Inferring disease transmission networks

Xiaofei Yang
Hong Kong Baptist University

Follow this and additional works at: https://repository.hkbu.edu.hk/etd_oa

Recommended Citation
Yang, Xiaofei, "Inferring disease transmission networks" (2014). Open Access Theses and Dissertations. 88.
https://repository.hkbu.edu.hk/etd_oa/88

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at HKBU Institutional Repository. It has been accepted for inclusion in Open Access Theses and Dissertations by an authorized administrator of HKBU Institutional Repository. For more information, please contact repository@hkbu.edu.hk.
Inferring Disease Transmission Networks

YANG Xiaofei

A thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Philosophy

Principal Supervisor: Prof. Jiming LIU

Hong Kong Baptist University

March 2014
Declaration

I hereby declare that this thesis has been composed by myself under the guidance of my principal supervisor, Prof. Jiming LIU, after registration for the degree of MPhil at Hong Kong Baptist University. The thesis has not been previously included in any thesis, dissertation or report submitted to any institution for a degree, diploma or other qualification. All the reported results have been generated and verified using the methods as described. All sources of information have been acknowledged by means of references to the relevant publications.

Signature: ____________________________

Date: March 2014
Abstract

To investigate how an infectious disease spreads, 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 within the diffusion network. 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, that is, to consider disease transmission networks in which nodes represent subpopulations, and links indicate their interrelationships. 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. Therefore, this thesis addresses the problem of inferring disease transmission networks at a metapopulation level. 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 recover most of the ground-truth disease transmission networks using only surveillance data.
Acknowledgements

I would like to thank all the professors and colleagues who have helped me during my 6-year’s study in Hong Kong Baptist University.

In particular, I would like to thank my supervisor Prof. Jiming LIU. My research interest was initiated by his excellent teaching in my undergraduate courses. His inspiring guidance and constructive suggestions in my master’s studies had laid a solid foundation for my future work.

I would also like to thank Prof. ZHOU Xiao Nong, Dr. William Kwok Wai CHEUNG, Prof. YUEN, Pong Chi for their insightful advice and encouragement. Thanks also go to the departmental secretaries and technicians for their kind assistance.

I give my special thanks to my parents and my girlfriend He Jing for their persistent support throughout my studies.
# Table of Contents

Declaration i

Abstract ii

Acknowledgements iii

Table of Contents iv

List of Tables vii

List of Figures viii

1 Introduction 1

1.1 Overview and Objectives ........................................... 1

1.2 Contributions and Significance ..................................... 5

1.3 Outline of the Thesis ............................................... 7

2 Related Work 9

2.1 Information Diffusion Network Inference ......................... 11

2.1.1 Information Diffusion Models ................................11
2.1.2 Link Prediction Problem .............................................. 12
2.1.3 Diffusion Network Inference ................................. 13
2.2 Disease Transmission Networks ................................. 16
   2.2.1 Disease Transmission Models ............................ 16
   2.2.2 Disease Transmission over Networks ....................... 22
   2.2.3 Inference Based on Surveillance Data ...................... 28
2.3 Summary ................................................................. 34

3 Problem Statement ..................................................... 37
   3.1 Definitions ............................................................. 38
   3.2 Linear Transmission Model ........................................ 42
   3.3 Network Inference Problem ....................................... 44
   3.4 Summary ................................................................. 45

4 The Proposed Network Inference Method ......................... 46
   4.1 Basic Ideas ............................................................. 47
   4.2 Partial Correlation Network Construction .................... 47
      4.2.1 Pearson Correlation Analysis ............................. 50
      4.2.2 First-Order Partial Correlation Analysis ................. 51
   4.3 Back-Tracking Bayesian Learning ............................... 53
      4.3.1 Marginal Likelihood Function ............................. 53
      4.3.2 Expectation-Maximization Computation ................... 55
      4.3.3 Back-Tracking Technique .................................. 56
   4.4 Discussions .......................................................... 56


4.4.1 Stopping Criteria ........................................... 56
4.4.2 Computational Complexity of the Algorithm .............. 58
4.5 Summary ......................................................... 59

5 Experiments ......................................................... 60
5.1 Experiments Based on Synthetic Data ......................... 61
  5.1.1 Experimental Setting ...................................... 61
  5.1.2 Baseline Method ........................................... 65
  5.1.3 Results ..................................................... 66
  5.1.4 Sensitivity Analysis ....................................... 76
5.2 Experiments Based on a Real-world Dataset .................. 93
  5.2.1 Dataset Description ....................................... 93
  5.2.2 Experimental Setting ..................................... 95
  5.2.3 Results .................................................... 97
5.3 Summary ......................................................... 112

6 Conclusion and Future Work ..................................... 114
  6.1 Main Contributions ........................................... 114
  6.2 Future Work .................................................. 115

Curriculum Vitae ..................................................... 129
List of Tables

1.1 An example of surveillance data ............................................. 4

2.1 Differences and considerations of different network-based studies ............. 35

3.1 Surveillance data generated based on the synthetic disease transmission net-
work shown in Fig. 3.1 (sorted based on infection time). .................... 39

3.2 Preprocessed surveillance data generated based on Table. 3.1 (sorted based
on infection time) .................................................................................. 40

5.1 Out-degrees for synthetic networks .................................................. 75
List of Figures

1.1 An illustration of individual-based and metapopulation-based disease transmission networks [1]. Blue cycles on the left-hand side represent spatial sub-populations in a metapopulation-based disease transmission networks. Green triangles, red diamonds and yellow squares represent susceptible, infected, and recovered individuals within each subpopulation, respectively. .................. 3

2.1 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. ................................................................. 15
2.2 A graphic representation of the Ross model. The transmission dynamics within the human population is described by an SIS model, indicating that infected people can recover and become susceptible again. Compared to humans, the life span of the vector is much shorter; therefore an SI model is used to describe the transmission dynamics within the vector population. The solid lines indicate the transmission within the same species. The dash lines represent the transmission interaction between humans and vectors.

2.3 A snapshot of the factors that affect malaria transmission. Based on three time scales (e.g., short, middle and long), we classify them into five categories: malaria parasites, epidemic vector, interventions, socioeconomic factors and climate change. Directed connections between any two impact factors indicate their direct relationships.

2.4 An illustrative model that is used to simulate malaria transmission in a population level. Similar to Ross model, it is also divided into two parts. Based on the reflection of people who get infected, the model that describes human transmission dynamics extends the infected state into several compartments in very detail compared with Ross model (Fig. 2.2). As for the vector, a compartment for the latently infected is introduced to simulate the impact factor of parasite incubation time.
2.5 We use the infectious disease of H7N9 as an example to display the interactive interface of HealthMap. As shown in the figure, all the infection cases are classified based on their states (i.e., confirmed or suspected, death or not death) with distinct colors. At the bottom of (a), all the related alert news are listed. Once a user clicked the news, as shown in (b), a pop-up panel will appear and display the corresponding details.

2.6 An example of utilizing SatScan to analyze West Nile virus incidence cases in US [67]. Both spatial and temporal dynamic changes of infection cases are captured by SatScan. Different levels of severities are displayed.

3.1 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).

4.1 A flowchart of NetEpi. The first step is to obtain approximate network structures based on the first-order partial correlation analysis. The second step is to carry out the Back-Tracking Bayesian Learning.
4.2 The possible transmission relationships among three nodes [78]. 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. (c) shows the possible correlation results among nodes $i$, $j$, and $k$.

5.1 An example of core-periphery networks with 128 nodes, and about 500 directed edges. For the sake of presentation, the directions of edges are not shown in the figure. As shown in the figure, a cluster of yellow nodes can be observed in the core of the network. The red links are the connections started or ended at those core nodes. Other black nodes with fewer connections are distributed in the periphery area.

5.2 An example of hierarchical community networks with 128 nodes, and about 500 directed edges. For the sake of presentation, the directions of edges are not shown in the figure. There are several small communities formed within the network. These communities are connected, forming one big cluster.

5.3 An example of random graphs with 128 nodes, and about 500 directed edges. For the sake of presentation, the directions of edges are not shown in the figure. There are not obvious patterns in this network.
5.4 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.

5.5 The precision-recall curves for synthetic core-periphery networks. It is obvious that NetEpi outperforms baseline the method in all cases.

5.6 The precision-recall curves for synthetic hierarchical community networks. It is obvious that NetEpi outperforms the baseline method in all cases.

5.7 The precision-recall curves for synthetic random graphs. It is obvious that NetEpi outperforms the baseline method in all cases.

5.8 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. 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.
5.7 Sensitivity analysis for core-periphery networks with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3. 

5.8 Sensitivity analysis for hierarchical community networks with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.

5.9 Sensitivity analysis for random graphs with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.

5.10 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. The curve with the size of a quarter of the number of network nodes is displayed as a blue solid line. The curve 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.
5.11 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. 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.

5.12 Snapshot of 51 townships located in Yunnan province, China, along the border between China and Myanmar. The placemarks represent townships, which are colored based on their numbers of infection cases: (i) red nodes: \((200, +\infty)\); (ii) purple nodes: \((150, 200]\); (iii) green nodes: \((100, 150]\); (iv) yellow nodes: \((50, 100]\); (v) blue nodes: \((0, 50]\). This picture is created using Google Earth.

5.13 The inferred malaria transmission network-based on the surveillance data in Yunnan province, China. There are in total 2928 infection cases reported in 51 townships during 2005. The width of the edge indicates the weight of transmission pathways. The color of each node represents its disease epidemic severity. Red edges represent a malaria endemic caused by the internal transmission component, the external influence component or their combination, while black edges represent a malaria endemic caused by the neighborhood transmission component.
5.14 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 head to the border. . . . . . 98

5.15 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 Fig. 5.14. . . . . . . . . . . . . . . . 99

5.15 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 Fig. 5.14. . . . . . . . . . . . . . . . 100

5.16 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. To interpret this result, more detailed information is needed. . 102

5.17 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) - (c) show the curves for townships selected from Fig. 5.16. . . . . . . . . . . . . . . . 103

5.18 Townships 2, 3, 24, 25 and 26 are located adjacent to the border between China and Myanmar. Therefore, their self-connected edges are more likely to represent the localized malaria endemic caused by the imported cases. . . . . 104

5.19 Although township 13 is located adjacent to the border between China and Myanmar, it is in a valley and almost isolated from the traffic to the border. Moreover, there is a river crossing this township. Therefore, its self-connected edge is likely due to local transmissions. . . . . . . . . . . . . . . 105
5.20 Townships 42, 45, 46, 47, 48 and 51 are located adjacent to the border between China and Myanmar. Therefore, their self-connected edges are due to the imported cases. 

5.21 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) - (c) show the curves for townships selected from Fig. 5.20.

