported cases, because they are located at the border between China and Myanmar (Figure 12). Figure 11(J)-(L) also validate

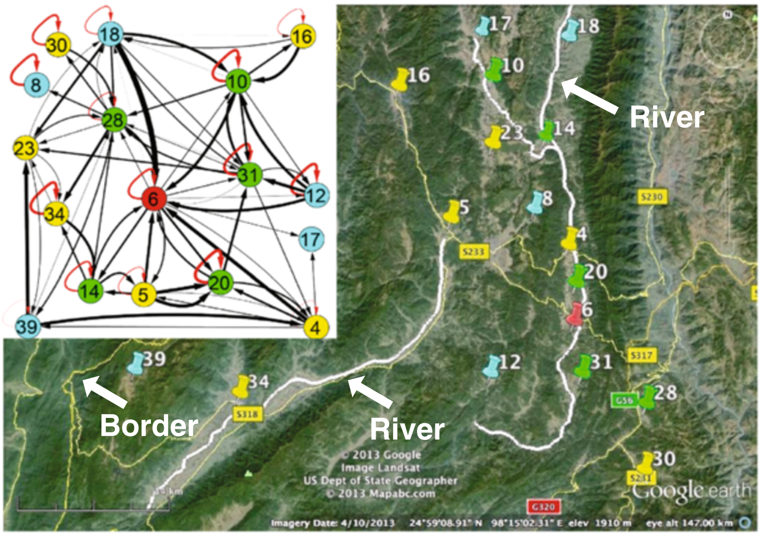


Figure 9 Townships that form a big community as shown in the upper-left subfigure are correlated by their locations that are distributed either in the upstream and downstream of rivers, or close to the highways that connect each other.

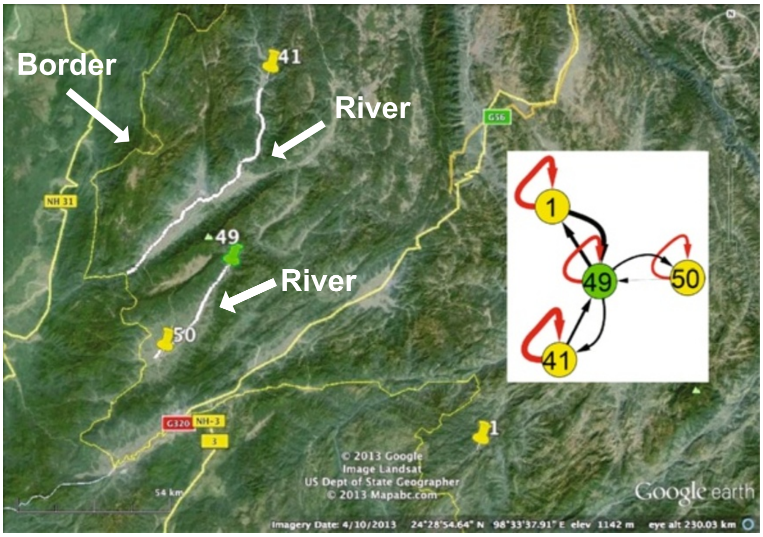


Figure 10 Townships in this figure are located relatively far from each other, except 49 and 50. Their connections may result from sharing the same source of the imported cases.

that their reported cases appear consecutively but are not similar with each other.

There are 47 rather than 51 townships in the inferred malaria transmission network. The four missing nodes have neither self-connected edges nor neighborhood-connected edges. The sum of their infection cases is 81, which is a very small proportion of all the infection cases. Therefore, we think their disease transmission dynamics are primarily accidentally imported cases. It seems that although some townships have similar temporal trends, they are not connected, for example, townships 18 (Figure 11(E)) and 50 (Figure 11(L)). The reason could be the choices of both the time window and the time lag. However, because this real-world dataset is very sparse, it is often difficult to choose the right values. In addition, although some townships are located very close to each other, and on the same rivers, they are not connected within the inferred malaria transmission network; for example, townships 34 and 39 in Figure 9 are not connected because their transmission pathways are not significant or their malaria endemics are mainly affected by the imported cases that disguise the impact of the other factors. To interpret them, currently available information about transportation, rivers, and geographical locations may not be adequate, as the transmission pathways are the *comprehensive results of all impact factors*. Moreover, the roads that are locally formed and managed are not displayed in the map, and they may play significant roles in malaria transmission. Missing reports and data sparsity may also affect the results. However, our method can still detect some hidden connections that may draw the attention of policy makers.

