Delft University of Technology
Master's Thesis in Electrical Engineering

Network Reconstruction for Epidemic Processes

Yue Tang
4620232

Network Architectures & Services Group

M.Sc. Thesis No: PVM 2018-098

TU Delft Delft University of Technology

Network Reconstruction for Epidemic Processes

Master's Thesis in Electrical Engineering

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands

Yue Tang
y.tang-5@student.tudelft.nl

M.Sc. Thesis No: PVM 2018-098

29th November 2018

Author
Yue Tang (y.tang-5@student.tudelft.nl)

Title
Network Reconstruction for Epidemic Processes

MSc presentation
29th November 2018

Graduation Committee	
Prof. dr. ir. Piet Van Mieghem	Delft University of Technology
Dr. Jaron Sanders	Delft University of Technology
Prof. Wioletta M. Ruszel	Delft University of Technology
Daily Supervisor	
Bastian Prasse	Delft University of Technology
Long Ma	Delft University of Technology

Abstract

Epidemic models are applied to describe epidemic processes such as the spreading of infectious viruses, opinions and fake news on real-life or online social networks, and to analyse the epidemic processes mathmatically. The viral state evolution is closely related to the underlying network topology. Therefore, the network topology is of vital importance to describing the viral state of each individual in a network.

This master thesis focuses on the network reconstruction problem of the NIMFA approximation of the Susceptible-Infected-Susceptible (SIS) epidemic process. Given the viral state series generated by the NIMFA epidemic process, we aim to estimate the adjacency matrix A of the underlying network given that the spreading parameters are known. The discrete-time NIMFA model, whose accuracy of modeling the SIS process is evaluated in Chapter 3, is applied in this thesis to describe the spreading process in the network, for it has the advantage of a lower computational complexity than the SIS model. The scope of the networks ranges from random network models to real networks (e.g., a subpart of the facebook network).

In this thesis, we estimate the adjacency matrix of the network from the viral states by a constrained linear least-squares formulation. Our algorithm gives an accurate estimate of the adjacency matrix provided that sufficiently many epidemic outbreaks are observed. By numerical evaluations of the network reconstruction method for random and real-world networks, we analyze the relationship among the accuracy, the number of outbreaks and the size of the network.

Keywords: SIS, NIMFA, network reconstruction, epidemic processes on networks

Preface

This master thesis is the final work during my studies for obtaining a Master of Science (MSc) degree in Electrical Engineering, Telecommunications and Sensing System at the EEMCS faculty of the Delft University of Technology. The thesis was finished at Network Architecture and Service (NAS) group, under the supervision of professor Piet Van Mieghem, Bastian Prasse and Long Ma. The successful completion of this work would not have been possible without the support and guidance from them.

Firstly, I would like to thank my supervisors. I would like to thank Professor Piet Van Mieghem for giving me this opportunity to carry out this work, and for his advices in shaping my research goal and giving feedback of the intermediate results obtained during this work. I would like to thank my daily supervisors Bastian Prasse and Long Ma. Their suggestions are of great help to me and their patient guidance helps me overcome many difficulties that I encountered during this period.

Besides, I would like to thank Jaron Sanders and Wioletta M. Ruszel for being part of my thesis committee. I also would like to thank all the other members in NAS group for offering me kindly help and encouragement.

Finally, I would like to thank my parents for always offering me support and understanding. I can have the courage to do anything I want, since I know they are always standing behind me. And I also would like to thank my friends for their accompany and encouragement.

Yue Tang
Delft, November 2018

Contents

List of Figures

List of Symbols

β	Infection rate
β_T	Infection probability, $\beta_T = T\beta$
δ	Curing rate
δ_T	Curing probability, $\delta_T = T\delta$
$\hat{A}$	Estimate of the adjacency matrix A
λ_1	The largest eigenvalue of the network's adjacency matrix A
$\mathcal{L}$	Link set in the graph G
$\mathcal{N}$	Node set in the graph G
τ	Effective Infection Rate, $\tau = \frac{\beta}{\delta}$
$\tau_c^{(1)}$	Epidemic threshold, $\tau_c^{(1)} = \frac{1}{\lambda_1}$
A	Adjacency matrix
a_{ij}	Element in the adjacency matrix A in the i-th row and the j-th column
d_i	Degree of node i
G	Graph, $G = (\mathcal{N}, \mathcal{L})$
K	Number of outbreaks
N	Number of nodes
n	Number of observations
p	Link probability, $p \in [0, 1]$ for Erdős-Renyi networks
T	Sampling time
V_i	The viral state matrix for the i-th epidemic outbreak

$v_i(t)$ The viral state of node i at time t, $t \in \mathbb{R}^+$ in the continuous-time NIMFA model

$v_i[k]$ The viral state of node i at time k, $k \in \mathbb{N}$ in the discrete-time NIMFA model

$x_i(t)$ The viral state of node i at time t, $t \in \mathbb{R}^+$ in the continuous-time SIS model

$x_i[k]$ The viral state of node i at time k, $k \in \mathbb{N}$ in the sampled-time SIS model

Chapter 1

Introduction

1.1 Background

The study of epidemics on networks involves a myriad of phenomena, such as the spreading of infectious disease, opinions and fake news on real-life or online social networks. It is of vital importance to understand what is going on in the epidemic process, so that effective measures can be taken to prevent disasters such as the death due to serious infectious diseases and the paralysis of the communication network casued by computer viruses. However, it is impractical to understand the epidemic processes by tracing the exact dynamics of the epidemic process in large complex networks. Epidemic models are applied to understand the viral dynamics, and, thus, further enable people to control the outbreak of the epidemic.

Before the concept of network science is raised, traditional epidemic models hardly considered the underlying network such as the Yule process and the linear birth and death process [1]. These traditional epidemic models cannot describe complex epidemic processes accuratly. Nowadays, compartmental models are the most commonly used epidemic models to describe complex epidemic processes. They divide the population into compartments base on the state of each individual. Individuals in the same compartment has the same viral state [4]. Two of the most fundamental examples of compartment models are the Susceptible-Infected-Susceptible (SIS) model and the Susceptible-Infected-Removed (SIR) model.