5.22 Townships 35, 37, 38 and 40 are located in the same valley, and along the same river. The surrounding environment provides a suitable condition for the malaria vector to develop and reproduce. Besides, there are two highways (S233 and S318) cross these townships that also connect to the border. Therefore, the malaria endemics within these townships are likely to be caused by a combination of the internal transmission dynamics and the imported cases.

5.23 Townships 36 and 43 are located close to a river, which provides them with a suitable condition that helps the vector to develop and reproduce.

5.24 Townships 9, 22 and 27 may be influenced by the internal transmission dynamics because of their locations adjacent to rivers. The rest of the townships shown in this figure may be influenced by the imported cases, because of the convenient traffics connect them to the border.
Chapter 1

Introduction

This chapter first demonstrates the necessity of studying disease transmission and control. It applies related theoretical studies to real-world practice, to illustrate the gaps in previous research. Specifically, the research is conducted according to what is required for real-world applications, which is to infer disease transmission networks at a metapopulation level and provide policy makers with intuitive insights and practical intervention suggestions. Accordingly, the thesis makes three major contributions: (i) builds a generalized linear disease transmission model that considers all the possible transmission pathways at a metapopulation level, (ii) it inversely infers hidden underlying disease transmission networks using spatiotemporal surveillance data, and (iii) it solves this inverse problem over Directed Cyclic Graphs. The chapter ends with an outline of the structures of this thesis.

1.1 Overview and Objectives

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. Therefore,
various intervention strategies have been applied, including but not limited to the following.

- **Vaccination**: Based on the understanding of the biological mechanisms of disease
  transmission, vaccines are developed and distributed to limit severe disease outbreaks.

- **Contact Deduction**: Monitoring disease transmission at both spatial and temporal scales,
  which is also called surveillance, provides valuable information for predicting, locating,
  isolating and protecting susceptible populations.

- **Prioritization and Optimization of Intervention Strategies**: The analysis of disease
  severity within a population can be extended so that intervention strategies, objective
  areas, and susceptible populations can be prioritized.

Another strategy, contact tracing, is also widely used to prevent disease outbreaks [2]. This
is a network-based approach conducted at an individual level. Potentially susceptible individ-
uals are identified and monitored to minimize the chances of infection. The network-based
approach not only helps to differentiate mixed populations [3], it can also be transformed
into other simulation models [4]. This approach is very close to the one adopted in this
thesis, 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 [5]); instead they rep-
resent patches of subpopulations (e.g., provinces, cities, and townships) and various trans-
mission pathways among them (e.g., highways and air travel routes), as shown in Fig. 1.1.
Figure 1.1: An illustration of individual-based and metapopulation-based disease transmission networks [1]. Blue cycles on the left-hand side represent spatial subpopulations in a metapopulation-based disease transmission networks. Green triangles, red diamonds and yellow squares represent susceptible, infected, and recovered individuals within each subpopulation, respectively.
Table 1.1: Example of surveillance data

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Infection Time</th>
<th>Infection Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
<td>23</td>
<td>M</td>
<td>31-10-2012</td>
<td>New York</td>
</tr>
<tr>
<td>Mary</td>
<td>20</td>
<td>F</td>
<td>02-12-2012</td>
<td>LA</td>
</tr>
<tr>
<td>Tom</td>
<td>25</td>
<td>M</td>
<td>14-11-2012</td>
<td>Boston</td>
</tr>
<tr>
<td>Ray</td>
<td>22</td>
<td>M</td>
<td>21-11-2012</td>
<td>Miami</td>
</tr>
</tbody>
</table>

Both individual-based and metapopulation-based studies of disease transmission networks are useful for the following reasons:

- It is helpful to analyze epidemic phase transition behavior [6] (where “epidemics are defined as outbreaks that affect a non-zero fraction of the population in the limit of large system size.” [3]);

- The dominant factors that underlie the spread of a disease epidemic can be investigated [7];

- Suggestions are provided to practically control epidemics by cutting off transmission pathways and/or isolating certain local regions [8].

Existing disease transmission studies that deal with the above two types of transmission networks all share the same limitation, that is, they assume that network structures are given in advance; for example, contact structures for influenza spreading [9] [10] or airlines for the spread of H1N1 [11] and SARS [12]. In these studies, information about which person or location will be infected next is given. However, in an actual epidemic, only the spatiotemporal surveillance datasets containing the infection times and locations of reported infection
cases are obtained [13]. 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 [14]).

1.2 Contributions and Significance

The network inference problem has been recently and widely studied in the research domain of information diffusion. 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 [15] [1].
- It is difficult to simulate complicated individual human behavior and collect large amounts of personal information [7] [16] [17].
- Controlling disease transmission at a metapopulation level is more practical from the viewpoint of front-line practitioners and policy makers [18].

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 infected people within the same subpopulation.

2. Unlike information diffusion or individual-based disease transmission networks, disease transmission at the 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 that 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 [1]. Partial results of this research have been reported in [19].

1.3 Outline of the Thesis

The thesis is organized as follows.

• Chapter 2 introduces related work including the problem of inferring information diffusion networks, disease transmission models, and recently conducted work on inferring disease transmission networks.

• Chapter 3 defines the three components abstracted from all the possible disease transmission pathways. Then a generalized metapopulation-based linear disease transmission model is proposed. The problem of inferring disease transmission networks is also formulated.

• Chapter 4 proposes a network inference method called NetEpi, which uses the first-order partial correlation analysis to obtain approximate transmission network structures. Then, a Bayesian learning framework combined with a back-tracking technique is used to fine-tune the results based on the partial correlation networks developed in previous step.

• Chapter 5 presents the experimental results, including an evaluation of NetEpi with both synthetic and real-world datasets. A probability-based baseline method is proposed to make comparisons. To analyze the robustness of NetEpi, a parameter sensi-
tivity analysis is conducted.

- Chapter 6 concludes this thesis and discusses future research.
Chapter 2

Related Work

As mentioned in Chapter 1, this study focuses on network-based disease transmission, because according to [3], most existing studies of disease transmission assume that populations are mixed completely, and therefore each individual has the same probability of infecting all other individuals in the population. This is definitely not a reasonable assumption, as the number of neighbors or friends each person knows is limited. This assumption of complete and homogeneous mixing also makes population-based disease transmission models lack heterogeneity. The exposure rates of individuals, therefore, cannot be accurately differentiated. In addition, the different models that are used to simulate distinct diseases, such as patch models for measles, distance-transmission models for foot-and-mouth disease, or multi-group models for pandemic influenza, can be transformed into network-based models [4].

Consequently, previous studies of network-based information diffusion and disease transmission are reviewed in this chapter. The process of information diffusion over a network is very similar to disease transmission at an individual level. This chapter first introduces
the models that are used to simulate information diffusion. Then discusses the link prediction problem, which is somewhat analogous to the network inference problem. Studies of inferring information diffusion networks are also reviewed. In the second part, an analysis of the frameworks and functionalities of two popular disease transmission models is presented. The detailed reasons for focusing on network-based disease transmission are given and related work is reviewed. Recent studies of inferring disease transmission networks at an individual level are also reviewed. Finally, a summary of the previous research in these fields is given.
2.1 Information Diffusion Network Inference

Most studies of information diffusion are concerned with maximizing the spread of information [20] or detecting outbreaks of hot topics [21]. The process of information diffusion over networks, for example, computer virus, viral marketing, etc. is very similar to network-based disease transmission. This section reviews related studies, and then demonstrates that the methods of inferring information diffusion networks cannot be used to infer disease transmission networks.

2.1.1 Information Diffusion Models

To simulate information diffusion over networks, nodes are usually defined as either informed or uninformed. An informed node has adopted an innovation, re-posted some new information, and etc., whereas uninformed node has not. In other words, these two states are similar to the infected and susceptible states in disease transmission, respectively. According to previous studies, there are two widely adopted models: the linear threshold and independent cascading models [22] [23].

- In the linear threshold model, nodes are given thresholds with a range between 9 and 1, drawn from a certain probability distribution. For a specific uninformed node $i$, once the sum of the weights over the connected edges of all informed neighbors becomes larger than the threshold that $i$ possesses, it becomes informed.

- In the independent cascading model, each node has a probability value $p$. Once one of a specific node $i$’s uninformed neighbors, for example, node $j$, becomes informed, there will be a trial with the predefined probability value $p$ that node $i$ will become
informed. If node $j$ fails to inform $i$ the first time, it will never try again [24].

### 2.1.2 Link Prediction Problem

Another approach somewhat similar to the network inference problem is link prediction. This problem usually infers network’s future added edges based on a network growth model that depicts the current available network structures. The network model “is useful to the extent that it can support meaningful inferences from observed network data” [25]. Several methods have been proposed [25], and some of the most common are as follows.

- **Graph Distance**: The edge weight is calculated based on the shortest distance between two connected nodes.

- **Common neighbors**: The more overlapping neighbors two nodes possess, the more likely it is that they are connected.

- **Jaccard’s Coefficient**: The edge weight is calculated based on the similarity between two nodes.

- **Preferential Attachment**: Nodes that are newly added to the existing networks have higher probabilities of connecting to those nodes that have more neighbors.

  Newly added nodes are more likely to connect to existing nodes that have more neighbors.

- **Katz**: all the paths connecting two nodes are summed. Each path is multiplied with an exponential function, which gives short paths more weight.
• **Hitting Time**: A random walk is initiated between two nodes. The edge weight between them is affected by the required number of steps.

• **SimRank**: Two nodes are more likely to be connected if they are within the same group or community. In human terms this means that two persons may become friends if they share common interests.

This model-based link prediction problem can also be extended to infer missing links or missing data [26] [27]. Typically, it is used to generate synthetic networks with different topologies, for example, the preferential attachment model [28], Kronecker model [29], duplication model [30], and forest fire model [31].

The main difference between link prediction and network inference is that the former usually starts with an overview of the currently available network structures. Newly added edges are inferred based on these existing structures. In contrast, in the network inference approach introduced in the next section, the networks are completely unknown, and the only available clue is the trace left from information cascading.

### 2.1.3 Diffusion Network Inference

The challenge when inferring an information diffusion network is to “identify the optimal network that best explains the observed infection times, given the times when nodes adopt pieces of information or become infected” [32]. Recently various methods have been proposed to solve this problem in an approximate way.

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 [32]. 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 that is introduced in [20] 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 [33]. In addition, noticing that the structure of a social network is sparse, Myers and Leskovec introduced penalty functions into the objective function to improve its accuracy. 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 [34].