Discussion

There are two key control parameters that play significant roles in the inference results of NetEpi. One is the time window that is used in the partial correlation networks, and the other is the number of observations needed to accurately infer the disease transmission networks. In the following, the influences of those two parameters are discussed individually.

To construct a partial correlation network, it is necessary to select an appropriate time window. Based on a real-world situation, time windows of one day, one week, two weeks, three weeks, one month, five weeks, and one and a half months are selected. In addition, a measurement is defined to evaluate the results:

$$m_{tw} = \left(\frac{s}{|E| + 1} \right)^{\frac{|E^p|}{|E|^2+1}}, \tag{21}$$

where s is the number of edges appearing in both the ground-truth network G and the partial correlation network G^p, $|E|$ is the number of edges in G, $|E^p|$ is the number of edges contained in G^p, and tw refers to the selected time window. It is desirable that G^p contains more edges that appear in G, and less edges that do not appear in G, therefore we use m_{tw} to measure the trade-off between s and $|E^p|$. Based on the experiments and theoretical analysis, the ranges of s and $|E^p|$ are as follows:

$$\begin{cases} s \in [0, |E|] \\ |E^p| \in [|E|, |E|^2] \, . \end{cases} \tag{22}$$

Therefore, the value of m_{tw} increases as s increases or $|E^p|$ decreases. The ideal case is $s = |E^p| = |E|$, so that

