Inferring Metapopulation Based Disease Transmission Networks

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

[1] Department of Computer Science, Hong Kong Baptist University
Hong Kong SAR, China
{xfyang09,jiming,william}@comp.hkbu.edu.hk
[2] National Institute of Parasitic Diseases, China CDC, Shanghai, China
zhouxn1@chinacdc.cn

Abstract. To investigate how an infectious disease spreads, it is most desirable to discover the underlying disease transmission networks based on surveillance data. Existing studies have provided some methods for inferring information diffusion networks, where nodes correspond to individual persons. However, in the case of disease transmission, to effectively develop intervention strategies, it would be more realistic and reasonable for policy makers to study the diffusion patterns at the metapopulation level, that is, to consider disease transmission networks where nodes represent subpopulations, and links indicate their interrelationships. Such networks are useful to: (i) investigate hidden factors that influence epidemic dynamics, (ii) reveal possible sources of epidemic outbreaks, and (iii) practically develop and improve strategies for disease control. Therefore, based on such a real-world motivation, we aim to address the problem of inferring disease transmission networks at the metapopulation level. Specifically, we propose an inference method called NetEpi (Network Epidemic), and evaluate the method by utilizing synthetic and real-world datasets. The experiments show that NetEpi can recover most of the ground-truth disease transmission networks based only on the surveillance data. Moreover, it can help detect and interpret patterns and transmission pathways from the real-world data.

Keywords: Network inference, disease transmission networks, metapopulation, Bayesian learning, partial correlation networks.

1 Introduction

Infectious disease transmission has been studied with a network based approach and at an individual level [1]. However, existing studies often assume network structures are given in advance (e.g., air travels for the spread of H1N1 [2]), suggesting that it is possible to know which individual could be infected next. In reality, what is possible to observe is only the spatiotemporal surveillance data, containing infection times and locations of reported infection cases. This data

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 385–399, 2014.

provides no knowledge of the hidden transmission pathways that denote the routes of disease propagation among geographical locations. This real-world situation directly poses a significant challenge to the policy makers in applying intervention strategies at appropriate times and locations. In this regard, inferring disease transmission networks (DTNs) becomes an important and urgent research problem in epidemiological studies, which is our key objective.

The network inference problem has been widely studied in the information diffusion domain and is usually conducted at an individual level. Based on empirical time-series data of when people get informed, the static network inference is transformed into a combinatorial optimization problem [3]. Formulating it as a MAX-k-COVER problem, Rodriguez et al. have proven that selecting the top k edges that maximize the likelihood of the static information diffusion network (IDN) structure is NP-hard. They introduced a greedy algorithm based on the submodularity properties to approximate the optimal solution. Myers and Leskovec formulated a similar problem with heterogeneous edge weights into a convex optimization problem, and proposed a maximum likelihood method to solve it [4]. In addition, having noticed that the structure of a social network is sparse, they introduced penalty functions into the objective function to improve the accuracy. In a recently published study on inferring DTNs at the individual level [5], Teunis and Heijne used a pairwise kernel likelihood function to incorporate disease related information, and trained and applied the model using a real-world dataset collected from a university hospital.

The above work has provided insights into solving network inference problems at an individual level. However, inferring DTNs is more meaningful and practical at a metapopulation level, where nodes and edges represent patches with subpopulations (e.g., cities) and transmission pathways among them (e.g., transportation) rather than individual persons and their pairwise connections (e.g., social contacts). This is due to the considerations of: (i) the appropriateness of simulating disease transmission in both spatial and temporal scales [6], (ii) difficulties in simulating complicated human behaviors and collecting a huge amount of personal information [1], and (iii) the practice of controlling disease transmission from the view point of policy makers [7]. However, this treatment leads to two additional challenges: (i) nodes within metapopulation based DTNs can in addition connect to themselves, indicating susceptible people get infected by infected people within the same subpopulation, and (ii) metapopulation based disease transmission follows Directed Cyclic Graphs (DCGs) rather than Directed Acyclic Graphs (DAGs) as in information diffusion or individual based DTNs. Even if a large proportion of a certain subpopulation is infected, the remaining susceptible persons that have not been temporally infected will still have chances of being infected later.

Inferring metapopulation based DTNs is not only desirable but also challenging. As far as we know, there has not been such work done before. In this paper, we will address this problem, and more specifically, make three contributions: (i) to build a generalized linear disease transmission model that considers all possible transmission pathways at the metapopulation level, (ii) to develop an

inference method, called NetEpi, that infers hidden DTNs based only on the spatiotemporal surveillance data, and (iii) to solve the network inference problem over DCGs rather than DAGs. We believe such work is also practically meaningful since it helps computationally predict large-scale infectious disease spread and provide policy makers with insights into optimizing intervention strategies.