The above-mentioned work has provided insights into how to solve the network inference problem at an individual level. However, they all assume that individuals can only be informed or infected once within a single information cascade (as in the linear threshold model and the independent cascading model [24]). That is to say, if a person is not informed or infected at the first contact, he or she will never become informed or infected. This assumption is not reasonable when inferring disease transmission networks at a metapopulation level, where people in different states (e.g., susceptible, infected, and recovered) are heterogeneously mixed within the same subpopulation. An illustrative example is shown in Fig. 2.1. Given the same synthetic network of $G$, information diffuses in the way shown in Fig. 2.1 (b). It is obvious that there are no cyclic loops during the information propagation. Usually, the information follows a path or a spanning tree. In comparison, disease transmission at a
Figure 2.1: 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.
metapopulation level follows the directed cyclic graph pattern as shown in Fig. 2.1 (c). In addition, nodes within metapopulation-based disease transmission networks have another type of connection; that is self-connected edges representing susceptible subpopulations within each node can be infected by their co-mixed infected subpopulation. Given the above two real-world considerations, methods for inferring information diffusion networks cannot be simply used to infer disease transmission networks at a metapopulation level.

2.2 Disease Transmission Networks

Network-based disease transmission can be classified into two categories, individual and metapopulation-based studies. Although they use the same disease transmission models, their concerns are different. These are discussed separately below, as well as those studies that use surveillance data.

2.2.1 Disease Transmission Models

To simulate disease transmission, various models have been proposed. Among them, the compartmental model and the spatiotemporal model have been widely adopted and investigated.

Two of the most traditional and fundamental compartmental models that characterize the spread of disease are the Susceptible-Infected-Susceptible (SIS) model and the Susceptible-Infected-Recovered (SIR) model [7]. The SIS model is used to simulate infectious diseases with repeated infections; in this model infected people may recover and then quickly become susceptible or lose immunity. Some sexually transmitted diseases, for example chlamydia
follow this its simulation model. Its mathematical form is as follows:

\[
\begin{align*}
\frac{dS}{dt} &= gI - \lambda S \\
\frac{dI}{dt} &= \lambda S - gI,
\end{align*}
\] (2.1)

where \( \lambda \) is the probability of transiting from the susceptible state to the infected state; and \( g \) is the probability of recovering and becoming susceptible again.

In comparison, the SIR model simulates diseases with lifelong or long-term immunity. That is to say, during the period of simulation, the recovered population will not get re-infected. A typical example is pertussis [36]. Its mathematical formula is as follows.

\[
\begin{align*}
\frac{dS}{dt} &= bN - \lambda S - dS \\
\frac{dI}{dt} &= \lambda S - gI - dI \\
\frac{dR}{dt} &= gI - dR,
\end{align*}
\] (2.2)

where \( N \) is the population size; \( b \) is the birth rate; and \( d \) is the death rate. The other parameters are the same as those in the SIS model.

Although these models have been proven to be efficient and successful [7], they neglect many detail, such as the uniqueness of different infectious disease, and the heterogeneity of individuals. Therefore, many derivative models have been developed to reflect the particular transmission processes of distinct diseases and to differentiate various groups of affected populations.

Take the vector-borne disease malaria as an example. It is transmitted between vectors and humans. Traditional SIS and SIR models do not work for this vector-borne disease, because they only describe the state transition processes within a population of either vector or humans. To overcome this limitation, Ross [37] proposed an SIS-SI model (shown in Fig. 2.2), which is now called the Ross model. This model divided humans and vector populations into
Figure 2.2: A graphic representation of the Ross model. The transmission dynamics within the human population is described by an SIS model, indicating that infected people can recover and become susceptible again. Compared to humans, the life span of the vector is much shorter; therefore an SI model is used to describe the transmission dynamics within the vector population. The solid lines indicate the transmission within the same species. The dash lines represent the transmission interaction between humans and vectors.

two groups, and simulated them by combining an SIS model with an SI model. The transmission process of malaria is affected by many coupled impact factors including but not limited to those shown in Fig. 2.3. To study them, epidemiologists and researchers in public health and other fields have proposed several derivative models based on the Ross model [38]. Studies in [39] and [40] added one more compartmental state to describe the incubation time of the malaria parasite. The infected state was further differentiated into two categories intended to model patients with both microscopically and sub-microscopically observable malaria parasite infections [41]. A comprehensive compartmental model considering factors of human immunity, parasite incubation time, asymptomatic infection, symptomatic infection, and etc. was proposed in 2007 [42]. This model provided not only a quantitative analysis, but also a
Figure 2.3: A snapshot of the factors that affect malaria transmission. Based on three time scales (e.g., short, middle and long), we classify them into five categories: malaria parasites, epidemic vector, interventions, socioeconomic factors and climate change. Directed connections between any two impact factors indicate their direct relationships.
framework for applying intervention strategies [43].

Fig. 2.4 gives an example of a model with similar detailed information. It is obvious that there are many more compartments than in the traditional Ross model (Fig. 2.2). The model that is used to simulate human transmission dynamics integrates states that describe symptoms, immunity, detection scale of malaria parasite, and etc., whereas the model describing the vector transmission dynamics integrates states that consider parasite incubation time. The force of infection no longer characterizes only the transmission among infected states and susceptible states between human and vector transmission dynamics. It includes all the transmission probabilities that infectious human (including states of “Semi-immune”, “Asymptomatic infection”, “Symptomatic infection”, “Clear disease”, and “Sub-patent infection”) can infect susceptible vectors.

Another approach to simulate disease transmission is based on a regression model as follows:

\[
\begin{align*}
Y_{it} & \sim Poisson(\mu_{it}) \\
\log(\mu_{it}) &= \beta X + \zeta_{it} + s_i + \epsilon_{it},
\end{align*}
\]

Here, impact factors are treated as input variables \(X\). Taking malaria as an example again, \(X\) may include factors such as temperature, rainfall, or elevation [44] [45] [46] [47]. The output variable \(Y\) is a measurement of disease transmission severity, for example the number of infection cases. \(\beta\) is the coefficients of the input variables. To make accurate and insightful predictions, spatial and temporal components are also embedded in the regression model, which are represented as \(s\) and \(\zeta\), respectively. Usually, \(s\) is modeled using a zero mean multivariate normal distribution. Correlations between different locations are defined as an exponential function decayed with increasing distance; this can be simplified using a
Figure 2.4: An illustrative model that is used to simulate malaria transmission in a population level. Similar to Ross model, it is also divided into two parts. Based on the reflection of people who get infected, the model that describes human transmission dynamics extends the infected state into several compartments in very detail compared with Ross model (Fig. 2.2). As for the vector, a compartment for the latently infected is introduced to simulate the impact factor of parasite incubation time.
indication function (equal to 1 if geographically adjacent, 0 otherwise) [46] [48]. The temporal component $\zeta$ is modeled as a first-order autoregressive process [49] [50]. $\epsilon$ is the error term. Parameter $i$ is the index of the target region, and $t$ is the time step within the simulation. A Bayesian approach is often used to compute the coefficients of the corresponding impact factors [44] [45] [46] [47].

2.2.2 Disease Transmission over Networks

2.2.2.1 Individual-based Disease Transmission Networks

At a metapopulation level, disease transmission networks are representations of the “spatial structure of environment, transportation infrastructures and movement patterns” [1]; at an individual level they also represent social networks. Individual-based disease transmission networks provide an analysis and visualization platform for both contract tracing (“Identify all potential transmission contacts from a source individual” [7]) and infection tracing (All infection cases are linked to form a transmission network [7]). Studies of individual-based disease transmission networks are necessary to: (i) find the influences of network structures on disease transmission, (ii) provide methods of collecting data to construct networks, and (iii) analyze the effectiveness of intervention strategies.

(a). Influences of network structures on disease transmission

The effect of community structures within social networks on the spread of disease has been studied [8]. The synthetic networks generated by either manually adjusted community structures or real-world datasets extracted from Facebook have been used as baseline net-
works in previous studies. These networks are neither directed nor weighted. An individual-based SIR model is used to generate the model of disease outbreaks. Keeling and Eames investigated the effects of different network topologies on disease transmission, and identify the following network structures.

- **Random Networks**: The network edges are formed randomly. There are no obvious clusters or community structures. The nodes are more likely to be homogeneous.

- **Lattices**: Individuals are randomly distributed in two-dimension grids. Those located in adjacent grids are set to be connected. Networks reflect highly localized structures. The landscape of disease transmission is wave-like.

- **Small-World Networks**: Networks combine the properties of both random networks and lattices. A small proportion of connections are randomly created based on lattices, allowing disease transmission over long ranges.

- **Spatial Networks**: Networks are flexible and are adjusted to form the various network topologies mentioned above. The edges are usually formed using connection kernels.

- **Scale-Free Networks**: The network edges are formed based on the preferential attachment, which means that newly added nodes are more likely to attach to those nodes with a large number of neighbors. The degrees of nodes follow the power-law distribution.

Bansal [51] and Asano [52] proposed algorithms to derive individual-based synthetic networks, which are usually treated as the baseline disease transmission networks [53]. Study in [53] indicate that to accurately predict disease outbreaks, it is necessary to incorporate
network clusters and path lengths into the model. The influences of temporal networks on disease transmission were studied in [54]. The edges in the temporal networks were not static. That is to say, they could disappear after a certain amount of time, or just appear temporarily. Therefore, the durations of edges plays a significant role in disease transmission in temporal networks. The longer edges persisted, the more likely it is that infection will occur.

(b). Data collection for constructing transmission networks

In one study, to build a contact networks from real-world practice, Radio Frequency Identification (RFID) devices were distributed to the attendees at a conference [55]. The RFIDs exchanged information whenever two attendees got within a predefined threshold, for example, 1 or 2 meters and a contact edge was formed between them. Another study used wireless sensor networks to reconstruct a social network in a high school [56]. Seven hundred and eighty-eight persons were tested, and people within three meters of each other were set to form an edge. Benavides et al. made use of smartphones with Bluetooth functionalities to generate social contact networks [57]. Data on contact durations and locations were collected and processed to form a transmission network. The SEIR model was applied over the network to investigate the influence of personal contacts on disease transmission.

(c). Evaluation of intervention strategies

Various vaccine strategies, such as random vaccination (select people to vaccinate at random), and targeted vaccination (vaccinate those persons who are more likely to be infected) were simulated in [56] over a real-world contact network that was extracted using wireless sensor network technology. Based on the simulation results, targeted vaccination outper-
formed random vaccination. However, as stated in the article, many medical, social, and ethical factors were not considered. The results of these studies provided policy makers with theoretical analysis only. Two types of vaccination strategies were proposed in [58]. The first one recommended vaccinating people with infectious neighbors, whereas the second one, although similar to the first, suggested giving infectious individuals limited times for infecting other people. A synthetic random graph was combined with the SIR model to generate a disease transmission path, and the basic reproduction number was used as a measurement of the effectiveness of the vaccination strategies [59]. The results showed that isolation outperforms contact tracing when the degree of each node was higher, and the number of cluster structures and length of incubation time was less and shorter, respectively.