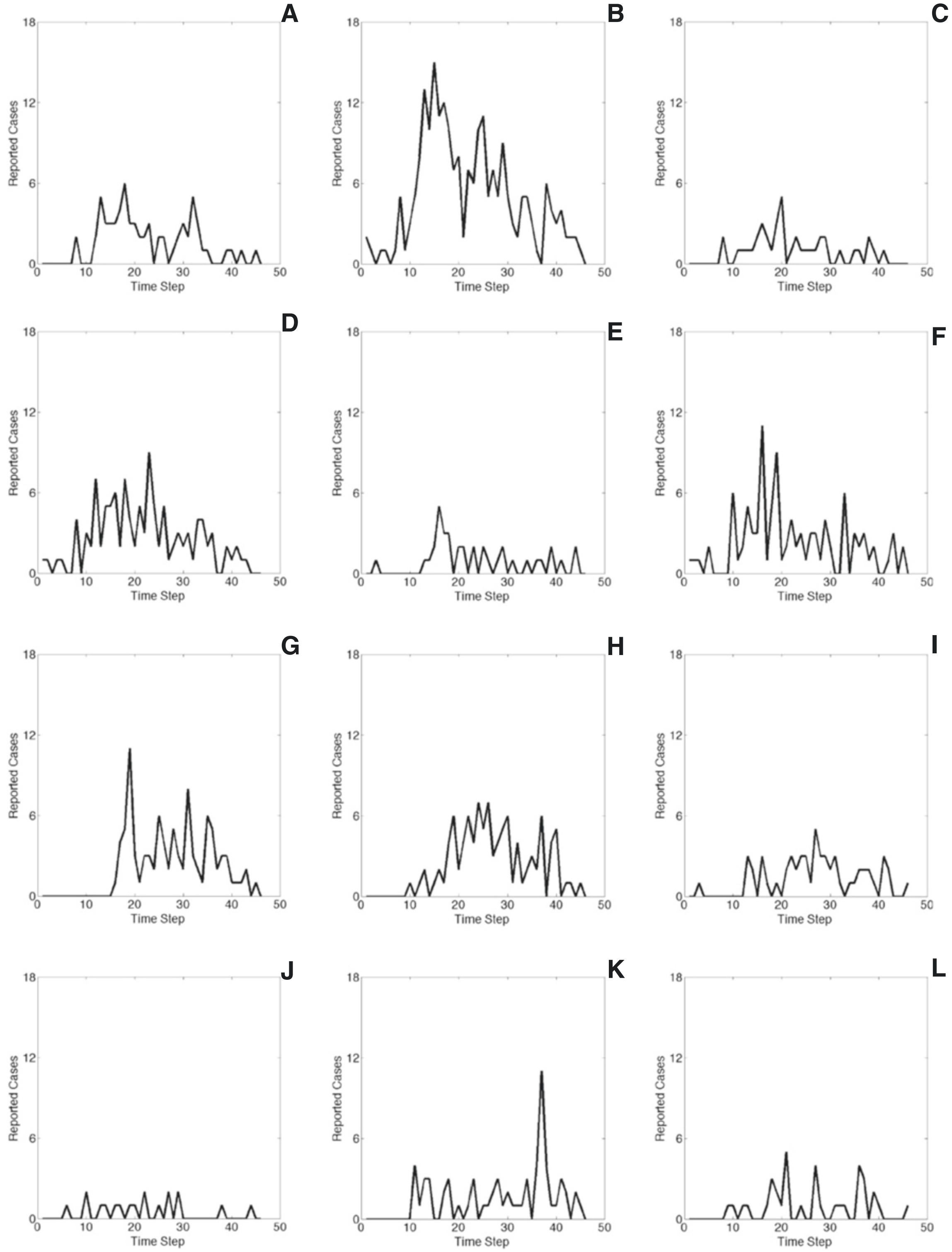


Figure 11 The reported cases for the selected nodes in 2005. In order to present them clearly, we aggregate the reported cases on an eight-day basis. **(A) - (F)** show the curves for townships selected from Figure 9. Their administrative names are Shangying Township, Wuhe Township, Beihai Township, Qushi Township, Jietou Township, and Zhenan Township, respectively. **(G) - (I)** show the curves for townships selected from Figure 10. Their administrative names are Mengdui Township, Qingping Township, and Zhangfeng Township, respectively. **(J) - (L)** show the curves for townships selected from Figure 12. Their administrative names are Tongbiguan Township, Mengyue Township, and Longba Township, respectively.

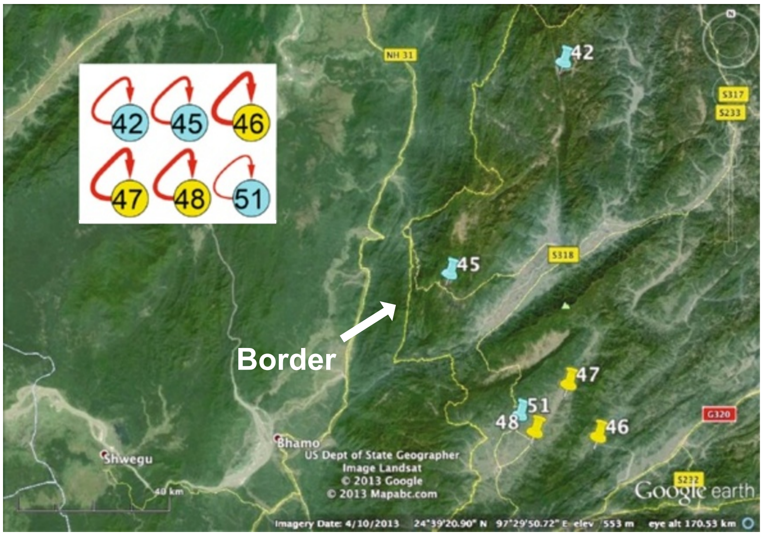


Figure 12 Townships 42, 45, 46, 47, 48 and 51 are located adjacent to the border between China and Myanmar. Therefore, their self-connected edges are possibly due to the imported cases.

m_{tw} significantly approximates 1. It should be noted that when s approximates or equals $|E|$, m_{tw} approximates 1 as well, and even $|E^p|$ is very large (but still much smaller than $|E|^2$). Here such cases are not punished, as finding all the ground-truth edges, or the majority of them, is more important than the constructed partial correlation network with a larger size.

For all the 24 synthetic transmission networks, we take the individual average values of the analyzed results of the 10 independent datasets. Based on the results shown in Figure 13, the relationships between trade-off measurement m_{tw} and time window tw are categorized into four classes.

"N" Shape: Examples of this type of relationship are shown in Figure 13(D). The trade-off measurement value in such case usually achieves the maximum at a time window with less or moderate values, for example, 7 or 14 days. The partial correlation networks also contain fewer edges under such a time window. m_{tw} decreases at the beginning because the increasing rate of s is slow compared to the fast increasing rate of $|E^p|$. m_{tw} gradually increases later because stronger correlations are identified under the conditions of increasing time window values.