The topological structure of the network is crucial for the research on epidemic processes. The knowledge of the underlying topology enables the analysis of the epidemic processes based on the epidemic models and the design of control strategies. Hence, one important question of studying the epidemic process is to know the topology of the network, which underlies the epidemic process. The network reconstruction problem is the task of estimating the network by observing the viral state series of an epidemic process. Network reconstruction is applied in many areas, such as deducing gene regulatory networks from expression data in biological networks [5]. This work solves the network reconstruction problem that given the viral state series generated by discrete-time NIMFA epidemic model, the adjacency matrix A needs to be estimated.

1.2 Thesis Layout

This thesis consists of five chapters:

Chapter 2 provides background knowledge of graph theory, including basic metrics and parameters applied to analyze network topologies, as well as typical network models. Chapter 3 introduces the Susceptible-Infected-Susceptible (SIS) model in continuous-time and sampled-time, and the N-Intertwined Mean Field Approximation (NIMFA) in continuous- and discrete-time. Chapter 4 introduces a reconstruction algorithm to estimate the adjacency matrix based on the discrete-time NIMFA model. The network reconstruction results are presented, as well as the numerical evauluation of the reconstruction accuracy. Finally, in Chapter 5, conclusions as well as an outlook to future work is given.

Chapter 2

Network Science

Networks such as telecommunication networks, computer networks, biological networks, social networks, economic and financial networks are commonly used in people's daily life. In general, a network represents a group of objects and their interconnections. Network science is dedicated to studying the general properties of such complex networks, such as the degree distribution and the shortest path for a better understanding of the network. This chapter introduces the fundamentals of graph theory, which is the root of network science. Network metrics and some commonly used network models are introduced.

2.1 Graph Theory

Graph theory is used to analyze the performance of the network. The idea of graph theory was first proposed by Leonhard Euler in 1736 [6] to solve the Seven Bridges of Königsberg problem. This problem asks whether it is possible to find a route which goes through each of the seven bridges of Königsberg only once. The modeling of the land and the bridges by means of nodes and links respectively was the foundation of graph theory. Based on Euler's idea, Johann Benedict Listing first introduced the "topology" to describe the structure of the network [6].

In graph theory, a graph $G(\mathcal{N}, \mathcal{L})$ consists of a set $\mathcal{N}$ of N nodes connected by a set $\mathcal{L}$ of L links. When a graph does not contain a self-loop or an overlapped link, it is referred to as the simple network. For instance, a particular graph is given by the *complete graph*: in a complete graph K_N with N nodes, any two nodes is connetced with each other. Figure 2.1 illustrates a complete graph K_8 with $N = 8$ nodes. The complete graph K_N has a total number of $L = \frac{N(N-1)}{2}$ links, which is the maximum number of links in the simple graph with N nodes.

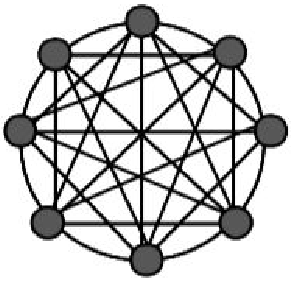

Fig. 2.1. The complete graph K_8 [1]

In this work, we confine ourselves to unweighted and undirected networks without self-loops. The topology of a graph with N nodes and L links is described by the $N \times N$ adjacency matrix A. For an unweighted and undirected network, the elements a_{ij} of the adjacency matrix A are binary numbers, which describe the presence or the absence of the link between two nodes:

$$a_{ij} = \begin{cases} 1, & \text{if there is a link between node } i \text{ and node } j \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

where $i = 1, ..., N$ and $j = 1, ..., N$. Since the graph is undirected, it holds that $a_{ij} = a_{ji}$. The diagonal entries a_{ii} are 0 since there are no self-loops. Therefore, the adjacency matrix A is symmetric.

The degree d_i of node i is defined as the number of neighbors of node i:

$$d_i = \sum_{j=1}^{N} a_{ij} \tag{2.2}$$

The degree of a node indicates the associated strength of this node with the other nodes in the network. For instance, on instagram or twitter, a high degree node means that the user has many connections to the other users.

The largest eigenvalue of the adjacency matrix A is denoted by λ_1 and its corresponding eigenvector is denoted by x_1, and it holds:

$$Ax_1 = \lambda_1 x_1 \tag{2.3}$$

2.2 Network Models

Instead of analyzing a specific network by a graph, network models are applied to analyze a group of networks with similar characteristics. This section introduces three important and well-studied network models: the Erdős-Renyi (ER) network, the small-world network and the scale-free network. Typical outcomes of these three types of networks are illustrated in Figure 2.2.

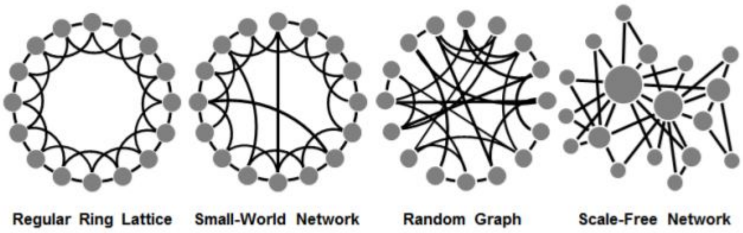

Fig. 2.2. Typical outcomes of the Erdős-Renyi, the small-world, and the scale-free random graph models [2]

.

Erdős-Renyi network

The Erdős-Renyi (ER) model was introduced by Paul Erdős and Alfred Renyi in 1959 [7]. In the ER model, a graph with N nodes is denoted as $G_p(N)$. Nodes are connected independently with the link probability $p \in [0, 1]$. If the link probability equals $p = 0$, then there is no link in the network. As the link probability p increases from 0 to 1, it is more likely to have a link between two nodes. The ER model generates a complete graph when the probability p equals to 1.

Small-world network

The small-world property found by Watts and Strogatz in 1998 [8] states that on average, arbitrary two nodes in the small-world network are connected by short paths. This property can be interpreted that any two people in the world can get to know each other through an average of 6 intermediate friends in the social network [9]. As it is illustrated in [9], a small-world network is generated from the regular graph, where each node in the graph has the same degree, by randomly rewiring the links with a probability $p \in [0, 1]$. When $p = 1$, the network tends to behave similarly to the Erdős-Renyi network.