### 2.2.2.2 Metapopulation-based Disease Transmission Networks

In comparison, when studying disease transmission networks at a metapopulation level, it is appropriate to simulate disease transmission at both temporal and spatial scales [1] [15]. Moreover, using the metapopulation level overcomes the difficulties of simulating complicated individual human behavior and the unavailability of large amounts of personal information [7] [16] [17]. From the view point of front-line practitioners and policy makers, controlling disease transmission at a metapopulation level is more reasonable and practical [18]. Previous studies of metapopulation-based disease transmission networks have mainly focused on (i) finding the effects of network structures on disease transmission, (ii) analyzing network invasion thresholds and disease behavior caused by spatial heterogeneity, and (iii)
investigating effective intervention strategies.

(a). *Inferring the structures of disease transmission networks*

A metapopulation-based network can be represented in an undirected form, because populations may return to their starting points after a period of time [60]. The influence of directness on disease transmission was discussed in [17]. Moreover, the module which was defined as “a set of densely interconnected nodes that are only sparsely connected to nodes of other modules” was also investigated in this study. The metapopulation network was generated in a synthetic way. Their results showed that modularity could delay disease outbreaks, and affect the scales of outbreaks. As for the directness, it increased the size of the disease outbreaks of whether modularity existed in the given metapopulation-based network.

(b). *Network invasion thresholds caused by spatial heterogeneity*

Theoretical analyses were conducted in [1] to study the disease invasion thresholds of metapopulation-based networks. This study provided a mathematical approach to computing the global invasion threshold, which “determines the minimum number of individuals traveling among subpopulations in order to have the infection of a macroscopic number of subpopulations” [1]. Moreover, their results indicated that the invasion threshold for metapopulation-based disease transmission was strongly related to the disease dependent parameters and the edge weights of networks. A similar invasion threshold analysis of metapopulation-based networks was conducted in [61]. Colizza and Vespignani aimed to find the relationship between the global invasion threshold (computed based on reproduction number) and edge weights. Their results showed that when the threshold value was low,
disease outbreaks only occurred if the edge weights were high.

A mathematical study was conducted in [62], which analyzed the reasonability of the decisions that people made during disease outbreaks. Two strategies were discussed: safe moves and unsafe moves. People who practiced safe moves tried to avoid cities with severe disease outbreaks on their routes. Unsafe moves were the opposite behavior. Based on the simulations over a synthetic network and the SIR model, their results described an interesting phenomenon. Safe moves, which were reasonable for individuals, caused outbreaks throughout the whole network. In contrast, unsafe moves only caused local endemics. Rozhnova et al. used a multiplied cosine function to embed the seasonality factors into an SIR model, and conducted experiments over a metapopulation-based network. Their theoretical analysis and results showed that the combination of spatial heterogeneity rose by metapopulation and the seasonality factors could significantly increase the persistence of disease transmission. The agent-based model and the metapopulation-based model were compared in [16] to analyze their differences in a simulation of disease transmission in Italy. To construct the metapopulation network, high-resolution census data was used with traffic data.

(c). Effectiveness of intervention strategies

The extension of network-based disease transmission from the individual level to the metapopulation level poses a dilemma. How should the limited intervention resources be distributed among spatial subpopulations (i.e., should they be preferentially distributed to severer regions or low level regions) [63]. This dilemma was discussed in [63]. A metapopulation-based network with only two nodes was considered as the fundamental disease transmission network. In contrast to the common sense notion that more resources should be allocated
to more severely affected regions, this study suggested that resources should be deployed to the region with the largest number of susceptible individuals. Optimizing intervention strategies were also discussed in [18]. Ndeffo et al. tried to balance social equitability with efficiency. The experimental results suggested that the tradeoff was highly dependent on the social contexts.

2.2.3 Inference Based on Surveillance Data

As the objective of this study is to infer disease transmission networks purely from surveillance data, this section reviews existing studies that have used surveillance data. These studies have examined (i) the development of surveillance software, (ii) computational modeling, detection and prediction, and (iii) the construction and inference of disease transmission networks.

(a). Surveillance software

The online software package called HealthMap, shown in Fig. 2.5, was designed to improve health surveillance by making use of online information resources [64]. Its input data consists of media news, validated official alerts, and etc. [65]. These data sources are organized along various dimensions, and analyzed using text-processing algorithms. The results are clustered based on geographical locations and disease types, and overlaid on a map for interactive manipulations. Similar event-based surveillance systems are EpiSPIDER, Global Public Health Intelligence Network, etc. There was a comprehensive examination of them in [66].

A disease surveillance software program called SaTScan provided a technique for scan-
Figure 2.5: We use the infectious disease of H7N9 as an example to display the interactive interface of HealthMap. As shown in the figure, all the infection cases are classified based on their states (i.e., confirmed or suspected, death or not death) with distinct colors. At the bottom of (a), all the related alert news are listed. Once a user clicked the news, as shown in (b), a pop-up panel will appear and display the corresponding details.
Figure 2.5: We use the infectious disease of H7N9 as an example to display the interactive interface of HealthMap. As shown in the figure, all the infection cases are classified based on their states (i.e., confirmed or suspected, death or not death) with distinct colors. At the bottom of (a), all the related alert news are listed. Once a user clicked the news, as shown in (b), a pop-up panel will appear and display the corresponding details.
Figure 2.6: An example of utilizing SatScan to analyze West Nile virus incidence cases in US [67]. Both spatial and temporal dynamic changes of infection cases are captured by SatScan. Different levels of severities are displayed.
ning both temporal and spatial scales [68]. As shown in Fig. 2.6, this helped to identify spatiotemporal clusters of disease. Poisson, Bernoulli, and various other models were used to simulate the infection cases. The application of this procedure could aid policy makers in detecting severe disease epidemic locations and time intervals. Spatiotemporal scan statistics have been widely used in disease surveillance. The computations can use univariate and multivariate scan statistics based on both the observed and expected number of infection cases [69]. In terms of visualization, the results of spatiotemporal scan statistics are displayed as geographical overlaid cylinders, with the height indicating the time duration, and the radius representing the severity.

(b). Computational modeling, detection, and prediction

In a project designed to provide the timely detection of disease outbreaks, the surveillance data of neurological chief complaints (CCs) was collected from the emergent department during the 6/1/02-5/31/03 to train a prediction model [70]. The model was based on a least square regression, and incorporated time lag, long-term effect, seasonal effect, etc. The results showed that the proposed model was accurate and gave timely predictions for the outbreaks of CCs in the following years. A stochastic inference model that captured both spatial and temporal variations was carried out in [71]. Various impact factors were integrated into a Poisson likelihood function. In this approach, surveillance data was often used as an application of the proposed models, and the major research focus was to analyze significant impact factors. Inspired by web search query data, a search-term based surveillance tool was used in [72]. Influenza query data was obtained from Yahoo!. Only data from the US was included, and the data processing procedure was given in the article. A linear model,
with additional variables of time trends over, characterized the relationship between cases of influenza and the corresponding search query. Their results accurately predicted influenza outbreaks 1-3 weeks in advance.

(c). Transmission network construction and inference

To study disease spread of livestock, data recording the movements of livestock in the UK were used to build a directed contact network [73]. In this contact network, nodes represented independent markets, and edges represented the movements of livestock. By comparing the network properties of the constructed contact network with random networks, Volkova et al. showed that the constructed contact network had small-world effects. Moreover, during the intense trading season, the contact network became more structured. By running a SEI model over the contact network, intervention strategies that targeted control over highly connected nodes were found to be efficient. Inferring both individual-based and metapopulation-based transmission networks has seldom been done before. A recent study [74] inferred 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. To estimate the parameters in the model, they applied a Markov chain Monte Carlo procedure. Moreover, they also tested this method with a real-world dataset, which had 264 individual infection cases collected at a university hospital over a five-year period.

It is obvious that almost all these studies (except [74]) assume that disease transmission networks are given beforehand, or alternatively, can be constructed from resources reflecting single impact factor. Compared to real-world situations in which the transmission pathways...
are hidden and there are many interdependent impact factors, the above assumption is not reasonable. In addition, no previous studies have considered the practice of inferring disease transmission networks at a metapopulation level. From the view point of front-line practitioners and policy makers, practical applications of theoretical studies are necessary. Therefore, this thesis inversely infers metapopulation-based disease transmission networks from real-world based surveillance data.

2.3 Summary

This chapter reviewed existing studies on the problems of inferring information diffusion networks and disease transmission networks. The reasons for focusing on network-based disease transmission were as follows:

- individuals are allowed to have distinct degrees of contacts, and few of them are permitted to infect the whole network;

- to produce heterogeneity within the whole population, the exposure rates are differentiated; and

- network-based models can be transformed into other simulation models, such as patch models, distance-transmission models and etc.

Previous research in the domain of information diffusion was reviewed and the simulation models widely used in information diffusion, (i.e., linear threshold model and independent cascading model), was proposed for inferring disease transmission networks at a metapopulation level. For simplicity, this is summarized in Table. 2.1.
Table 2.1: Differences and considerations of different network-based studies

<table>
<thead>
<tr>
<th>Level</th>
<th>Information Diffusion</th>
<th>Individual Based Disease Transmission</th>
<th>Metapopulation Based Disease Transmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Re-infection within one cascade</td>
<td>[20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34]</td>
<td>[7] [8] [16] [17] [54] [55] [56] [57] [58] [59] [74]</td>
<td>[1] [15] [16] [17] [18] [60] [61] [62] [63] [73] [75]</td>
</tr>
<tr>
<td>Individual behaviours</td>
<td>Usually no</td>
<td>Depending on disease types</td>
<td>Yes</td>
</tr>
<tr>
<td>Self-connected edges</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Spread over Directed Cyclic Graphs</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Simulation in a spatiotemporal scale</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Data collection</td>
<td>Difficult</td>
<td>Difficult</td>
<td>Easy</td>
</tr>
<tr>
<td>Front-line practice and policy making</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Network inference methods</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

In the second part of this chapter, two traditional disease transmission models, the compartmental model and the regression based spatiotemporal model were introduced. Previous
studies of both individual and metapopulation-based disease transmission networks were reviewed; most of these studies assumed that the disease transmission networks are given in advance, or can be constructed by considering only one type of resources (e.g., transportation network, live stock movements). However, in reality, this is not the case. Transmission networks are the outcome of multiple interactions between various impact factors. Significantly, in real-world situation they are hidden. That is to say, what people observe is only the surveillance data at both spatial and temporal scales. People do not know how the disease transmits between locations, or when it will be resurgent.