"S" Shape: Examples of this type of relationship are Figure 13(A), (B), (C), (F), (H) and (I). As the length of the time window increases, more edges in the ground-truth networks appear in the partial correlation networks. The correlations of these edges are consolidated as the time-series data are smoothed. At a given point, for example, a time window of 14 days, the majority of the strong correlations have been identified, so that even as the length of time window continues to increase, the number of strongly correlated edges remains stable.

"V" Shape: Examples of this type of relationship are Figure 13(E) and (G). In such cases, the trade-off measurement value reaches the maximum at the very beginning ($tw = 1$), then decreases dramatically to a valley, and increases afterwards. The climax at the start is caused by the low values of both s and $|E^p|$. A proportion of ground-truth edges have not been found out when the time window is equal to one day. Moreover, the sizes of the corresponding partial correlation networks are also small. As in the "N" shape, m_{tw} decreases to a valley because the increasing rate of s is slow compared to the fast increasing rate of $|E^p|$. The subsequent increase is the same as in the "N" shape.

Another important control parameter is the number of observations (size of surveillance dataset), which is the parameter M. Intuitively, the more data there are, the better the inferred results should be. However, it is usually difficult to obtain complete and sufficient surveillance data because of missing reports, immature surveillance systems, etc. In addition, although a huge amount of data can be collected, big data still poses challenges for both data storage and data analysis. Consequently, experiments testing the influence of the size of the surveillance dataset on the accuracy of NetEpi are conducted.

If the size of the surveillance dataset is much smaller than the length of the time window that NetEpi uses, the construction of the partial correlation networks will fail. Therefore, this research assumes that the size of surveillance dataset should be at least larger than the length of the time window. Specifically, the detailed relationships