The paper is organized as follows. The metapopulation based DTN inference problem is formulated in Section 2. A two-step inference method (NetEpi) is introduced in Section 3. NetEpi is evaluated by using both synthetic and real-world datasets in Section 4. Finally, we make conclusions in Section 5.

2 Problem Statement

A ground-truth DTN is defined as $G = <V, E>$, where the set of nodes is denoted as $V = \{v_i \mid i = 0, 1, 2, ..., N\}$. i is the index of a specific node. v_0 represents the external source node for the imported cases that would potentially cause local epidemics [8] (Imported cases are the laboratory-confirmed infection cases where people have traveled to disease endemic regions within days before the onset of the disease [8]). v_i $(i = 1, 2, ..., N)$ correspond to the rest of nodes within the target region. $E = \{\boldsymbol{e}_i \mid i = 1, 2, ..., N\}$ denotes the set of edges with weights $W = \{\boldsymbol{w}_i \mid i = 1, 2, ..., N\}$. $\boldsymbol{e}_i$ is the set of incoming links for node i, and $\boldsymbol{w}_i$ is the corresponding weight vector. Source node v_0 has no incoming links. The physical meanings of these edges that have non-zero weights describe the generalized transmission pathways that *temporally correlate* subpopulations in terms of their infection observations. In reality, G cannot be directly obtained. What is often collected is surveillance data, which can be represented as $D = \{<v_i, ic_i, t_i> \mid i = 0, 1, 2, ..., N, t \in T\}$ after aggregating infection cases based on locations and infection times. v_i corresponds to a geographical location (e.g., a city, or a township), ic_i is the aggregated number of infection cases, and t_i indicates a time step. T is the considered time period of disease transmission.

We refer to the estimated DTN as G^*, and consider three types of transmission pathways: (i) internal transmission component (ITC), which indicates that infected people, directly (e.g., in the air-borne disease of influenza) or indirectly (e.g., in the vector-borne disease of malaria), infect susceptible people within the same subpopulation, (ii) neighborhood transmission component (NTC), where disease transmits, through physically connected highways, adjacent borders, etc., among several subpopulations (it signifies the interactions happening between infected people in different subpopulations), and (iii) external influence component (EIC), which represents the source of imported cases from distant endemic regions or countries. In G, it is an external node connected to all the other nodes.

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 total number of infection cases

can be written as a linear combination of the above three components plus an error term ε that captures the unpredicted biases, as follows:

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

where ${itc_i}^t$, ${ntc_i}^t$, and ${eic_i}^t$ refer to the numbers of infection cases from ITC, NTC, and EIC to node i ($i \neq 0$) at time step t, respectively. N_i is the number of the neighbors of node i. The error term ε follows a zero-mean normal distribution, $\varepsilon \sim N(0, \beta)$. Eq. 1 characterizes the temporal dynamics of infection cases at each location. To be noticed, in the real world, once a patient is diagnosed to be infected, treatments and interventions (e.g., medication and isolation) would be taken by the physicians or hospitals. Thus, the infection cases at the current time step would be set to be isolated in the following time steps.

Given an observed surveillance dataset $D = \{<v_i, ic_i, t_i> \mid i = 0, 1, 2, ..., N, t \in T\}$, we intend to infer E of G and their corresponding weights W. The likelihood function for a specific node i based on Eq. 1 is:

$$\mathcal{L}(\boldsymbol{w}_i, \beta | ic_i) = \prod_{t=1}^{T} \frac{1}{(2\pi\beta)^{(1/2)}} e^{-\frac{1}{2\beta}(ic_i^t - w_{ii} \times ic_i^{t-1} - \sum_{j}^{N_i^*} w_{ji} \times {ic_j}^{t-1} - w_{0i} \times {ic_0}^{t-1})} \tag{2}$$

where N_i^* indicates the number of the estimated neighbors of node i within G^*. This set of neighbors can be written as $V_i^* = \{v_j \mid j = 0, 1, 2, ..., N \text{ and } w_{ji} \neq 0\}$. β is the variance of the normal distribution for ε. Therefore, we transform the network inference problem into an optimization problem, which is to find an optimal combination of neighbors with accurate weights for a specific node i. Specifically, to infer network G^*, we aim to maximize the likelihood function, given as:

$$\mathcal{L}(W, \beta | D) = \prod_{i=1}^{N} \mathcal{L}(\boldsymbol{w}_i, \beta | ic_i) \tag{3}$$