These assumptions and limitations create a gap between theoretical studies of disease transmission and the real-world need for practical disease control. Therefore, this thesis inversely infers disease transmission networks using only spatiotemporal surveillance data.
Chapter 3

Problem Statement

In this chapter, we define our key terms and give illustrative examples. Then we discuss the disease transmission model we use to infer metapopulation-based disease transmission networks. The model is a generalized linear transmission model that considers all the transmission pathways between or within subpopulations, including the external influence component, internal transmission component and neighborhood transmission component. Their abstracted representations in the transmission networks and mathematical forms are shown as well. Finally, we formulate the problem of inferring metapopulation-based disease transmission networks.
3.1 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. Using Fig. 3.1 as an example, suppose an infectious disease transmits over a part of this network; then the aggregated surveillance data of this network would have a table form like that shown in Table 3.1. Note that the data are generated for demonstration purposes and are based on only a portion of the whole disease transmission network.

After aggregating infection cases based on locations and infection times, a dataset $D = \{ <v_i, ic_i, t_i> | i = 0, 1, 2, ..., N, t \in T \}$ is collected (example shown in Table. 3.2). $i$ 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 $V = \{ v_i | i = 0, 1, 2, ..., N \}$ 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 [76]). $v_i (i = 1, 2, ..., N)$ correspond to the rest of nodes within the target region.
Table 3.1: Surveillance data generated based on the synthetic disease transmission network shown in Fig. 3.1 (sorted based on infection time).

<table>
<thead>
<tr>
<th>Name (id_p)</th>
<th>Infection Time (it_p)</th>
<th>Infection Location (loc_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported/confirmed case 1</td>
<td>1</td>
<td>Node 1</td>
</tr>
<tr>
<td>Reported/confirmed case 2</td>
<td>2</td>
<td>Node 2</td>
</tr>
<tr>
<td>Reported/confirmed case 3</td>
<td>2</td>
<td>Node 2</td>
</tr>
<tr>
<td>Reported/confirmed case 4</td>
<td>2</td>
<td>Node 3</td>
</tr>
<tr>
<td>Reported/confirmed case 5</td>
<td>2</td>
<td>Node 3</td>
</tr>
<tr>
<td>Reported/confirmed case 6</td>
<td>2</td>
<td>Node 4</td>
</tr>
<tr>
<td>Reported/confirmed case 7</td>
<td>2</td>
<td>Node 5</td>
</tr>
<tr>
<td>Reported/confirmed case 8</td>
<td>3</td>
<td>Node 1</td>
</tr>
<tr>
<td>Reported/confirmed case 9</td>
<td>3</td>
<td>Node 1</td>
</tr>
<tr>
<td>Reported/confirmed case 10</td>
<td>3</td>
<td>Node 2</td>
</tr>
<tr>
<td>Reported/confirmed case 11</td>
<td>3</td>
<td>Node 2</td>
</tr>
<tr>
<td>Reported/confirmed case 12</td>
<td>3</td>
<td>Node 2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[ E = \{ e_i \mid i = 1, 2, \ldots, N \} \] denotes the set of directed edges with different weights \( W = \{ w_i \mid i = 1, 2, \ldots, N \}. \)
\( e_i = \{ e_{ji} \mid j = 0, 1, 2, \ldots, N \} \) is the set of incoming links for node \( i \) and \( w_i = \{ w_{ji} \mid j = 0, 1, 2, \ldots, N \} \) 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 \textit{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 Fig. 3.1). As the data de-
Table 3.2: Preprocessed surveillance data generated based on Table. 3.1 (sorted based on infection time)

<table>
<thead>
<tr>
<th>Number of Infection Cases ( (ic) )</th>
<th>Node 0</th>
<th>Node 1</th>
<th>Node 2</th>
<th>Node 3</th>
<th>Node 4</th>
<th>Node 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection Time ( (t = 1) )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Infection Time ( (t = 2) )</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Infection Time ( (t = 3) )</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Infection Time ( (t = 4) )</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Infection Time ( (t = 5) )</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>Infection Time ( (t = 6) )</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>16</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Infection Time ( (t = 7) )</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>21</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>Infection Time ( (t = 8) )</td>
<td>1</td>
<td>1</td>
<td>14</td>
<td>29</td>
<td>7</td>
<td>35</td>
</tr>
<tr>
<td>Infection Time ( (t = 9) )</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>37</td>
<td>8</td>
<td>44</td>
</tr>
<tr>
<td>Infection Time ( (t = 10) )</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td>46</td>
<td>9</td>
<td>54</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

scribes a real-world situation, the assumption is that infected people can infect susceptible people within the same subpopulation (shown in Fig. 2.1). 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. Here, these two types of external transmission effects are differentiated into two categories. We define them respectively as the neighborhood transmission component and the external influence component. Their definitions are given as follows.

**Definition 2. (Internal Transmission Component):** Within each node (subpopulation), pre-
Figure 3.1: 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).
viously infected people may correlate to newly infected people without outside disturbances. This component is disease independent. Air-borne diseases such as influenza, vector-borne diseases such as malaria, and other infectious diseases all have this property. In $G$, for node $i$, this component is represented as an edge linking to itself with weight $w_{ii}$.

**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 [77]. 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 $w_{0i}$.

### 3.2 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 [15], 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:

\[
\begin{align*}
  itc_i^t &= w_{ii} \times ic_i^{t-1} \\
  ntc_i^t &= \sum_{j} w_{ji} \times ic_j^{t-1} \\
  eic_i^t &= w_{0i} \times ic_0^{t-1},
\end{align*}
\]  

(3.1)

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 \( \varepsilon \). \( \varepsilon \) is used to capture unpredicted biases. The assumption is that it follows a zero-mean normal distribution, \( \varepsilon \sim N(0, \beta) \) and is the same for all nodes:

\[
\begin{align*}
  ic_i^t &= itc_i^t + ntc_i^t + eic_i^t + \varepsilon \\
        &= w_{ii} \times ic_i^{t-1} + \sum_{j} w_{ji} \times ic_j^{t-1} + w_{0i} \times ic_0^{t-1} + \varepsilon.
\end{align*}
\]  

(3.2)

Equations 3.1 and 3.2 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 treatments and interventions; 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.
3.3 Network Inference Problem

To inversely infer the existence of the edges within the hidden disease transmission network $G$ and their corresponding weights $W = \{w_i \mid i = 0, 1, 2, ..., N\}$, given an observed surveillance dataset $D = \{(v_i, ic_i, t_i) \mid i = 0, 1, 2, ..., N, t \in T\}$. It is not accurate to infer disease transmission networks by following the cascading process in the information diffusion [32]. Because the disease transmission process at the metapopulation level does not follow the Directed Acyclic Graphs pattern (example shown in Fig. 2.1).

Therefore, to recover the network structure $G$, it is necessary to first write the likelihood function for a specific node $i$ based on Eq. 3.2:

$$L(w_i, \beta \mid ic_i) = \prod_{t=1}^{T} \frac{1}{(2\pi\beta)^{1/2}} e^{-\frac{1}{2\pi}(ic_i^t - w_{ii} \times ic_{i}^{t-1} - \sum_{j} w_{ji} \times ic_{j}^{t-1} - w_{0i} \times ic_{0}^{t-1})}, \quad (3.3)$$

where all the parameters are the same as in Eq. 3.2, except we use $N_i^*$ here rather than $N_i$ to indicate the number of estimated neighbors of node $i$ within the inferred network $G^*$. This set of neighbors can be written as $V_i^* = \{v_j \mid j = 0, 1, 2, ..., N$ and $w_{ji} \neq 0\}$. $\beta$ is the variance of the normal distribution for the error term $\varepsilon$. From this equation, the network inference problem is transformed 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 likelihood function becomes:

$$L(W, \beta \mid D) = \prod_{i=1}^{N} L(w_i, \beta \mid ic_i), \quad (3.4)$$

and the objective is to maximize the likelihood function shown in Eq. 3.4. To evaluate the estimated network $G^*$, precision-recall curves are used. Specifically, both the existences of
edges and their corresponding weights in the synthetic network \( G \) and the estimated network \( G^* \) are compared.

### 3.4 Summary

This chapter defined disease transmission networks. Then, considering the real-world situation of disease transmission, the traditional SI model was extended from the individual level to the metapopulation level to build a linear transmission model that incorporates all the possible transmission pathways. These are (i) the internal transmission component, which recognizes the correlations between previously infected people and current infected people within the same subpopulation, and thus has self-connected edges, (ii) the neighborhood transmission component, which describes the temporal correlations among infected people that distribute in different subpopulations, and (iii) the external influence component, which describes the introduction of imported cases and has edges starting from an external node that are connected to all the other nodes. Their mathematical forms were given as well. An error term that follows the zero-mean normal distribution is embedded to capture unpredicted biases. Finally, the problem of inferring metapopulation-based disease transmission networks was formulated. Specifically, given the preprocessed surveillance dataset \( D = \{<v_i, ic_i, t_i> | i = 0, 1, 2, ..., N, t \in T\} \), the inference problem was transformed into a problem of maximizing a likelihood function.
Chapter 4

The Proposed Network Inference Method

In this chapter, we propose a network inference method called NetEpi. As a first step, we use the pairwise correlation analysis to remove a proportion of the false positive edges in network $G$. However, the properties of disease transmission over networks highly bias the results by concluding that there are high correlations between nodes without connections. We consequently adopt a first-order partial correlation analysis to overcome this problem and to obtain approximate partial correlation networks. In the second step, we use an Expectation-Maximization based Bayesian learning method combined with a back-tracking technique to further fine tune the partial correlation networks, and estimate their corresponding edge weights. At the end of this chapter, we describe the stopping criteria of NetEpi, and analyze its time complexity.
4.1 Basic Ideas

As discussed in previous chapters, the objective of this research is to infer the hidden disease transmission network $G$ at a metapopulation level using only the surveillance data. Therefore, for our proposed method, NetEpi, the only input is the preprocessed dataset $D$. In Fig. 4.1, the framework of NetEpi is presented, including the data preprocessing procedure. Note that the procedure for using the first-order partial correlation to obtain the approximate network structures is centralized compared to the back-tracking Bayesian learning. In the second step, the edge connections for each individual node is inferred and their results are combined to form the final estimated disease transmission network $G^*$. For a clear understanding of the framework, the second step of NetEpi is presented separately in Fig. 4.1. In the following sections, we describe NetEpi in detail.

4.2 Partial Correlation Network Construction

Using dataset $D$, an approximate network structure over which the disease transmission is constructed. This step reduces the trivial computation cost for the algorithm in our next step and removes a proportion of the false positive edges. Correlation analysis is used to do this. In the following, the commonly used method for correlation analysis, which is the Pearson correlation, is illustrated. Then the illustrations for inferring disease transmission networks are demonstrated, and a first-order partial correlation analysis to address these limitations.
Figure 4.1: A flowchart of NetEpi. The first step is to obtain approximate network structures based on the first-order partial correlation analysis. The second step is to carry out the Back-Tracking Bayesian Learning.
Figure 4.1: A flowchart of NetEpi. The first step is to obtain approximate network structures based on the first-order partial correlation analysis. The second step is to carry out the Back-Tracking Bayesian Learning.
Figure 4.2: The possible transmission relationships among three nodes [78]. 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. (c) shows the possible correlation results among nodes $i$, $j$, and $k$.