Scale-free network

The scale-free property found by Barabási and Albert in 1999 [10]. Many real-world networks such as the social network show a scale-free property that has a power law degree distribution:

$$p(k) \sim k^{-\gamma} \quad (2 < \gamma < 3), \tag{2.4}$$

where $p(k)$ is the fraction of nodes having degree k in the network and γ is a exponential parameter whose value satisfies $2 < \gamma < 3$ in most real-world networks [11]. The Barabási-Albert (BA) model generates the scale-free networks following the rule of preferential attachment [12], which stating that new nodes are added to the network iteratively by connecting to the initially existing nodes, where the node with larger degree has a higher probability of being connected.

Chapter 3

Epidemic Models

Epidemic models describe the dynamical evolution of the contagion process within a group of individuals. Most epidemic models are compartmental models, which assume that every individual of the population can be assigned to different classes depending on the stage of the disease [13].

Two fundamental compartmental models are the Susceptible-Infected-Susceptible (SIS) model and the Susceptible-Infected-Removed (SIR) model. In both models, a susceptible node can be infected by its infected neighbors. In the SIS model, an infected node can be cured and become a susceptible node, which can be infected by its neighbors again. On the other hand, in the SIR model, a cured node has immunity (and is hence removed) and will not be infected again. In this work, the SIS model is considered.

For the task of the network reconstruction, it is most convenient to describe the SIS process in discrete-time. However, it is proved in [14] that the maximu-likelihood network reconstruction from the complete sampled-time SIS nodal state infection information is NP-hard. Thus, we consider the discrete-time N-Intertwined Mean-Field Approximation (NIMFA) of the SIS process, which has the advantage of reducing the computational complexity by calculating just N linear equations instead of calculating 2^N linear equations in the SIS process with N nodes.

This chapter introduces the Susceptible-Infected-Susceptible (SIS) epidemic model. Section 3.1 and Section 3.2 introduce the system model of the continuous-time SIS process, the sampled-time SIS process and the NIMFA model respectively. Section 3.3 gives a numerical evaluation on the accuracy of the two approximate models, the sampled-time SIS model and the discrete-time NIMFA model, with respect to the original continuous-time SIS process.

3.1 SIS Epidemic Model

Arguably one of the most fundamental epidemic models is the Susceptible-Infected-Susceptible (SIS) model [13], whereby each node is either in a susceptible or an infected state. It can be described in continuous-time and in discrete-time, which we will introduce in Section 3.1.1 and Section 3.1.2, respectively.

3.1.1 Continuous-time SIS Process

The viral state of a node i at continuous time t is denoted by a Bernoulli random variable $x_i(t) \in \{0, 1\}$. The values $x_i(t) = 0$ and $x_i(t) = 1$ represent that node i is in the susceptible state and in the infected state at time $t \geq 0$, respectively. The SIS process assumes that the curing process of each node i and the infection process of each link are both Poisson process with the *recovery rate* δ and the *infection rate* β [1]. As shown in Figure 3.1, a susceptible node i can be infected by its neighbors with a total infection rate $\sum_{j=1}^{N} a_{ij} x_j \beta$, and an infected node i can be cured with a recovery rate δ.

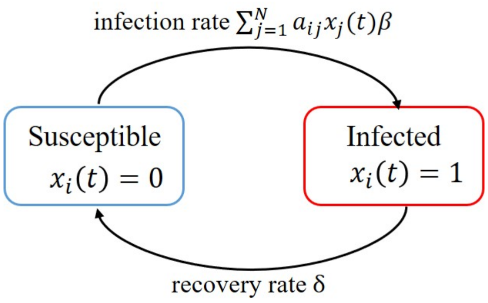

Fig. 3.1. The state transition of an SIS epidemic process

The *effective infection rate* is defined as $\tau = \frac{\beta}{\delta}$, which is crucial to the epidemic behavior, and a phase transition of the epidemic process occurs. "The epidemic threshold τ_c is defined as the border between an exponential die-out phase and a non-zero fraction of infected nodes in the metastable state" [1]. For any finite sized network and for no self-infections ($\epsilon = 0$), the lower bound of the SIS epidemic threshold τ_c is given by [15]

$$\tau_c \geq \tau_c^{(1)} = \frac{1}{\lambda_1}, \tag{3.1}$$

where λ_1 is the largest eigenvalue of the adjacency matrix A. If the effective infection rate satisfies $\tau \geq \tau_c^{(1)}$, then the virus will spread over the network and prevail for a very long time and an epidemic outbreak can be observed. On the contrary, if $\tau < \tau_c^{(1)}$, then the number of infected nodes decreases exponentially fast [13, 16]. The prevalence in the SIS epidemic process after an infinitely long time tends to zero, where the absorbing state $x_i(t) = 0$ for all nodes i is reached [17].

Figure 3.2 shows the expected fraction of the infected nodes in an examplary SIS process with a high effective infection rate τ. It starts with a small number fraction of nodes being infected. The number of infected nodes increases exponentially fast during the outbreak phase. Then it reaches to the meta-stable state and keep for a long time. Then after an exponentially long time with respect to N, the epidemic dies out, where the absorbing state is reached.

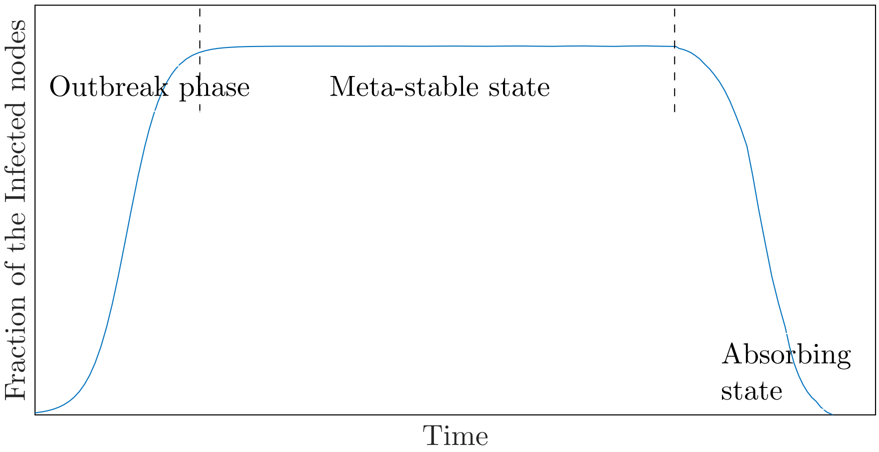

Fig. 3.2. The expected fraction of the infected nodes in the SIS process.