3 The Proposed Network Inference Method

3.1 Partial Correlation Network Construction

Given D, we first hope to construct an approximate network structure. It will reduce the trivial computations for our second step as well as filter out a proportion of false positive edges. Using the pearson correlation to build such networks is intuitive but not workable in the case of disease transmission. As shown in Fig. 1(a), disease transmission may follow a path from i to k, then to j. Even though i and j are not directly connected, they may still be correlated. Therefore, in the approximate network structure, denoted as G^p, they may be connected

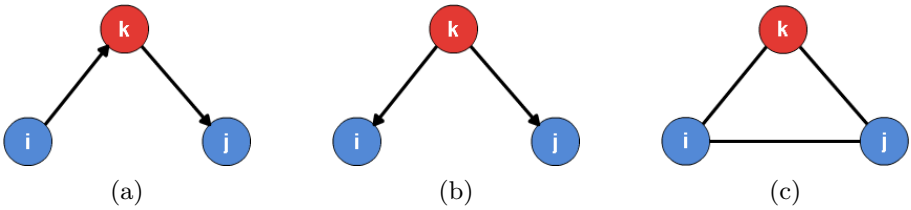


Fig. 1. Possible transmission relationships among three nodes [9]. Blue nodes are the targets of correlation analysis. The red one is the intermediate node. (a) shows that there is no directed edge between nodes i and j. Disease transmission follows a path from node i to k, then to j. (b) shows node k transmits to nodes i and j simultaneously and independently. (c) shows the pearson correlation results for (a) and (b).

as illustrated in Fig. 1(c). The same problem also exists in the case of Fig. 1(b), where i and j are the children of k considering the disease transmission dynamics.

To avoid such situations, we carry out first-order partial correlation analysis, which measures the dependence between two variables, while removing or fixing a third variable. In this regard, to compute it between nodes i and j, we remove or fix the impact of another node k, where $k = 0, 1, 2, ..., N$ and $k \neq i, j$. From the results, we choose coefficients that indicate strong correlations with significant p-values. It should be mentioned that partial correlation usually does not provide edge directions [10]. Therefore, to infer directed edges, we analyze the time-series data with a time lag (e.g., one day or one week). Then, the direction is from the node using the previous time-series data to the node using the current one. The partial correlation coefficient between nodes i and j after fixing the variable of node k is $\rho_{ij.k} = (\rho_{ij} - \rho_{ik}\rho_{jk})(\sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2})$, where ρ_{ij}, ρ_{ik}, and ρ_{jk} are the covariances. This method removes many false positive edges as well as generates an approximate partial correlation network (PCN), G^p.

3.2 Back-Tracking Bayesian Learning

It should be noted that some edges in G^p still do not exist in G. A possible solution is to set the weights of these false positive edges with values of zero during the inference process. This is similar to the removal of irrelevant basis components as in basis pursuit for dimensionality reduction [11]. In our proposed inference method, we base our second step on the Sparse Bayesian Learning (SBL) framework [12]. 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 node i, we divide preprocessed surveillance dataset D into two subsets: an $M \times 1$ vector of $\boldsymbol{y} = \{<v_i, ic_i, t_i> \mid t_i = 2, 3, ..., M+1, M \in T\}$ and an $M \times |N^p|$ matrix of $\boldsymbol{x} = \{<v_j, ic_j, t_j> \mid j \in N^p, t_j = 1, 2, ..., M, M \in T-1\}$. 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 in Section 2 as follows:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{w}^T + \varepsilon \tag{4}$$

where $\boldsymbol{w} = \{w_j \mid j \in N^p\}$ is a vector indicating all possible incoming links estimated based on G^p. ε is an error term. Under the framework of SBL, both $\boldsymbol{w}$ and ε follow a zero-mean Gaussian distribution with variances of $\boldsymbol{\alpha}$ and β, respectively. They are defined respectively as: $p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{N^p} N(w_i|0, \alpha_j^{-1})$ and $p(\varepsilon) = N(0, \beta)$. Because we have no prior knowledge about $\boldsymbol{w}$ and ε, it is reasonable to set them with non-informative prior distributions (e.g., Gamma distribution). $\boldsymbol{\alpha}$ and β are assumed to have the same hyperparameters for all nodes.

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

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \beta) = \frac{likelihood \times prior}{normalize\ factor} = \frac{p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)} \tag{5}$$

which is a Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \beta^{-1}\boldsymbol{\Sigma}\boldsymbol{x}^T\boldsymbol{y}$, $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \beta^{-1}\boldsymbol{x}^T\boldsymbol{x})^{-1}$ where $\boldsymbol{\Lambda} = diag(\alpha_1, \alpha_2, ..., \alpha_{N^p})$. "type-II maximization likelihood" maximization combined with a maximum a posteriori probability (MAP) estimate transforms the whole problem into that of maximizing the marginal likelihood function of:

$$p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta) = \int p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} \tag{6}$$