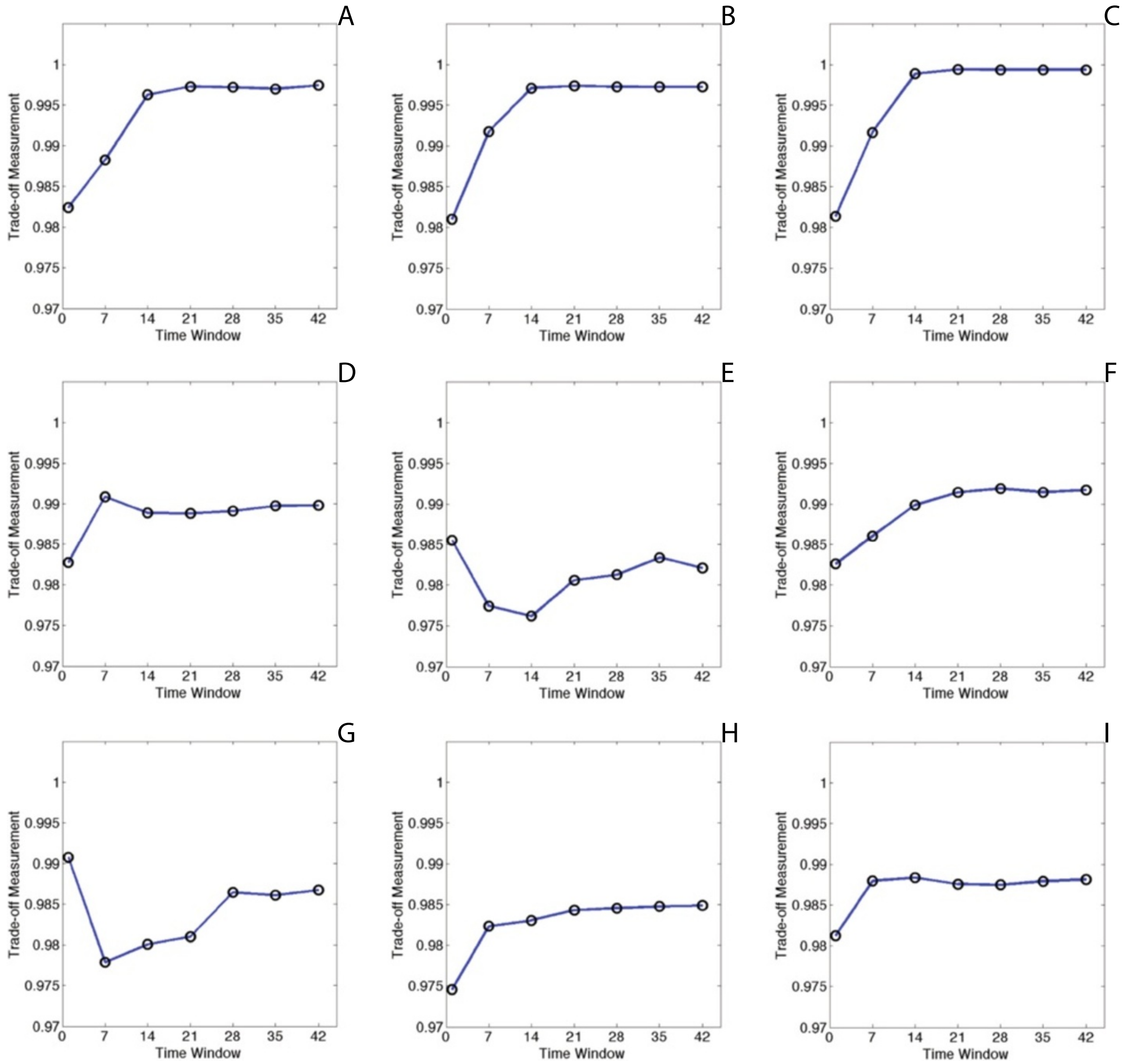


Figure 13 Sensitivity analysis for the choice of time window on synthetic networks with different sizes and topologies. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of m_{tw} computed from Eq. 21. **(A) - (C)** show core-periphery networks with size of 64 nodes with 100 edges, 128 nodes with 180 edges, and 256 nodes with 350 edges. **(D) - (F)** show hierarchical community networks with size of 64 nodes with 100 edges, 128 nodes with 180 edges, and 256 nodes with 350 edges. **(G) - (I)** show random graphs with size of 64 nodes with 100 edges, 128 nodes with 180 edges, and 256 nodes with 350 edges.

between the size of surveillance data M, length of selected time window tw, number of network nodes N and the scale parameter φ should be as follows:

$$M - tw + 1 \geq \frac{N}{\varphi} \tag{23}$$

The left-hand side of the above equation is the size of the time-series dataset after smoothing it under time window tw. The right-hand side is the size of the available surveillance dataset to be tested. Obviously, this criteria guarantees that no matter how long the selected time window is, given a target scale related to the number of network nodes, it is often possible to find a lower bound for the surveillance data that will ensure that the partial correlation analysis is workable. For example, given a network with 128 nodes ($N = 128$) and a time window of 35 ($tw = 35$), if NetEpi is performed when the surveillance dataset is almost half the size ($\varphi = 2$) of the number of

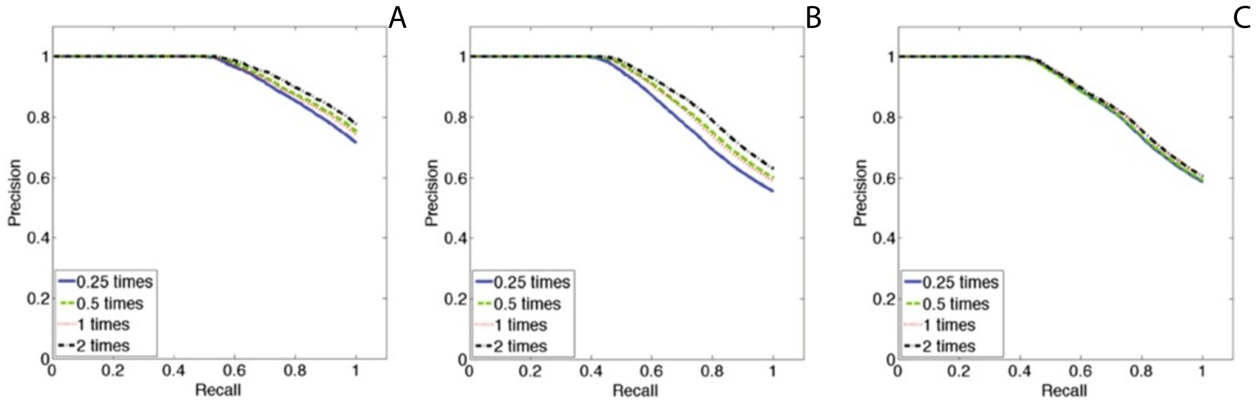


Figure 14 Sensitivity analysis for the choice of observation or surveillance data with different size. (A) - (C) show the results of networks with different topologies but the same size of 128 nodes and 200 edges. Network in **(A)** is a core-periphery network. Network in **(B)** is a hierarchical community network. Network in **(C)** is a random graph. The curve with the size of a quarter of the number of network nodes is displayed as a blue solid line. The curve with the size of a half the number of network nodes is displayed as a green dashed line. The curve with the size of the same number of network nodes is displayed as a red dotted line. The curve with the size of two times the number of network nodes is displayed as a black dash-dot line.

nodes, then the size of the training surveillance dataset should at least be 98.