3.1.2 Sampled-time SIS Process

"The transition probability of the continuous-time Markovian SIS process from state i at time τ to state j at time $t + \tau$ is denoted by $P_{ij}(t)$, which is independent of τ since the SIS process is stationary" [18]. The sampled-time SIS process is a sampled-time Markov chain with sampling time T, which is a discrete-time Markov chain [1]. "The transition probabilities P_{ij} from state i to state j of the sampled-time Markov chain are given by the first-order Taylor-expansion $P_{ij} = P'_{ij}(0)T$" [18]. Hence, in the sampled-time SIS process, the infection probability per link is $\beta_T = \beta T$ and the curing probability is $\delta_T = \delta T$.

There are three transitions in the sampled-time SIS process [18], which are listed below:

1. A node i changes from the infected state at time k to the susceptible

state at time $k+1$, which occurs with probability:

$$\Pr\left[x_i[k+1]=0|x_i[k]=1\right]=\delta T \tag{3.2}$$

2. A node i changes from the susceptible state at time k to the infected state at time $k+1$, which occurs with probability:

$$\Pr\left[x_i[k+1]=1|x_i[k]=0\right]=\beta\sum_{j=1}^{N}a_{ij}x_j[k]T \tag{3.3}$$

   where N is the number of nodes in the network.

3. No node changes its state from time k to time $k+1$, which occurs with probability:

$$\Pr\left[x[k+1]=x[k]\right]=1-\delta T u^T x[k]-\sum_{j=1}^{N}\beta T x_j[k]\sum_{i=1}^{N}\left(1-x_i[k]\right)a_{ij}. \tag{3.4}$$

3.2 The N-Intertwined Mean-Field Approximation (NIMFA)

The exact SIS process is given by a continuous-time Markov chain with a state space with 2^N elements for a network with N nodes. The state transition probabilities in the continuous-time Markov chain are described by a set of linear equations. Since the number of possible states of the system grows exponentially with the number of nodes N, this exact description of the SIS process is not practical. Therefore, several methods have been developed to approximate the SIS model in order to reduce the computational complexity to make an analysis feasible. The N-Intertwined Mean-Field Approximation (NIMFA) [15] and the Heterogeneous Mean-Field method (HMF) [19] are two widely-used approximation methods.

Pastor-Satorras and Vespignani [19] introduced the Heterogeneous Mean-Field method, in which the SIS process is approximated based on the degree distribution of the underlying graph. It assumes that the infection probabilities of nodes with the same degree are the same [20]. The HMF approximation considers the probability that a node with degree k is infected at time t during the epidemic process. ”However, the state of each node is not taken into account” [21].

The N-intertwined epidemic approximation is derived by observing the viral state of *each* node with the only assumption that the state of neighboring nodes is stochastically independent [21]. It replaces the binary viral state $x_i(t)$ in the SIS process by its expectation $v_i(t) = E[x_i(t)]$. It was shown in [21] that NIMFA is currently the most accurate approximation of the SIS model in any network.

The NIMFA approximation is derived as follows, which is shown in Figure 3.3. The governing equation of the SIS process for node i is

$$\frac{dx_i(t)}{dt} = -\delta x_i(t) + (1 - x_i(t))\beta \sum_{j=1}^{N} a_{ji} x_j(t), \tag{3.5}$$

for nodes $i = 1, ..., N$. By taking the expected value of equation (3.5), we obtain

$$\frac{dE[x_i(t)]}{dt} = -\delta E[x_i(t)] + \beta \sum_{j=1}^{N} a_{ji} E[x_i(t)] - \beta \sum_{j=1}^{N} a_{ji} E[x_i(t) x_j(t)] \tag{3.6}$$

In NIMFA, by assuming independence, which implies

$$E[x_i(t) x_j(t)] = E[x_i(t)] E[x_j(t)] \tag{3.7}$$

and by denoting $v_i(t) = E[x_i(t)] = \Pr[x_i(t) = 1]$, equation (3.6) becomes

$$\frac{dv_i(t)}{dt} = -\delta v_i(t) + \beta\,(1 - v_i(t)) \sum_{j=1}^{N} a_{ij} v_j(t), \quad i = 1, 2, ..., N \tag{3.8}$$

The discrete-time NIMFA model is derived from the continuous-time NIMFA (3.8). Applying Euler's method [22] gives:

$$\frac{dv(t)}{dt} \simeq \frac{v(t+T) - v(t)}{T}, \quad \text{for } T \simeq 0 \tag{3.9}$$

Inserting equation (3.9) in equation (3.8), yields the discrete-time NIMFA model [23]:

$$v_i[k+1] = -\delta T v_i[k] + v_i[k] + (1 - v_i[k])\beta T \sum_{j=1}^{n} a_{ij} v_j[k], \tag{3.10}$$

where $k \in \mathbb{N}$ denotes the discrete-time step, and T is the sampling time.

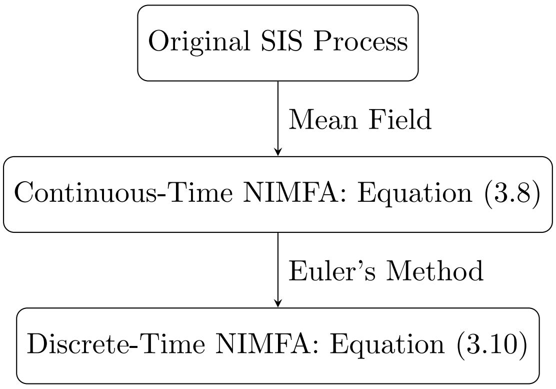
Original SIS Process
Mean Field
Continuous-Time NIMFA: Equation (3.8)
Euler's Method
Discrete-Time NIMFA: Equation (3.10)

Fig. 3.3. The flow chart of the derivation of the discrete-time NIMFA

If the sampling time T satisfies $T \leq \frac{1}{\delta}$ and $T \leq \frac{1}{\beta(N-1)}$, then the viral state $v_i[k]$ of the discrete-time NIMFA (3.10) stays in $[0, 1]$ for every time $k \geq 0$ as shown by [23].