4.2.1 Pearson Correlation Analysis

The Pearson correlation coefficient $\rho_p$ is a measurement of the dependence between two variables $X$ and $Y$. Its results are easy to interpret. If $\rho_p$ approximates $+1$, then $X$ and $Y$ are positively proportional. If $\rho_p$ approximates $-1$, then if $X$ increases, $Y$ decreases. If $\rho_p$ approximates $0$, $X$ and $Y$ are almost independent. Its mathematical form is:

$$\rho_p = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2 \sum_{i=1}^{N} (Y_i - \bar{Y})^2}}. \quad (4.1)$$

Although the Pearson correlation is used to analyze the correlation between two selected nodes $i$ and $j$, a problem arises in the analysis of disease transmission networks. As shown in Fig. 4.2 (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 as illustrated in in Fig. 4.2
(c). The same problem exists in the case of Fig. 4.2 (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, as described in the following section.

4.2.2 First-Order Partial Correlation Analysis

The first-order partial correlation is a measurement of the dependence between two vari-
ables $X$ and $Y$, after removing or fixing a third variable $Z$. To compute it between nodes $i$
and $j$, the effect of another node $k$, where $k = 0, 1, 2, ..., N$, 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 [78] [79]. Therefore, to infer a directed
relationship, in this study, we analyze the partial correlation with a time lag. (It is not limited
to a one-unit time lag; other options are allowed. This thesis uses a time lag of one unit as an
example.) 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). The direction is 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:

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

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

The algorithm for constructing the partial correlation networks is given in Alg. 1. By using this method, the influences of the intermediate node and parent node (node $k$ shown in Figs. 4.2 (a) and (b)) can be removed. Therefore, a proportion of false positive solutions can be removed, and an approximate partial correlation network $G^p$ can be obtained.

**Algorithm 1 Construct partial correlation networks**

**Input:** $D$: Preprocessed surveillance dataset; $\rho_0$: Correlation coefficient threshold;

**Output:** $G^p$: Partial correlation network;

1: Divide $D$ into two subsets with time lag of one time unit;

2: Compute pairwise pearson correlation coefficients;

3: for all $i = 0, 1, 2, \ldots N$ do

4: for all $j = 0, 1, 2, \ldots N$ do

5: for all $k = 0, 1, 2, \ldots N$ do

6: Compute first-order partial correlation coefficient $\rho$ based on Eq. 4.2;

7: if $\rho \geq \rho_0$ then

8: Put $k$ into the neighborhood list of node $i$;

9: Combine all neighborhood lists to form $G^p$;

10: return $G^p$;
4.3 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 [80]. In the proposed inference method, the Bayesian learning is based on the Sparse Bayesian Learning (SBL) framework [81]. Related work has been widely and well reported in signal processing studies [80]. 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.

4.3.1 Marginal Likelihood Function

For a specific node $i$, the preprocessed surveillance dataset $D$ is divided into two subsets: an $M \times 1$ vector of $y = \{<v_i, ic_i, t_i> | t_i = 2, 3, ..., M + 1, M \in T \}$ and a $M \times |N^p|$ matrix of $x = \{<v_j, ic_j, t_j> | j \in N^p, t_j = 1, 2, ..., M, M \in T - 1 \}$. $M$ is the size of output variable $y$ and input variable $x$. $N^p$ 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 $y$, $x$, and other parameters. If not specifically stated, all the parameters are formulated for node $i$. Here, we use a time lag of 1 between $y$ and $x$. The relationship between $y$ and $x$ can be formulated based on the
generalized linear transmission model introduced in section 3.2 as follows

\[ y = xw^T + \varepsilon, \]  

(4.3)

where \( w = \{ w_j \mid j \in N^p \} \) is a vector indicating all possible incoming links estimated based on \( G^p \). \( \varepsilon \) is an error term. Under the framework of SBL, both \( w \) and \( \varepsilon \) follow a zero-mean Gaussian distribution with variances of \( \alpha \) and \( \beta \), respectively. They are defined as:

\[ p(w|\alpha) = \prod_{j=1}^{N^p} N(w_j | 0, \alpha_j^{-1}), \]  

(4.4)

\[ p(\varepsilon) = N(0, \beta). \]  

(4.5)

Because there is no prior knowledge of \( w \) and \( \varepsilon \), it is reasonable to set them with non-informative prior distributions, such as a Gamma distribution. Here, \( \alpha \) and \( \beta \) are assumed to have the same hyperparameters for all nodes.

Given the observation data \( y \) and the prior distribution \( \alpha \) and \( \beta \), the posterior distribution of \( w \) is:

\[
p(w|y, \alpha, \beta) = \frac{\text{likelihood} \times \text{prior}}{\text{normalize factor}}
= \frac{p(y|w, \alpha, \beta)p(w|\alpha, \beta)}{p(y|\alpha, \beta)}
= \frac{p(y|w, \beta)p(w|\alpha)}{p(y|\alpha, \beta)},
\]  

(4.6)

which is a Gaussian distribution \( N(\mu, \Sigma) \) with

\[
\mu = \beta^{-1} \Sigma x^T y
\]  

(4.7)

\[
\Sigma = (\Lambda + \beta^{-1} x^T x)^{-1},
\]  

(4.8)

where \( \Lambda = \text{diag}(\alpha_1, \alpha_2, ..., \alpha_{N^p}) \). “Type-II maximization likelihood” maximization combined with a maximum a posteriori probability (MAP) estimate [81] transforms the whole
problem into the following marginal likelihood function:

\[ p(y|\alpha, \beta) = \int p(y|w, \beta)p(w|\alpha)dw. \]  

(4.9)

4.3.2 Expectation-Maximization Computation

Writing Eq. 4.9 into a logarithm form \( \mathcal{L}(\alpha) \), we have:

\[
\mathcal{L}(\alpha) = \log p(y|\alpha, \beta) = \log \int p(y|w, \beta)p(w|\alpha)dw = \frac{1}{2} [M \log 2\pi + \log |C| + y^T C^{-1} y]
\]

(4.10)

with

\[
C = \beta I + x \Lambda^{-1} x^T
\]

(4.11)

The derivatives of Eq. 4.10 with respect to \( \alpha_j \) and \( \beta \) are [82]:

\[
\frac{\partial \mathcal{L}(\alpha)}{\partial \log \alpha_j} = \frac{1}{2} (1 - \alpha_j \Sigma_{jj} - \alpha_j \mu_j^2)
\]

(4.12)

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

(4.13)

Setting Eqs. 4.12 and 4.13 to zero, the estimations of \( \alpha_j \) and \( \beta \) become:

\[
\alpha_j^{\text{new}} = \frac{1 - \alpha_j \Sigma_{jj}}{\mu_j^2}
\]

(4.14)

\[
\beta^{\text{new}} = \frac{M - \sum_{j=1}^{J} (1 - \alpha_j \Sigma_{jj})}{\|y - x\mu\|^2}
\]

(4.15)

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.
4.3.3 Back-Tracking Technique

In the disease transmission network $G$, only positive links indicating the existence of transmission pathways exist. However, the prior distribution shown in Eq. 4.4 may cause $w$ to be negative. To avoid this, a constraint limiting $w$ 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 $\mu$, $\Sigma$, $\alpha_j$, and $\beta$ are updated, any $\alpha_j$ 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 this procedure is shown in Alg. 2.

4.4 Discussions

4.4.1 Stopping Criteria

As stated in [32], it is not trivial nor practical to find all the edges within the network $G$, or the exact time required to stop the inference program. Thus, once the program iterates to the maximum permitted steps, or the update of the marginal likelihood function converges to a significantly small value, the learning procedure is stopped.
Algorithm 2 Back-Tracking Bayesian Learning

Input: $D$: Preprocessed surveillance dataset; $G^p$: Partial correlation network;

Output: $G^*$: Inferred disease transmission network;

1: Divide $D$ into two subsets with time lag of one time unit;

2: for all node $i = 0, 1, 2, ..., N$ do

3: Initialize parameters for prior distributions;

4: Construct marginal likelihood function $p_i(y|\alpha, \beta)$ (shown in Eq. 4.9);

5: while not reaching stopping criteria do

6: for all node $j \in N^p$, and $i \neq j$ do

7: Compute contributions to $p_i(y|\alpha, \beta)$;

8: Select node with maximum contribution;

9: Re-estimate all weights of current neighbors of node $i$;

10: if all weights are not less than zero then

11: Update neighborhood list;

12: else

13: Remove neighbors with weights less than zero, and put them into blacklist;

14: Roll back $p_i(y|\alpha, \beta)$;

15: Combine all neighborhood lists to form $G^*$;

16: return $G^*$;
4.4.2 Computational Complexity of the Algorithm

To compute the partial correlation network, the time complexity is $O(N^3)$ without counting the time required to compute the mean and variance of each time-series data. To speed up this process, dynamic programming is used to recursively compute the first-order partial correlation based on the result of the Pearson correlation coefficients or the zeroth-order partial correlation coefficients. In addition, once the time-series data of $D$ is large, a Discrete Fourier Transformation (DFT) [83] can be used to compress the data and shorten the computation time, by avoiding redundant computations.

The complexity of Bayesian learning is mainly a result of the computation of the parameters of $\Sigma$, which requires $O(N^3)$. The efficient incremental algorithm proposed in [84] can optimize this computation. In addition, the computation based on the partial correlation network $G^p$ can also reduce the computational time of this process. The induced computation needed for each node is significantly reduced from the number of all nodes in $G$ to those that are estimated neighbors in $G^p$. After integrating the back-tracking algorithm, the time complexity becomes exponential. However, based on the experiments in this thesis, the algorithm usually converges quickly. That is to say, the algorithm seldom tracks back to the nodes that are first selected, because the previous Bayesian learning selects those significantly contributing nodes at the very beginning, making the marginal likelihood function converge to a near optimum solution without needing a large space to increase; moreover, the process will be stable until it reaches the stopping criteria.
4.5 Summary

At the beginning of this chapter, the objective of this thesis, which was to infer hidden disease transmission networks at a metapopulation level using only a real-world surveillance dataset was emphasized. A metapopulation-based inference method, called NetEpi, which can overcome the problem of transmission over Directed Cyclic Graphs was proposed. NetEpi compromises two steps. The first step is to construct an approximate network structure that can remove a proportion of the false positive edges in the final estimated graph $G^*$. A first-order partial correlation analysis was used to avoid the problem caused by intermediate nodes or same parent nodes in the disease transmission process. In the second step of NetEpi, an Expectation-Maximization based Bayesian learning method combined with a back-tracking technique was used to further fine tune previous inferred edges and their corresponding weights. The stopping criteria and time complexity analysis were presented at the end of the chapter.
Chapter 5

Experiments

In this chapter, the proposed metapopulation-based network inference method, NetEpi, is tested on both synthetic and real-world datasets. The processes and settings that are used to produce the synthetic data are introduced. Two hundred and forty sets of synthetic data are generated with distinct topologies, sizes, and other control parameters. Then NetEpi is compared to a probability-based baseline method. The measurements of precision-recall curves, parameters sensitivity, etc. are taken to judge the performance of NetEpi and the baseline method. The results show that NetEpi outperforms the probability-based baseline method for all datasets. In the second section of this chapter, NetEpi is applied to a real-world dataset, specifically to the surveillance data for malaria transmission in Yunnan Province, China in 2005. Combined with the profiles of the geographical locations, the results of the inferred malaria transmission network can be used to interpret the local transmission pattern to a certain degree.
5.1 Experiments Based on Synthetic Data

To test the performance of NetEpi, synthetic data are used to evaluate the inferred disease transmission network $G^*$ against the ground-truth network $G$. The data generation proceeds as follows: a Kronecker Graphs model [85] is first used to generate a network structure with edges linking only neighborhood nodes (Definition 3). Next, using predefined probabilities, all the nodes are linked to an external node $v_0$, and self-connected edges are generated. Then, the transmission model given in Eq. 3.2 is interactively run for a sufficient number of iterations to generate a synthetic disease surveillance dataset.