Figures 14 and 15 show the results of experiments for six networks with different topologies (core-periphery networks, hierarchical community networks, and random graphs) and sizes (128 nodes with 200 edges and 256 nodes with 350 edges). For each network, different sizes of surveillance dataset are tested independently. All of them are tested under the time window of 35. The scale parameter φ is set to equal to 4, 2, 1 and 0.5, as shown in the precision-recall curves with the legends 0.25, 0.5, 1 and 2 times, respectively.

In all these experiments, although less surveillance data may bias the accuracy of NetEpi, the bias is not significantly obvious, even in Figure 15(B), as the missing data is not considered during the generation of the synthetic surveillance data. These results confirm that NetEpi can accurately find and estimate those edges that play important roles in disease transmission, even given minimal surveillance data.

Conclusions

This research bridges the gap between theoretical studies of disease transmission networks and real-world

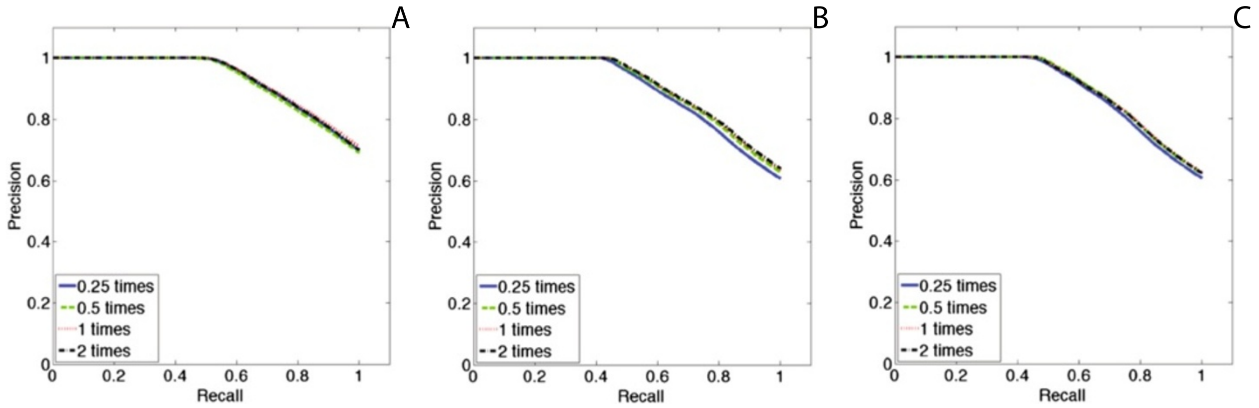


Figure 15 Sensitivity analysis for the choice of observation or surveillance data with different size. (A) - (C) show the results of networks with different topologies but the same size of 256 nodes and 350 edges. Network in **(A)** is a core-periphery network. Network in **(B)** is a hierarchical community network. Network in **(C)** is a random graph. The curve with the size of a quarter of the number of network nodes is displayed as a blue solid line. The curve with the size of a half the number of network nodes is displayed as a green dashed line. The curve with the size of the same number of network nodes is displayed as a red dotted line. The curve with the size of two times the number of network nodes is displayed as a black dash-dot line.

infectious disease transmission, by inversely inferring hidden disease transmission networks using only surveillance data. Specifically, it addresses this problem at a metapopulation level, which is more meaningful and practical for front-line practitioners and policy makers. To achieve this goal, a network inference method called NetEpi is developed. The proposed method and the experimental results provide policy makers with insights into discovering hidden transmission pathways among subpopulations and optimizing limited resources when implementing intervention strategies. In addition, this novel tool can be implemented as a part of surveillance-response system to actively detect and monitor low-transmission patterns [37].