3.3 Numerical Evaluation of the SIS Epidemic Models

The sampled-time SIS model and the NIMFA model are both approximations of the continuous-time SIS model. In the following, we discuss how these approximate models fit to the continuous-time SIS model by means of simulation.

We generate $M = 50$ random Erdős-Renyi networks with $N = 100$ nodes and with link probability $p = 0.5$, where disconnected networks are discarded. According to [14], the sampling time T for the sampled-time SIS process needs to satisfy

$$T \leq \frac{4}{N^2\beta + 4N\delta} \tag{3.11}$$

to ensure that the expression of the transition probabilities are in the interval [0, 1]. In the simulations, we choose the value of the sampling time T such that the bound (3.11) is satisfied with equality. We choose all nodes in the network to be infected initially $x_i[0] = 1$, for all nodes i, and set the recovery rate to $\delta = 1$. The results for the sampled-time SIS shown below are the average of $P = 1000$ realizations of the SIS process.

The average over $M = 50$ networks and $P = 1000$ realizations for the continuous-time SIS model at time k is given by:

$$\bar{x}_{\text{exact}}(kT) = \frac{1}{MP}\sum_{i=1}^{M}\sum_{j=1}^{P} x^{(ij)}(kT), \tag{3.12}$$

where $x^{(ij)}(kT)$ denotes the viral state of the continuous-time SIS process for a single realization j for the i-th network at time k.

The average over $M = 50$ networks and $P = 1000$ realizations for the sampled-time SIS model at time k is:

$$\bar{x}_{\text{sampled}}[k] = \frac{1}{MP} \sum_{i=1}^{M} \sum_{j=1}^{P} x_{\text{sampled}}^{(ij)}[k] \tag{3.13}$$

where $x_{\text{sampled}}^{(ij)}[k]$ denotes the viral state of the sampled-time SIS process for a single realization j for the i-th network at time k.

For the discrete-time NIMFA model, $\bar{v}[k]$ is the average over $M = 50$ networks at time k:

$$\bar{v}[k] = \frac{1}{M} \sum_{i=1}^{M} v^{(i)}[k], \tag{3.14}$$

where $v^{(i)}[k]$ denotes the viral state of the discrete-time NIMFA model for the i-th network at time k.

We compare the error of the approximated models to the original continuous-time SIS model with respect to the mean squared error (MSE) ε, which is given by:

$$\varepsilon_1 = \frac{1}{K} \sum_{k=1}^{K} (\bar{x}(kT) - \bar{x}_{\text{sampled}}[k])^2 \tag{3.15}$$

and

$$\varepsilon_2 = \frac{1}{K} \sum_{k=1}^{K} (\bar{x}(kT) - \bar{v}[k])^2 \tag{3.16}$$

for the sampled-time SIS model and the discrete-time NIMFA model respectively, where K is the total number of observations.

The impact of the effective infection rate τ and the sampling time T on the fit of the approximate models to the continuous-time SIS model is investigated in the following.

Effective Infection Rate τ

The epidemic process can be represented by the prevalence:

$$P[k] = \frac{1}{N} \sum_{i=1}^{N} x_i[k] \tag{3.17}$$

Figure 3.4 and Figure 3.5 show the prvalence of the three epidemic models above the epidemic threshold ($\tau = 5\tau_c^{(1)}$) and below the epidemic threshold ($\tau = 0.9\tau_c^{(1)}$). Each of the curves for the sampled-time SIS model and the discrete-time NIMFA model has an error e to the continuous-time SIS curve. We define these errors as $e[k] = \frac{1}{N}\sum_{i=1}^{N} x_i(kT) - \frac{1}{N}\sum_{i=1}^{N} v_i[k]$, and we obtain Figure 3.6 and Figure 3.7 for $\tau = 5\tau_c^{(1)}$ and $\tau = 0.9\tau_c^{(1)}$.

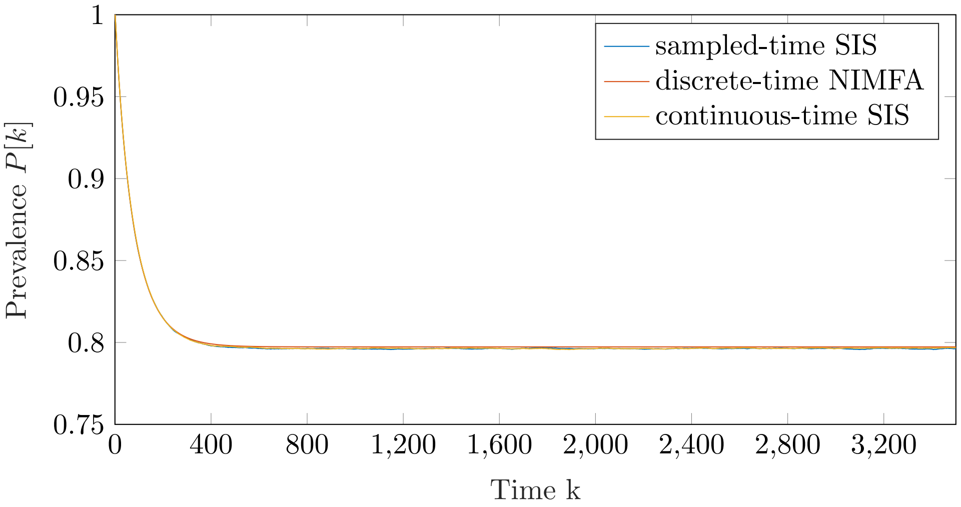

Fig. 3.4. The prevalence $P[k]$ for the sampled-time SIS model, the discrete-time NIMFA and the continuous-time SIS for $\tau = 5\tau_c^{(1)}$.