Writing Eq. 6 into a logarithm form $\mathcal{L}(\boldsymbol{\alpha})$, we have:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}) &= \log p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta) = \log \int p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} \\ &= -\frac{1}{2}[M \log 2\pi + \log |\boldsymbol{C}| + \boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}] \end{aligned} \tag{7}$$

with $\boldsymbol{C} = \beta I + \boldsymbol{x}\Lambda^{-1}\boldsymbol{x}^T$. The derivatives with respect to α_j and β are [13]:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \log \alpha_j} = \frac{1}{2}(1 - \alpha_j \Sigma_{jj} - \alpha_j \mu_j^2) \tag{8}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \log \beta} = \frac{1}{2}[\frac{M}{\beta} - \|\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\mu}\|^2 - trace(\boldsymbol{\Sigma}\boldsymbol{x}^T\boldsymbol{x})] \tag{9}$$

Setting Eqs. 8 and 9 to zero, the estimations of α_j and β become:

$$\alpha_j^{new} = \frac{1 - \alpha_j \Sigma_{jj}}{\mu_j^2} \tag{10}$$

$$\beta^{new} = \frac{M - \sum_{j=1}(1 - \alpha_j \Sigma_{jj})}{\|\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\mu}\|^2} \tag{11}$$

The above iterative estimation procedure is solved by using the Expectation-Maximization algorithm. In each iteration, we estimate the contributions to the marginal likelihood function for all nodes in G^p. The one with maximum contribution is selected as the candidate neighbor. Then, its corresponding weight is computed. As to be noted, in G, we only have positive links. However, the prior distribution may cause $\boldsymbol{w}$ to become negative. To avoid this, a constraint of limiting $\boldsymbol{w}$ to be positive is introduced. To incorporate this constraint, we use a back-tracking technique. During the EM learning procedure, the update of marginal likelihood function and other parameters proceeds sequentially. Consequently, each time $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, α_j, and β are updated, we select those α_j that fail the constraint, and put their corresponding indices into a blacklist. The program is rolled back to the previous step and proceeds by selecting nodes that do not exist in the blacklist. The algorithm is shown in Alg. 1.

Algorithm 1. Back-Tracking Bayesian Learning

Require: D: Preprocessed surveillance dataset; G^p: Partial correlation network;
Ensure: G^*: Inferred disease transmission network;

1. Divide D into two subsets with time lag of one time unit;
2. **for all** node $i = 1, 2, ..., N$ **do**
3. Initialize parameters for prior distributions;
4. Construct marginal likelihood function $p_i(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)$ (shown in Eq. 6);
5. **while** not reaching stopping criteria **do**
6. **for all** node $j \in N^p$, and $i \neq j$ **do**
7. Compute contributions to $p_i(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)$;
8. **end for**
9. Select node with maximum contribution;
10. Re-estimate all weights of current neighbors of node i;
11. **if** all weights are not less than zero **then**
12. Update neighborhood list;
13. **else**
14. Remove neighbors with weights less than zero, and put them into blacklist;
15. Roll back $p_i(\boldsymbol{y}|\boldsymbol{\alpha}, \beta)$;
16. **end if**
17. **end while**
18. **end for**
19. Combine all neighborhood lists to construct G^*;
20. return G^*;

3.3 Discussions

As stated in [3], it is not trivial nor practical to find all the edges within G, or the exact time required to stop the inference program. Thus, once the program iterates to the maximum permitted iteration steps, or the update of the marginal likelihood function converges to a small value, we will stop the learning procedure. To compute the PCN, the time complexity is $O(N^3)$. To speed up this

process, we use dynamic programming to recursively compute the first-order partial correlation based on the result of zeroth-order partial correlation. As for the back-tracking Bayesian learning, the complexity of Bayesian learning is mainly distributed over the computation of parameters of $\boldsymbol{\Sigma}$, which requires $O(N^3)$. An efficient incremental algorithm proposed in [12] can optimize this computation. Besides, the computation based on G^p can also reduce this computational time. After integrating the back-tracking algorithm, the time complexity becomes exponential. However, based on our experiments, the algorithm usually converges fast. That is to say, the algorithm seldom tracks back to the nodes that are selected at the very beginning. This is caused by the previous Bayesian learning; it selects those significantly contributing nodes at the very beginning, making the marginal likelihood function converge to a near optimum solution without large space to increase, and stable until reaching the stopping criteria.