The current version of NetEpi does not consider the detailed impact factors of a specific disease. That is to say, the inferred disease transmission networks are comprehensive and abstract networks that integrate all the impact factors. Taking the inferred malaria transmission network as an example, the inferred edges can be interpreted as geographical locations, convenient traffic routes, suitable habitats for the vector, etc. Therefore, to investigate the transmission pathways in more detail, and to find out the exact interpretations for the inferred edges, it will be necessary to build specific transmission models for different diseases. Moreover, various impact factors should be carefully integrated.

Another direction for future work is to infer dynamic disease transmission networks. Currently, the assumption is that the hidden disease transmission networks do not change within a prefixed time period. However, in reality, the network may change as impact factors change over time. Therefore, inferring dynamic disease transmission networks is useful over a long-time scale, which is also more helpful for policy makers to design long-term intervention strategies.

Finally, the current back-tracking technique rolls back the optimization procedure roughly rather than smoothly, and converges to either the local optimum or the global optimum. Therefore, future work should modify this technique to improve precision.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the experiments: XY JL. Performed the experiments: XY. Analyzed the data and experimental results: XY JL WC XNZ. Contributed reagents/materials/analysis tools: XY JL XNZ. Wrote and revised the paper: XY JL WC XNZ. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge the funding support from Hong Kong Research Grants Council (HKBU211212) and from National Natural Science Foundation of China (NSFC81273192) for the research work being presented in this paper. They would also like to thank the reviewers, the handling editor, and the Editor-in-Chief for their constructive comments on the manuscript.

Author details

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China. [2] National Institute of Parasitic Diseases, China CDC, Shanghai, China.

Received: 18 June 2014 Accepted: 13 October 2014
Published: 17 November 2014

References

1. Eames KTD, Keeling MJ: **Contact tracing and disease control.** *Proc R Soc Lond B Biol Sci* 2003, **270**(1533):2565–2571.
2. Newman M. E: **Spread of epidemic disease on networks.** *Phys Rev E* 2002, **66**(1):016128.
3. Riley S: **Large-scale spatial-transmission models of infectious disease.** *Science* 2007, **316**(5829):1298–1301.
4. Eubank S, Guclu H, Anil Kumar VS, Marathe MV, Srinivasan A, Toroczkai Z, Wang N: **Modelling disease outbreaks in realistic urban social networks.** *Nature* 2004, **429:**180–184.
5. Pastor-Satorras R, Vespignani A: **Epidemic dynamics and endemic states in complex networks.** *Phys Rev E* 2001, **63**(6):066117.
6. Keeling JM, Eames TDK: **Networks and epidemic models.** *J R Soc Interface* 2005, **2**(4):295–307.
7. Salathé M, Jones JH: **Dynamics and control of diseases in networks with community structure.** *PLoS Comput Biol* 2010, **6**(4):1000736.
8. Hollingsworth TD, Ferguson NM, Anderson RM: **Will travel restrictions control the international spread of pandemic influenza?** *Nat Med* 2006, **12**(5):497–499.
9. Sebastian F, Marcel S, Vincent JAA: **Modelling the influence of human behaviour on the spread of infectious diseases: A review.** *J R Soc Interface* 2010, **7**(50):1247–1256.
10. Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A: **Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic.** *PLoS ONE* 2011, **6**(1):16591.
11. Hufnagel L, Brockmann D, Geisel T: **Forecast and control of epidemics in a globalized world.** *Proc Natl Acad Sci U S A* 2004, **101**(42):15124–15129.
12. Liu J, Yang B, Cheung W, Yang G: **Malaria transmission modelling: a network perspective.** *Infectious Diseases Poverty* 2012, **1**(1):1–8.
13. Leventhal GE, Kouyos R, Stadler T, von Wyl V, Yerly S, Böni J, Cellerai C, Klimkait T, Günthard HF, Bonhoeffer S: **Inferring epidemic contact structure from phylogenetic trees.** *PLoS Comput Biol* 2012, **8**(3):1002413.
14. Gomez-Rodriguez M, Leskovec J, Krause A: **Inferring networks of diffusion and influence.** In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '10*. New York, NY, USA: ACM; 2010:1019–1028.
15. Kempe D, Kleinberg J, Tardos E: **Maximizing the spread of influence through a social network.** In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '03*. New York, NY, USA: ACM; 2003:137–146.
16. Myers S, Leskovec J: **On the convexity of latent social network inference.** In *Advances in Neural Information Processing Systems 23*. Edited by Lafferty J, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A: Curran Associates, Inc.; 2010:1741–1749.
17. Myers SA, Zhu C, Leskovec J: **Information diffusion and external influence in networks.** In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'12*. New York, NY, USA: ACM; 2012:33–41.
18. Teunis P, Heijne JCM, Sukhrie F, van Eijkeren J, Koopmans M, Kretzschmar M: **Infectious disease transmission as a forensic problem: Who infected whom?** *J R Soc Interface* 2013, **10**(81):20120955. doi:10.1098/rsif.2012.0955.
19. Arino J: **Diseases in metapopulations.** In *Modeling and Dynamics of Infectious Diseases. Series in Contemporary Applied Mathematics, Volume 11*. Edited by Ma Z, Zhou Y, Wu J. Singapore: World Scientific; 2009:65–123.
20. Colizza V, Vespignani A: **Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations.** *J Theor Biol* 2008, **251**(3):450–467.
21. Ajelli M, Goncalves B, Balcan D, Colizza V, Hu H, Ramasco J, Merler S, Vespignani A: **Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models.** *BMC Infect Dis* 2010, **10**(1):190.