5.1.1 Experimental Setting

Three types of Kronecker Graphs [85] are constructed: (i) core-periphery networks (Fig. 5.1), 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 (Fig. 5.2) in which nodes form several small communities that are connected to form one large cluster, and (iii) random graphs (Fig. 5.3), 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. 3.2) is run ten times to generate independent synthetic datasets. For a single dataset, the transmission process is made
Figure 5.1: An example of core-periphery networks with 128 nodes, and about 500 directed edges. For the sake of presentation, the directions of edges are not shown in the figure. As shown in the figure, a cluster of yellow nodes can be observed in the core of the network. The red links are the connections started or ended at those core nodes. Other black nodes with fewer connections are distributed in the periphery area.
Figure 5.2: An example of hierarchical community networks with 128 nodes, and about 500 directed edges. For the sake of presentation, the directions of edges are not shown in the figure. There are several small communities formed within the network. These communities are connected, forming one big cluster.
Figure 5.3: An example of random graphs with 128 nodes, and about 500 directed edges. For the sake of presentation, the directions of edges are not shown in the figure. There are not obvious patterns in this network.
Figure 5.4: 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.

to cover all the edges of $G$. In total, there are three types of network topologies $\times 8$ different sizes $\times 10$ independent transmission processes $= 240$ datasets.

5.1.2 Baseline Method

To our best knowledge, there have 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 directed infection cases at $t = n$ will have connections to those nodes that have infection cases at $n + 1$ (shown
in Fig. 5.4). 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:

$$w_{ij} = \frac{ic_i^t ic_j^{t+1}}{\sum_{i=1}^{N} ic_i^t}.$$  \hspace{1cm} (5.1)

5.1.3 Results

To evaluate the inference results, the precision-recall curves are computed as shown in Figs. 5.5, 5.6, and 5.7. Similar to the definitions in [32], 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.

In core-periphery networks, nodes are located in the core region, as shown in Fig. 5.1. These nodes have more connections than those distributed in the periphery region. Therefore,
Figure 5.5: The precision-recall curves for synthetic core-periphery networks. It is obvious that NetEpi outperforms baseline the method in all cases.
Figure 5.5: The precision-recall curves for synthetic core-periphery networks. It is obvious that NetEpi outperforms baseline the method in all cases.
Figure 5.6: The precision-recall curves for synthetic hierarchical community networks. It is obvious that NetEpi outperforms the baseline method in all cases.
Figure 5.6: The precision-recall curves for synthetic hierarchical community networks. It is obvious that NetEpi outperforms the baseline method in all cases.
Figure 5.7: The precision-recall curves for synthetic random graphs. It is obvious that NetEpi outperforms the baseline method in all cases.
Figure 5.7: The precision-recall curves for synthetic random graphs. It is obvious that NetEpi outperforms the baseline method in all cases.
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 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 (shown in Fig. 5.3). 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:

\[
    d_{\text{avg}} = \frac{\sum_{i=1}^{N} d_i}{N},
\]

where \(d_i\) is the out-degree for node \(i\) and \(d_{\text{avg}}\) is the average out-degree for the whole network.

The out-degree statistics for all the 24 synthetic networks are listed in Table. 5.1.

For networks with the same topologies but a different number of nodes, NetEpi achieves better results when inferring smaller networks, as shown in Figs. 5.5, 5.6, and 5.7. 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
Figure 5.8: 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. 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.
Table 5.1: Out-degrees for synthetic networks

<table>
<thead>
<tr>
<th>Size</th>
<th>Core-Periphery Network</th>
<th>Hierarchical Community Network</th>
<th>Random Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 nodes, 100 edges</td>
<td>1.4154</td>
<td>1.5385</td>
<td>1.6923</td>
</tr>
<tr>
<td>64 nodes, 150 edges</td>
<td>1.7385</td>
<td>1.8308</td>
<td>1.8923</td>
</tr>
<tr>
<td>128 nodes, 180 edges</td>
<td>1.3876</td>
<td>1.4496</td>
<td>1.4651</td>
</tr>
<tr>
<td>128 nodes, 200 edges</td>
<td>1.5504</td>
<td>1.6357</td>
<td>1.5969</td>
</tr>
<tr>
<td>256 nodes, 350 edges</td>
<td>1.3619</td>
<td>1.5175</td>
<td>1.5097</td>
</tr>
<tr>
<td>256 nodes, 400 edges</td>
<td>1.5525</td>
<td>1.6537</td>
<td>1.6615</td>
</tr>
<tr>
<td>512 nodes, 720 edges</td>
<td>1.4016</td>
<td>1.5107</td>
<td>1.5029</td>
</tr>
<tr>
<td>512 nodes, 800 edges</td>
<td>1.5439</td>
<td>1.6199</td>
<td>1.6686</td>
</tr>
</tbody>
</table>

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 Fig. 5.8.

However, inferring metapopulation-based disease transmission networks is different than inferring individual-based information diffusion networks in 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 [86] (two types of vector-borne diseases 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 conducted to get a whole picture of large-size networks. It should be noted that the relationships between the higher-level nodes and the lower-level nodes are very similar to the relationships between metapopulation-based disease transmission networks and individual-based disease transmission networks shown in Fig. 1.1, except that here individual-based disease transmission networks are treated as lower-level metapopulation-based disease transmission networks.

Figure 5.6 shows the prediction results of NetEpi. Figs. 5.6 (a), (c), (e), and (g) show the number of infection cases in each simulation time step. Figs. 5.6 (b), (d), (f), and (h) show the differences between the predicted number of infection cases and the data generated by the synthetic data. As shown, 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. This confirms that NetEpi converges to an optimal solution, although it may not be the global one.

5.1.4 Sensitivity Analysis

The parameters that need to be defined include thresholds for maximizing marginal likelihood function, the number of nodes, the number of edges, the length of time window, and the size of the surveillance data. For the thresholds that are used to maximize the likelihood function, they are set to small values. This is reasonable because we assume the NetEpi approximates to the optimum when those parameters converge to small values and therefore, we do not conduct sensitivity analysis. As for the numbers of nodes and edges, we have already generated many datasets to test their influences.

There are two parameters that play significant roles in the inference results of NetEpi. One
(a) Core-Periphery Network with 64 nodes and 100 edges

(b) Core-Periphery Network with 64 nodes and 100 edges
(c) Hierarchical Community Network with 128 nodes and 180 edges

(d) Hierarchical Community Network with 128 nodes and 180 edges
(e) Random Graph with 256 nodes and 350 edges

(f) Random Graph with 256 nodes and 350 edges
Figure 5.6: NetEpi accurately captures the disease transmission trend.
is the time window that is used in the partial correlation networks. If it is too short, any bias
can distort the results. If it is too long, all the time-series data would be too smooth which
leads to higher similarities between each other. Another one is the number of observations
which is important considering the data quality. Usually, a large amount of data ensures
the reliability and accuracy of the results. However, we may not get a huge amount of data
depending on different applications or limitations. 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:

\[ m_{tw} = \left( \frac{s}{|E| + 1} \right) \frac{|E^p|}{|E| + 1}, \]  
(5.3)

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 con-
tained in \( G^p \), \( m_{tw} \) is the trade-off between \( s \) and \( |E^p| \), and \( tw \) is the selected time window.

Based on the experiments and theoretical analysis, the ranges of \( s \) and \( |E^p| \) are as follows:

\[
\begin{aligned}
    & s \in [0, |E|] \\
    & |E^p| \in [|E|, |E|^2].
\end{aligned}
\]  
(5.4)

Therefore, the value of \( m \) increases as \( s \) increases or \( |E^p| \) decreases. The ideal case is
\( s = |E^p| = |E| \), so that \( m \) significantly approximates 1. It should be noted that when \( s \)
approxBates or equals \( |E| \), \( m \) 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 Figs. 5.7, 5.8, and 5.9, the relationships between trade-off measurement $m$ and time window $tw$ are categorized into four classes.

1. “N” Shape: Examples of this type of relationship are shown in Figs. 5.7 (b), 5.8 (a), (b), (f), 5.9 (g), and (h). The trade-off measurement value in such cases 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. $m$ decreases at the beginning because the increasing rate of $s$ is slow compared to the fast increasing rate of $|E_p|$. $m$ incrementally increases later because stronger correlations are mined out under the conditions of increasing time window values.

2. “S” Shape: Examples of this type of relationship are Figs. 5.7 (a), (c), (d), (e), (f), (g), (h), 5.8 (e), (g), (h), 5.9 (c), (d), (e), and (f). 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 mined out, so that even as the length of time window continues to increase, the number of strongly correlated edges remains stable.
Figure 5.7: Sensitivity analysis for core-periphery networks with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.
Figure 5.7: Sensitivity analysis for core-periphery networks with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.
Figure 5.8: Sensitivity analysis for hierarchical community networks with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.
Figure 5.8: Sensitivity analysis for hierarchical community networks with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.
Figure 5.9: Sensitivity analysis for random graphs with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.
Figure 5.9: Sensitivity analysis for random graphs with different sizes. The horizontal axis is the selected time window with the unit of day. The vertical axis is the measurement value of $m$ computed from Eq. 5.3.
3. **“V” Shape**: Examples of this type of relationship are Figs. 5.8 (c), 5.9 (a), and (b). 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, $m$ 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.