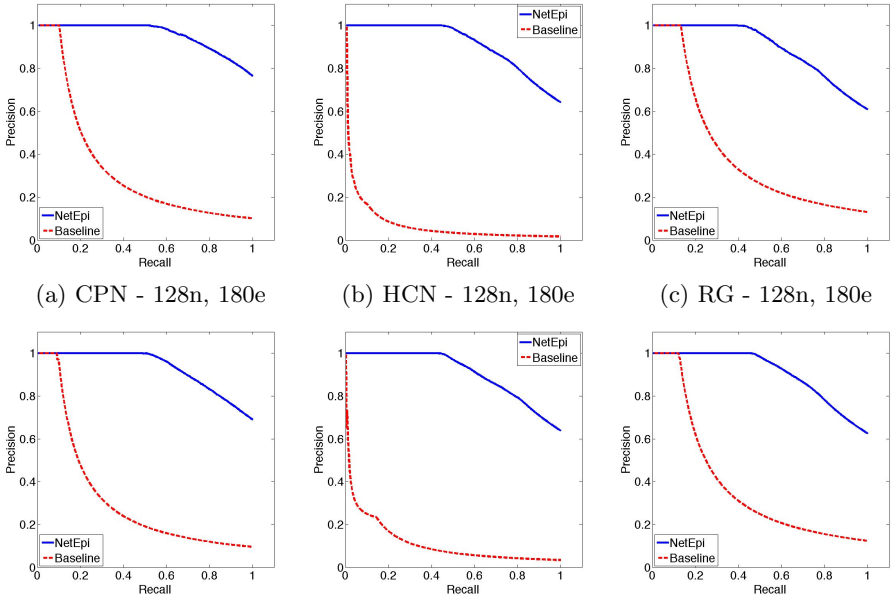
4 Experiments

4.1 Experiments Based on Synthetic Data

The synthetic data generation proceeds as follows: we first use Kronecker Graphs model [14] to generate a basic network structure. Then, we link all the nodes with an external node v_0, and generate self-connected edges with predefined probabilities. We iteratively run the transmission model, as given in Eq. 1, for a sufficient number of time steps to generate the disease surveillance data.

Experimental Setting. We construct 3 types of network structures: (i) core-periphery networks (CPNs), which have a cluster of nodes in the core of the network, (ii) hierarchical community networks (HCNs), where a proportion of the nodes form several small communities, and (iii) random graphs (RGs), which have no obvious pattern. Then, for each structure, we generate networks with different sizes: 64n with 100e, 150e ("n"and "e" are the abbreviations of "nodes" and "edges", respectively); 128n with 180e, 200e; 256n with 350e, 400e; 512n with 720e, 800e. For each of them 10 datasets are produced. Specifically within each generation process, we make sure that the transmission process cover all the edges in G. In total, there are 3 types of network topologies $\times$ 8 different sizes $\times$ 10 independent transmission processes = 240 datasets.

The Baseline Method. To our best knowledge, there have not been much prior work on inferring network structures over DCGs. 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 connections to those nodes have infection cases at $n+1$. The edge weight is affected by the number of infection cases and the number of infected nodes at the previous time step. We select the top k edges with the highest weights, and form the estimated disease transmission network G^* accordingly. The mathematical formula to compute the baseline edge weight is $w_{ij} = {ic_i}^t {ic_j}^{t+1} / \sum_{i=1}^{N} {ic_i}^t$.



(a) CPN - 128n, 180e (b) HCN - 128n, 180e (c) RG - 128n, 180e

(d) CPN - 256n, 350e (e) HCN - 256n, 350e (f) RG - 256n, 350e

Fig. 2. The precision-recall curves for synthetic DTNs. It is obvious that NetEpi outperforms the baseline method in all cases. The average degree for the above six networks are 1.3876, 1.4496, 1.4651, 1.3619, 1.5175, and 1.5097, respectively, from (a) to (f).

Result Evaluation. To evaluate the inference results, we compute the precision-recall curves as shown in Fig. 2. For the sake of space, we display only part of our experimental results here. The precision and recall are defined as "what fraction of edges in G^* is also present in G", and "what fraction of edges of G appears in G^*", respectively [3]. For nodes i and j, if both ground-truth edge e_{ij} and inferred edge e^*_{ij} exist, and the difference between the corresponding weights $|w_{ij} - w^*_{ij}|$ is less than a threshold, then we say the inferred edge is accurate.