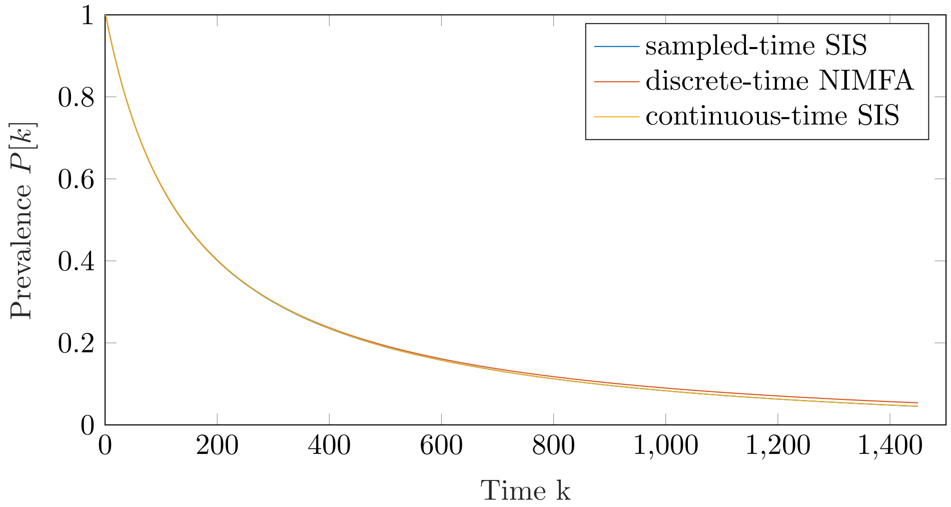

Fig. 3.5. The prevalence $P[k]$ for the sampled-time SIS model, the discrete-time NIMFA and the continuous-time SIS for $\tau = 0.9\tau_c^{(1)}$.

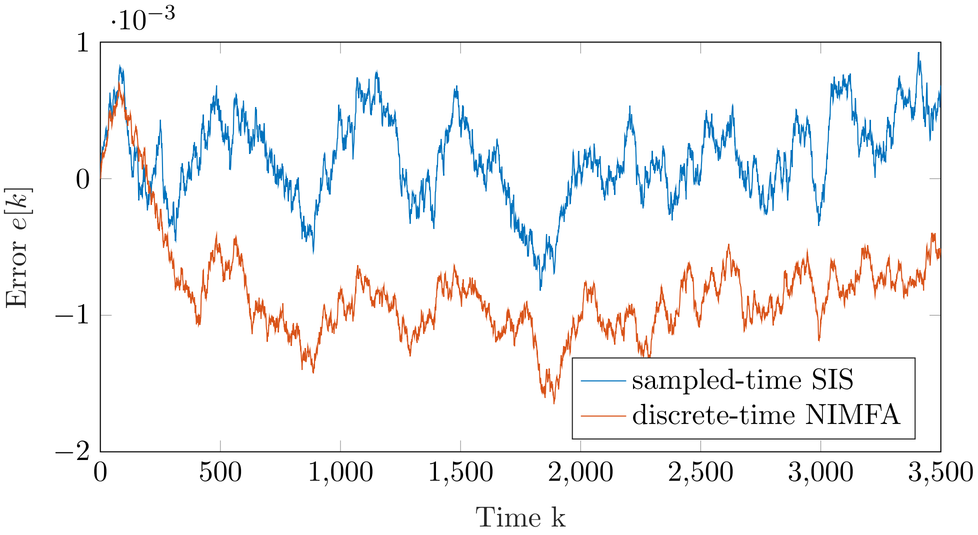

Fig. 3.6. Error $e[k]$ of the prevalence of the sampled-time SIS model and the discrete-time NIMFA compared to the continuous-time SIS model for $\tau = 5\tau_c^{(1)}$.

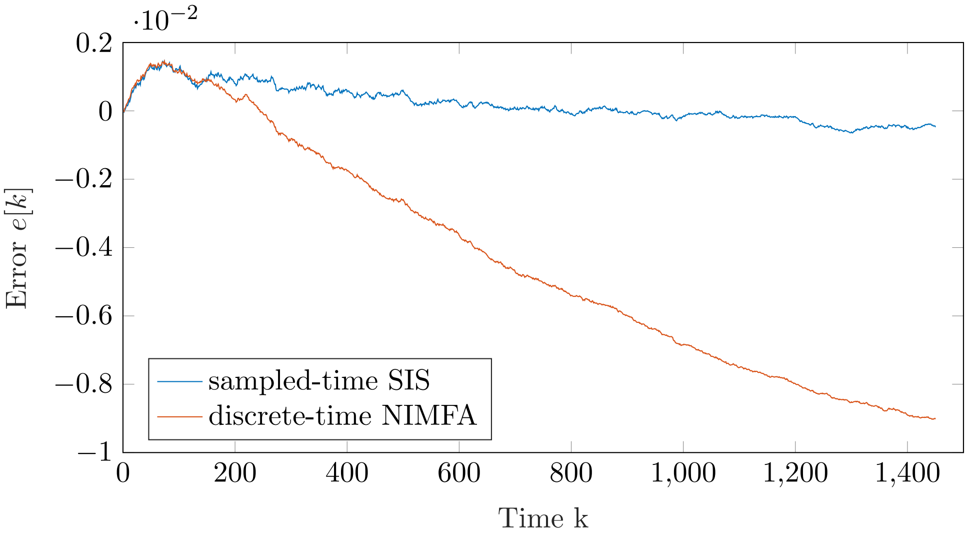

Fig. 3.7. Error $e[k]$ of the prevalence of the sampled-time SIS model and the discrete-time NIMFA compared to the continuous-time SIS model for $\tau = 0.9\tau_c^{(1)}$.

The MSEs ε of the two approximated models, the sampled-time SIS model and the discrete-time NIMFA model, are $\varepsilon_1 = 8.56 \times 10^{-11}$ and $\varepsilon_2 = 1.30 \times$

10^{-10}, respectively, for $\tau = 5\tau_c^{(1)}$. This result shows that the sampled-time SIS model outperforms the discrete-time NIMFA model with respect to approximating the continuous-time SIS process. For the case of $\tau = 0.9\tau_c^{(1)}$, the MSE is $\varepsilon_1 = 5.07 \times 10^{-11}$ and $\varepsilon_2 = 4.70 \times 10^{-8}$, respectively. Also for this situation, the sampled-time SIS model outperforms the discrete-time NIMFA model with respect to approximating the continuous-time SIS process. Comparing the two situations for $\tau = 5\tau_c^{(1)}$ and $\tau = 0.9\tau_c^{(1)}$, the former one has a smaller error for the NIMFA model.