22. Lentz HHK, Selhorst T, Sokolov IM: **Spread of infectious diseases in directed and modular metapopulation networks.** *Phys Rev E* 2012, **85**(6):066111.
23. Ndeffo Mbah ML, Gilligan CA: **Resource allocation for epidemic control in metapopulations.** *PLoS ONE* 2011, **6**(9):24577.
24. Yang X, Liu J, Cheung WKW, Zhou X-N: **Inferring metapopulation based disease transmission networks.** In *Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, Volume 8444*: Springer International Publishing; 2014:385–399. [http://dx.doi.org/10.1007/978-3-319-06605-9_32].
25. Yang X: *Inferring disease transmission networks*. Hong Kong: Baptist University; 2014. Master's thesis.
26. Shang C-S, Fang C-T, Liu C-M, Wen T-H, Tsai K-H, King C-C: **The role of imported cases and favorable meteorological conditions in the onset of dengue epidemics.** *PLoS Negl Trop Dis* 2010, **4:**775.
27. Dénes A, Kevei P, Nishiura H, Röst G: **Risk of infectious disease outbreaks by imported cases with application to the european football championship 2012.** *Int J Stochastic Anal* 2013, **2013:**.
28. Yuan Y, Li C-T, Windram O: **Directed partial correlation: Inferring large-scale gene regulatory network through induced topology disruptions.** *PLoS ONE* 2011, **6**(4):16835.
29. Lasserre J, Chung H-R, Vingron M: **Finding associations among histone modifications using sparse partial correlation networks.** *PLoS Comput Biol* 2013, **9**(9):1003168.
30. Wipf DP, Rao BD: **Sparse bayesian learning for basis selection.** *IEEE Trans Signal Process* 2004, **52**(8):2153–2164.
31. Tipping ME: **Sparse bayesian learning and the relevance vector machine.** *J Mach Learn Res* 2001, **1:**211–244. [http://dx.doi.org/10.1162/15324430152748236]
32. Tzikas DG, Likas CL, Galatsanos NP: **Sparse bayesian modeling with adaptive kernel learning.** *IEEE Trans Neural Netw* 2009, **20**(6):926–937.
33. Leskovec J, Faloutsos C: **Scalable modeling of real graphs using kronecker multiplication.** In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*. New York, NY, USA: ACM; 2007:497–504.
34. WHO: *World Malaria Report 2012*: World Health Organization; 2012.
35. Brasil P, de Pina Costa A, Pedro R, da Silveira Bressan C, da Silva S, Tauil P, Daniel-Ribeiro C: **Unexpectedly long incubation period of plasmodium vivax malaria, in the absence of chemoprophylaxis, in patients diagnosed outside the transmission area in brazil.** *Malaria J* 2011, **10**(1):122.
36. Hulden L, Hulden L, Heliovaara K: **Natural relapses in vivax malaria induced by anopheles mosquitoes.** *Malaria J* 2008, **7**(1):64.
37. Zhou X.-N, Bergquist R, Tanner M: **Elimination of tropical disease through surveillance and response.** *Infectious Diseases Poverty* 2013, **2**(1):1.

doi:10.1186/2047-2501-2-8
Cite this article as: Yang *et al.*: Inferring disease transmission networks at a metapopulation level. *Health Information Science and Systems* 2014 **2**:8.