In our experiments, NetEpi outperforms the baseline method in all 240 datasets. Specifically, for networks that have the same sizes but different topologies, NetEpi performs the best on the CPNs. Nodes located in the core region have more connections as compared with those in the periphery region. Therefore, to achieve an optimal solution, core-located nodes will have higher probabilities to possess a more number of combinations of neighbors. In other words, the probabilities to find a globally optimal solution for a single node will decrease as the number of its incoming edges increases. The accuracies of NetEpi over CPNs are consequently biased by the tradeoff between core-located and periphery-located nodes. In comparison, for HCNs, there is no longer a single core. In contrast, there are several sub-cores that individually form a sub-community. This structure makes the average number of combinations for each

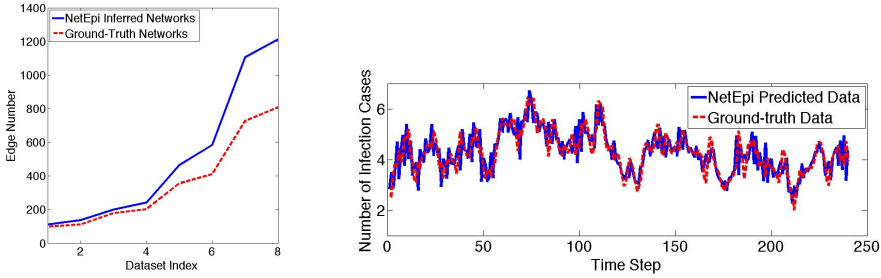


Fig. 3. (a) As the GTN size increases, the accuracy of NetEpi decreases. The number of false edges increases as well. (b) NetEpi accurately captures the disease transmission trend.

node increase and directly affect the inference accuracy. As for RGs, the number of connections for each node does not have a fixed pattern, and NetEpi achieves oscillating results. Here, we use an average out-degree distribution to illustrate the accuracy differences between networks with distinct topologies. It is defined as $d_{avg} = \sum_{i=1}^{N} d_i/N$ where d_i is the out degree for node i, d_{avg} is the average degree for the whole network. Our analysis results of the 24 synthetic networks show that the average degrees for CPNs are always smaller than those for HCNs. And, the average degrees for RGs present oscillating patterns (Fig. 2).

For networks with the same topologies but different sizes, NetEpi achieves better results on inferring smaller ones as shown in Fig. 2. During the inference process, the whole Ground-Truth Network (GTN) is treated as a complete network. Even given the approximate structure G^p, the complexity quadratically increases as the number of nodes increases. Meanwhile, as the edge number increases, the number of combinations of neighbors for each node to achieve optimal solutions increases as well, which directly interferes the inference results as in Fig. 3(a). However, the network sizes of metapopulation based DTNs are usually small at the administrative level. For example, for a global epidemic disease, WHO publishes statistical reports at the country level. Therefore, a possible solution to infer large-size networks is to perform hierarchical clustering based on geographical information. NetEpi is conducted from the highest level where each node represents a cluster of lower-level nodes. Then, within each high-level node, NetEpi is performed again to infer lower-level transmission networks. This process is repeatedly and sequentially conduced in order to get a whole picture of large-size networks. Fig. 3(b) shows an example of the prediction results of NetEpi. It is obvious that the predicted epidemic trend happening in the GTN is well captured by the inferred network. This validates that NetEpi converges to an optimal solution, although this may not be the global one.

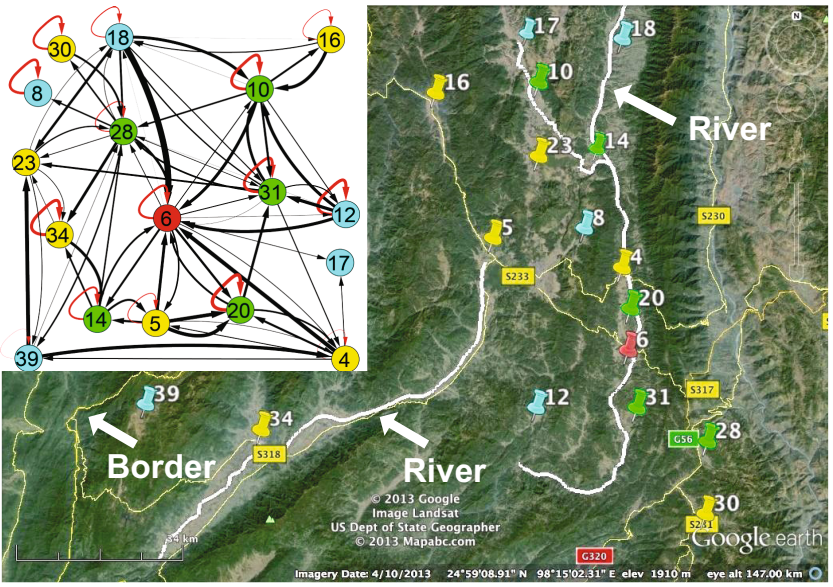
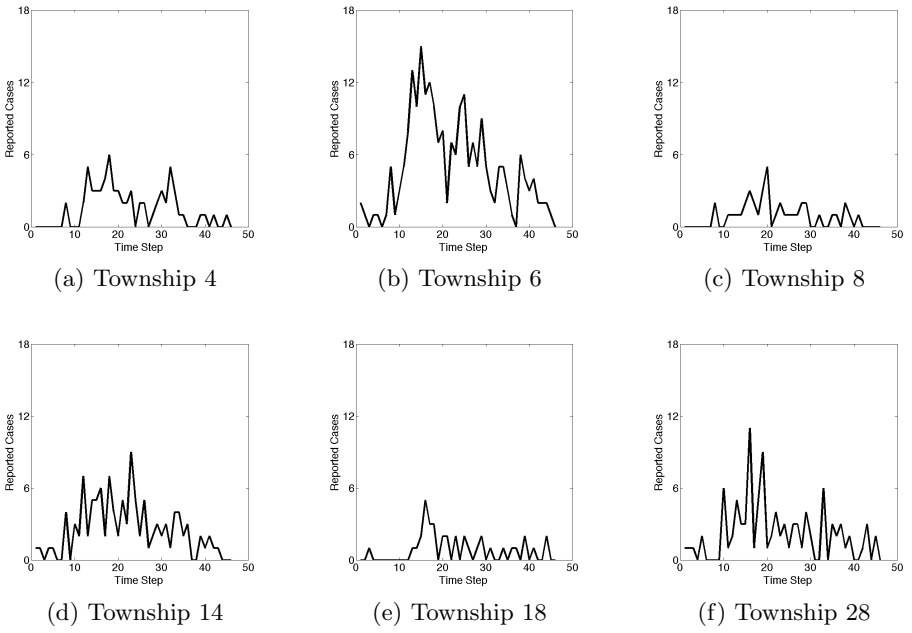