To discover the relationship between the MSE ε and the effective infection rate τ, we change the value of τ in the interval $[0.9\tau_c^{(1)}, 5\tau_c^{(1)}]$. In this situation, the sampling time is set to $T = \frac{4}{N^2\beta_{max}+4N\delta}$, where $\beta_{max} = 5\tau_c^{(1)}$. Figure 3.8 shows that the MSE ε of the sampled-time SIS model does not have an obvious change with the effective infection rate τ. For the MSE ε of the discrete-time NIMFA model, when the effective infection rate τ is around the epidemic threshold $\tau_c^{(1)}$, it has the worst fitness to the continuous-time SIS process. Otherwise, when τ is greater than $\tau_c^{(1)}$, the MSE ε of the discrete-time NIMFA model shows a downwards trend as τ increases. To visualize the influence of the effective infection rate τ on the fit of the sampled-time SIS process, we plot the samped-time SIS curve separately in Figure 3.9. There seems to be no obvious relationship on the sampled-time SIS curve with the change of the effective infection rate τ.

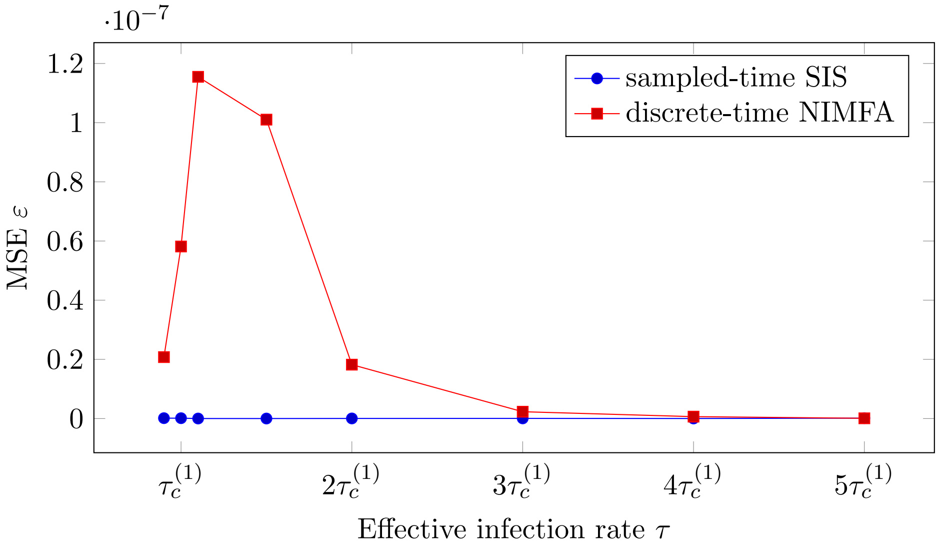

Fig. 3.8. Fitness ε of the sampled-time SIS model and the discrete-time NIMFA model with respect to the continuous-time SIS model, in dependency of the effective infection rate τ.

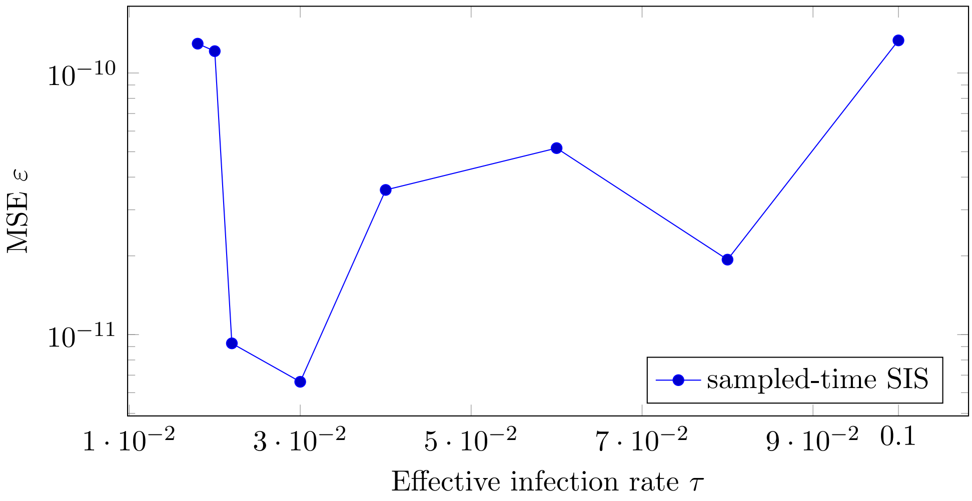

Fig. 3.9. Fitness ε of the sampled-time SIS model with respect to the continuous-time SIS model, in dependency of the effective infection rate τ on a semi-logarithmic scale.

Sampling Time T

In the following, the influence of the sampling time T on the fitness ε is investigated. We vary the sampling time T in the interval of $[\frac{1}{100}T_{\max}, T_{\max}]$ and obtain the relationship between the fitness ε and sampling time T in Figure 3.10. Figure 3.10 shows that the fitness ε becomes better when the sampling time T decreases. We fit the above fitness ε to a cubic polynomial curve and the result is shown in Figure 3.11. The function of the fitness ε and the sampling time T can be approximated as

$$\varepsilon(T) = 0.0088T^3 - 2.08 \times 10^{-5}T^2 + 1.743 \times 10^{-8}T \tag{3.18}$$

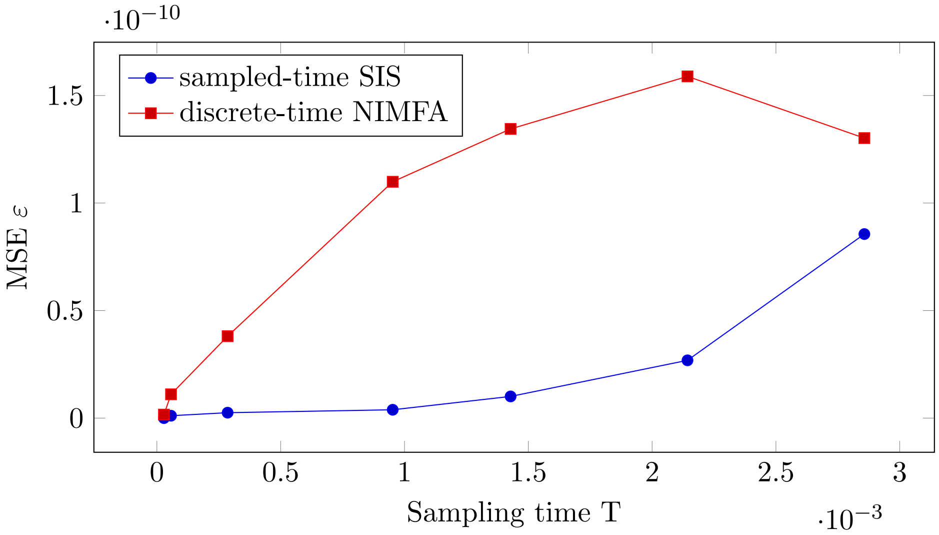

Fig. 3.10. Fitness ε of the sampled-time SIS model, the discrete-time and the continuous-time NIMFA model with respect to the continuous-time SIS model, in dependency of the sampling time T.