4. **“L” Shape**: There is only one example in the collected cases, Fig. 5.8 (d). The trade-off measurement value achieves the maximum and then decreases to a valley, as in the “V” Shape. However, $m$ does not increase afterwards. After checking the statistics of our experiments, it is clear that this is because the ground-truth edges with strong correlations were found within the 14-day time window, thus there is significantly less variation in the following experiments conducted with longer time window values. The number of false positive edges increases later, making $m$ decrease in the later experiments even though the time window is lengthened.

Another important control parameter is the number of observations (size of surveillance dataset), which is the parameter $M$ mentioned in Section 4.3.1. 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 thesis assumes that the size of surveillance dataset should be at least larger than the length of the time window. Specifically, the detailed relationships between the size of surveillance data $M$, length of selected time window $tw$, number of network nodes $N$ and the scaled parameter $s$ should be as follows:

$$M - tw + 1 \geq \frac{N}{s}$$ (5.5)

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 ($s = 2$) of the number of nodes. Then the size of the training surveillance dataset should at least be 98.

Figures 5.10 and 5.11 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,
Figure 5.10: 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. The curve with the size of a quarter of the number of network nodes is displayed as a blue solid line. The curve 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.
Figure 5.11: 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. 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.
different sizes of surveillance dataset are tested independently. All of them are tested under
the time window of 35. The scaled parameter $s$ 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 Fig. 5.11 (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.

5.2 Experiments Based on a Real-world Dataset

5.2.1 Dataset Description

The real-world dataset was provided by the Chinese Center for Disease Control and Pre-
vention. 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 disease in the Yunnan region. There were 2928 cases reported in 51 town-
ships in 2005. These townships are distributed along the border between China and Myanmar
(a high malaria-endemic country). The data is 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 are further classified into five cate-
gories 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), $(150, 200]$
Figure 5.12: Snapshot of 51 townships located in Yunnan province, China, along the border between China and Myanmar. The placemarks represent townships, which are colored based on their numbers of infection cases: (i) red nodes: $(200, +\infty)$; (ii) purple nodes: $(150, 200]$; (iii) green nodes: $(100, 150]$; (iv) yellow nodes: $(50, 100]$; (v) blue nodes: $(0, 50]$. This picture is created using Google Earth.
5.2.2 Experimental Setting

The dataset is known to be very 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 ~ 17 days [87]. However, studies have also reported that the incubation time can be longer, from several months to several years [87] [88]. Therefore, in this study, 21 days has been selected as the time window for inferring the hidden malaria transmission network, as it includes both the reported incubation time and the sensitivity analysis conducted previously.
Figure 5.13: The inferred malaria transmission network-based on the surveillance data in Yunnan province, China. There are in total 2928 infection cases reported in 51 townships during 2005. The width of the edge indicates the weight of transmission pathways. The color of each node represents its disease epidemic severity. Red edges represent a malaria endemic caused by the internal transmission component, the external influence component or their combination, while black edges represent a malaria endemic caused by the neighborhood transmission component.
5.2.3 Results

The inferred malaria transmission network is shown in Fig. 5.13. 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. The values of the edge weights are scaled up 20 times for the sake of clear presentation. As shown in Fig. 5.13, 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. For the sake of clear illustration and reading convenience, partial snapshots of Fig. 5.13, have been put separately in the following figures so that each figure contains townships with both their geographical information and their inferred results.

- **Small Communities:** Figures. 5.14 and 5.16 show that there are two communities in the malaria transmission network. The larger one (Fig. 5.14) 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.
Figure 5.14: 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 head to the border.
Figure 5.15: 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 Fig. 5.14.
Figure 5.15: 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 Fig. 5.14.

It can readily be noted from Fig. 5.14 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 Figs. 5.15 (a) - (f) in which reported cases are aggregated on an eight-day basis for clear presentation. As shown, although township 18 (Fig. 5.15 (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. 5.15 (b)). In comparison, the “mountain-valley-mountain” pattern of township 6 can only be partially matched with other townships (e.g., townships 4 (Fig. 5.15 (a)), 14 (Fig. 5.15 (d)) and 28 (Fig. 5.15 (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 and 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. 5.15 (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 (Fig. 5.16), 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 (Fig. 5.17 (a)) and 49 (Fig. 5.17 (b)) 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. For townships 2, 3, 24, 25, and 26, 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 (Fig. 5.18). Moreover, highway S231 crosses most of the infected areas, and directly connects them to the border. A similar situation can be seen in townships 13, 42, 45, 46, 47, 48, and 51 (Figs. 5.19 and 5.20). Figs. 5.21 (a) - (c) also validate that the reported cases of townships 1, 46 and 51 appear consecutively but are
Figure 5.16: 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. To interpret this result, more detailed information is needed.
Figure 5.17: 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) - (c) show the curves for townships selected from Fig. 5.16.
Figure 5.18: Townships 2, 3, 24, 25 and 26 are located adjacent to the border between China and Myanmar. Therefore, their self-connected edges are more likely to represent the localized malaria endemic caused by the imported cases.
Figure 5.19: Although township 13 is located adjacent to the border between China and Myanmar, it is in a valley and almost isolated from the traffic to the border. Moreover, there is a river crossing this township. Therefore, its self-connected edge is likely due to local transmissions.
Figure 5.20: Townships 42, 45, 46, 47, 48 and 51 are located adjacent to the border between China and Myanmar. Therefore, their self-connected edges are due to the imported cases.
Figure 5.21: 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) - (c) show the curves for townships selected from Fig. 5.20.
different from each other.

Townships 35, 37, 38, and 40, as shown in Fig. 5.22, are located in the same valley and along the same river (white line in Fig. 5.22). 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 these townships are affected by internal malaria transmission dynamics. In addition, these four townships are situated along two highways, S233 and S318, which connect to the border. It is likely that people in these townships cross the border frequently.

Townships 36 and 43 may also have malaria endemics caused by the adjacent rivers. Although they are situated in separate valleys, and are not a key traffic routes, they have relatively good habitats for the vector to develop. Thus, the internal transmission dynamic could be the dominant factor in the localized malaria endemics.

Townships 7, 9, 11, 15, 19, 22, 27, and 29 are situated relatively far from the border. However, the possibility of imported cases cannot be excluded, especially for 7, 11, 15, 19, and 29 because they are adjacent to highways. As for townships 9, 22, and 27, it is more likely that their transmission dynamics are caused by their surrounding environments. They are all near rivers.

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
Figure 5.22: Townships 35, 37, 38 and 40 are located in the same valley, and along the same river. The surrounding environment provides a suitable condition for the malaria vector to develop and reproduce. Besides, there are two highways (S233 and S318) cross these townships that also connect to the border. Therefore, the malaria endemics within these townships are likely to be caused by a combination of the internal transmission dynamics and the imported cases.
Figure 5.23: Townships 36 and 43 are located close to a river, which provides them with a suitable condition that helps the vector to develop and reproduce.
Figure 5.24: Townships 9, 22 and 27 may be influenced by the internal transmission dynamics because of their locations adjacent to rivers. The rest of the townships shown in this figure may be influenced by the imported cases, because of the convenient traffics connect them to the border.
imported cases. It seems that although some townships have similar temporal trends, they are not connected, for example, townships 18 (Fig. 5.15 (e)) and 50 (Fig. 5.17 (c)). 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 Fig. 5.14 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.

5.3 Summary

In this chapter, NetEpi has been tested both synthetic and real-world datasets to examine its performance. First, 24 synthetic disease transmission networks with different sizes and topologies were generated. Then synthetic data based on the generalized linear transmission model discussed in the previous chapter were produced. Specifically, for each network 10 independent datasets were generated. The precision-recall analyses showed that NetEpi performs better in all the datasets than a probability-based baseline method. Moreover, given
networks with the same size but distinct network topologies, NetEpi performs better in core-periphery networks than in hierarchical community networks or random graphs, because the out-degrees for core-periphery networks are smaller than those for the other two topologies. The average neighborhood combinations needed to achieve the optimal solution for each node are consequently reduced.

Experiments using synthetic networks were used to test the robustness of NetEpi over two important control parameters. First, NetEpi was run using different lengths of time windows. The results showed that the partial correlation networks usually achieve the best results when the time window equals to either 21 or 35 days. Longer time windows can introduce more false positive edges, whereas shorter time windows may miss many ground-truth edges. Second, the influence of the number of available surveillance data was examined. The experiments with datasets in which the number of observations were a quarter, half, 1 times, and 2 times the number of nodes suggested that this parameter had no significant influence, which further consolidates the robustness and accuracy of NetEpi.

Finally, NetEpi was applied to a real-world dataset that contained 2928 malaria infection cases from Yunnan province, China. The inferred malaria transmission network not only reflects the fact that many infection cases are imported cases; it is also consistent with the geographical road and river information. Specifically, the inferred network suggests the following interpretations: (i) the townships that are located adjacent to the border have more severe endemics; (ii) townships that are close to rivers where the malaria vector can quickly develop and reproduce offspring have higher rates of infection; and (iii) townships that are situated on the same rivers, or share the same highways have related infection rates.
Chapter 6

Conclusion and Future Work

6.1 Main Contributions

This thesis bridges the gap between theoretical studies of disease transmission networks and real-world 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. The main contributions of this thesis are summarized below.

1. We build a generalized linear disease transmission model that considers all the possible transmission pathways at a metapopulation level. The disease transmission dynamics caused by the internal environments, the interactions between subpopulations, and the
external influences from the imported cases have all been included.

2. We propose a method called NetEpi to inversely infer the hidden disease transmission networks. Using only the temporal and spatial information extracted from the available surveillance data, NetEpi is able to identify different types of transmission pathways and estimate their corresponding strengths.

3. Metapopulation-based disease transmission follows a Directed Cyclic Graphs pattern rather than a Directed Acyclic Graphs pattern, such as those in individual-based information diffusion and disease transmission. Even if a large proportion of a certain sub-population is infected, the remaining susceptible persons who have not been infected may become infected in the future. The proposed method can address this real-world requirement.

### 6.2 Future Work

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.

For example, we may use Poisson regression to embed environmental factors in malaria
transmission, and include this model in the internal transmission component. In addition, zero-inflated Poisson regression can also be used in handling sparse data. Another possible solution is to use white-box models, like compartmental models. For each subpopulation, we can use the same model framework but different datasets. However, this kind of method heavily depends on the accuracy of predefined parameters between different compartments. Moreover, because our dataset is very sparse, and most of the reported cases are imported cases, it may not be practical to use compartmental models to capture the internal transmission dynamics.

Another direction for our 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 and smoothly, and converges to either the local optimum or the global optimum. Therefore, future work should modify this technique to improve accuracy.
Bibliography


on Web Search and Data Mining, WSDM ’11, (New York, NY, USA), pp. 55–64, ACM, 2011.


Curriculum Vitae

Academic qualifications of the thesis author, Mr. YANG Xiaofei:

- Received the degree of Bachelor of Science (Hons.) (First Class) in Computing Studies from Hong Kong Baptist University, November 2011.

March 2014