Fig. 4. Townships that form a community in the inferred malaria transmission network

4.2 Experiments Based on a Real-World Dataset

Experimental Setting. The real-world dataset was provided by Chinese Center for Disease Control and Prevention. It contains reported malaria cases in Yunnan province, China. In total, there are 2928 cases reported in 51 townships in 2005. These townships are distributed along the border between China and Myanmar (a high malaria-endemic country) and classified into 5 categories based on the numbers of infection cases: $(200, +\infty)$ (red), $(150, 200]$ (purple), $(100, 150]$ (green), $(50, 100]$ (yellow), and $(0, 50]$ (blue). The dataset is very sparse, with missing data. Moreover, there is no complete information about the sources and identifications of imported cases. Thus, a fixed external node cannot be set up before 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. We are informed that there exist imported cases, and expect that 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 [15]. However, studies have also reported that the incubation time could be longer from several months to several years [15]. Therefore, we choose 21 days as the time window when inferring



(a) Township 4 (b) Township 6 (c) Township 8

(d) Township 14 (e) Township 18 (f) Township 28

Fig. 5. 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.

the underlying malaria transmission network, so that it compromises both the reported incubation time and the sensitivity analysis that we have conducted previously.

Result Interpretation. The inferred network contains two classes of nodes. Some of them connect to themselves as we expected, while the others form communities. Self-connected nodes occupy 50.98% of the whole network. This caters to our previous expectation. These nodes are located adjacent to the border between China and Myanmar, or connected with the border by highways, or situated close to rivers, which provide suitable environments for the vectors of malaria to reproduce. Therefore, the malaria endemic within these self-connected nodes are possibly caused by EIC, ITC, or their combinations.

As shown in Fig. 4, there is a community found in the inferred network. It contains most nodes that have severe endemic situations. Many townships are distributed along two rivers. Besides providing suitable habitat there, rivers also bring the larva of vectors from the upstream to the downstream. Therefore, the inferred edges of these nodes possibly represent partial influences from rivers, and impact of vectors' movements. In addition, 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 endemic in the region. Other townships 16, 28, and 30 are connected

with the others by highways (e.g., S231, S233, S317, and S318), indicating that their transmission pathways are possibly caused by transportation.

It can readily be noted from Fig. 4 that some inferred edges are thicker than others, denoting higher transmission influences. For example, e_{18-6} (the dash in the index is used for separation) is thicker than e_{14-6}, e_{4-6}, and e_{28-6}. We interpret this based on Fig. 5 (a) - (f) where reported cases are aggregated on an eight-day basis for clear presentation. As shown, although township 18 (Fig. 5(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 (Fig. 5(b)). In comparison, the "mountain-valley-mountain" pattern of township 6 can only be partially matched with other townships (e.g., townships 4 (Fig. 5(a)), 14 (Fig. 5(d)) and 28 (Fig. 5(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 steps 20 to 30 in the trend of township 6 cannot contribute to the valley appearing 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 (Fig. 5(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.

There are totally 47 rather than 51 townships contained in the inferred network. The 4 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 infection cases. Therefore, we think their disease transmission dynamics are caused by accidentally imported cases. In addition, although some townships are located very close to each other, and in the positions of the upstream or the downstream of the same river, they are not connected in the inferred network (e.g., townships 10 and 17). We believe this is because their transmission pathways are not significant or their malaria endemic is mainly affected by imported cases that overtook the impact of other factors. To interpret them, currently available information about highways, rivers, and geographical locations may not be fully adequate, because they represent the transmission pathways that are the *comprehensive results of all impact factors*. In addition, the roads that are locally formed and managed are not displayed in the map, which may also play significant roles in malaria transmission. Missing reports and data sparsity may affect the results as well. However, our method can still detect some hidden connections that may draw the attention of policy makers.

5 Conclusion

In this study, based on the need for real-world disease transmission pattern discovery, we have defined and addressed an inverse network inference problem. Given only the surveillance data, we have proposed a two-step network inference algorithm, called NetEpi. Having highlighted the major differences between the individual based network inference and the metapopulation based

network inference problems, we defined a linear disease transmission model over a Directed Cyclic Graph (DCG) containing three types of transmission pathways as often found in the real-world situations, namely, internal transmission component, neighborhood transmission component, and external influence component. We performed partial correlation analysis to construct an approximate network structure for the underlying disease transmission network, and then conducted back-tracking Bayesian learning to iteratively infer edges and estimate their corresponding weights. We have evaluated the proposed method by using synthetic data. The experimental results have shown that NetEpi outperforms a probability based baseline inference method, and performs well over a relatively small-scale network, which is sufficient for metapopulation based disease transmission network modeling in practice. Meanwhile, NetEpi achieves a reasonable accuracy over different network topologies. In addition, we have applied NetEpi to a real-world disease transmission dataset and have discovered certain meaningful community patterns as well as transmission pathways. Our future work will focus on inferring disease transmission networks in which there exist various underlying, sometimes dynamically-changing network structures. We will also consider other impact factors that may be disease-dependent. This work will further be applied to the real-world situations for policy makers to develop and implement intervention strategies for controlling disease transmission.

Acknowledgement. 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.

References

1. Keeling, M.J., Eames, K.T.: Networks and Epidemic Models. J. R. Soc. Interface. 2, 295–307 (2005)
2. Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V., Vespignani, A.: Human Mobility Networks, Travel Restrictions, and Global Spread of 2009 H1N1 Pandemic. PLoS ONE 6, e16591 (2011)
3. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring Networks of Diffusion and Influence. In: 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1019–1028. ACM Press, New York (2010)
4. Myers, S., Leskovec, J.: On the Convexity of Latent Social Network Inference. In: Advances in Neural Information Processing Systems, pp. 1741–1749 (2010)
5. Teunis, P., Heijne, J.C.M., Sukhrie, F., van Eijkeren, J., Koopmans, M., Kretzschmar, M.: Infectious Disease Transmission as a Forensic Problem: Who Infected Whom? J. R. Soc. Interface 10(81) (2013)
6. Arino, J.: Diseases in Metapopulations. In: Ma, Z., Zhou, Y., Wu, J. (eds.) Modeling and Dynamics of Infectious Diseases. Series in Contemporary Applied Mathematics, vol. 11, pp. 65–123. World Scientific (2009)
7. Ndeffo, M.M.L., Gilligan, C.A.: Resource Allocation for Epidemic Control in Metapopulations. PLoS ONE 6, e24577 (2011)

8. 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 Meterorological Conditions in the Onset of Dengue Epidemics. PLoS Negl. Trop. Dis. 4, e775 (2010)
9. Yuan, Y., Li, C.T., Windraw, O.: Directed Partial Correlation: Inferring Large-Scale Gene Regulatory Network through Induced Topology Disruptions. PLoS ONE 6, e16835 (2011)
10. Lasserre, J., Chung, H.R., Vingron, M.: Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks. PLoS Comput. Biol. 9, e1003168 (2013)
11. David, P.W., Bhaskar, D.R.: Sparse Bayesian learning for basis selection. IEEE Trans. Signal Processing. 52(8), 2153–2164 (2004)
12. Tipping, M.E., Faul, A.: Fast Marginal Likelihood Maximization for Sparse Bayesian Models. In: 9th International Workshop on Artificial Intelligence and Statistics, pp. 3–6 (2003)
13. Tzikas, D., Likas, C., Galatsanos, N.: Sparse Bayesian Modeling with Adaptive Kernel Learning. IEEE Trans. Neural Networks. 20(6), 926–937 (2009)
14. Leskovec, J., Faloutsos, C.: Scalable Modeling of Real Graphs using Kronecker Multiplication. In: 24th International Conference on Machine Learning, pp. 497–504. ACM Press, New York (2007)
15. 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. Malar. J. 10(1), 122 (2011)