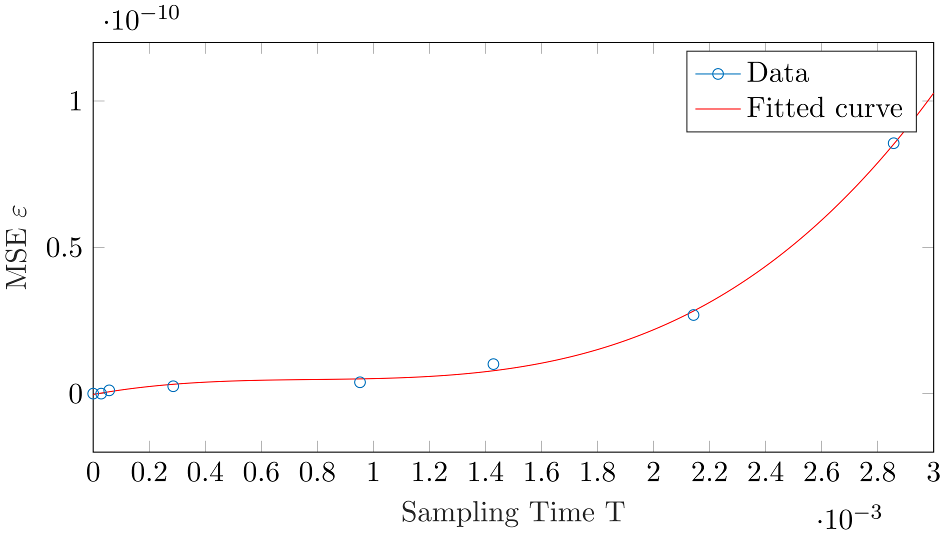

Fig. 3.11. The fitted cubic polynomial curve of the fitness ε and the sampling time T.

Chapter 4

Network Reconstruction for NIMFA

The objective of this thesis is to reconstruct the network topology given the knowledge of the viral states $v_i[k]$ of all nodes i in a discrete-time NIMFA model. Section 4.1 formulates the network reconstruction task as linear system and introduces the basic idea of the network reconstruction method introduced in [24]. Section 4.2 discusses the identifiability of the network. Section 4.3 gives the network reconstruction algorithm in detail. Finally, the numerical evaluations of the network reconstruction algorithm on the accuracy for both random and real-world networks are presented in Section 4.4.

4.1 Formulation as Linear System

As discussed in Chapter 3, compared to the sampled-time SIS model, the NIMFA model has a lower computational complexity, which makes an analysis easier. Therefore, the discrete-time NIMFA model is considered in the following, whose equations were introduced in Chapter 3. According to the derivations in [24], we define

$$b_i[k] = \frac{v_i[k+1] - v_i[k]}{1 - v_i[k]} \tag{4.1}$$

$$c_i[k] = \frac{v_i[k]}{v_i[k] - 1} \tag{4.2}$$

for every node $i = 1, 2, ..., N$. Then, the NIMFA equation (3.10) becomes

$$b_i[k] = \delta_T c_i[k] + \beta_T v^T[k] A^T_{\text{row},i}, \tag{4.3}$$

where $A_{\text{row},i}$ is the i-th row of the adjacency matrix A. As stated in [24]: For a network with N nodes, the equations (4.3) for node $i = 1, ..., N$ can be concatenated

$$\beta_T A v[k] = b[k] - \delta_T c[k], \tag{4.4}$$

with the $N \times 1$ vectors $v[k] = (v_1[k], ..., v_N[k])^T$, $b[k] = (b_1[k], ..., b_N[k])^T$ and $c[k] = (c_1[k], ..., c_N[k])^T$. Furthermore, for all the observation time instants $k = 1, ..., n$, the equations (4.4) can be combined as:

$$\beta_T A V = B - \delta_T C, \tag{4.5}$$

with the $N \times n$ matrices $V = (v[1]...v[n])$, $B = (b[1]...b[n])$ and $C = (c[1]...c[n])$. Finally, the network reconstruction problem becomes solving the above set of linear equations (4.5) for the adjacency metrix A.

If the $N \times n$ viral state matrix V has full row rank, $\text{rank}(V) = N$, then there is a unique solution to the normal equation (4.5) for the adjacency matrix $\hat{A}$:

$$\hat{A} = \frac{1}{\beta_T}(C - \delta_T B)V^T(VV^T)^{-1}, \tag{4.6}$$

which follows from the normal equations [25]. For convenience, we define the $N \times n$ matrix M:

$$M = \frac{1}{\beta_T}(C - \delta_T B), \tag{4.7}$$

which yields

$$\hat{A} = MV^{\dagger}, \tag{4.8}$$

where the Moore-Penrose pseudo-inverse [25] of the viral state matrics V is given by $V^{\dagger} = V^T(VV^T)^{-1}$.

4.2 Exact Network Reconstruction

4.2.1 Singular Value Decomposition (SVD)

Theorem 1 ***(Singular Value Decomposition(SVD)[26])***
If X is a real $m \times n$ matrix, then there exist orthogonal matrices $U = [u_1, ..., u_m] \in \mathbb{R}^{m \times m}$ and $Q = [q_1, ..., q_n] \in \mathbb{R}^{n \times n}$, such that

$$U^T X Q = \Sigma \in \mathbb{R}^{m \times n}, \tag{4.9}$$

The diagonal entries of the matrix Σ are the singular values of X and the vectors u_i and q_i are the i-th left singular vectors and the i-th right singular vectors, respectively.

The sigular value decomposition of the viral state matrix V is given by

$$V = U\Sigma Q^T \tag{4.10}$$

The numerical rank of the viral state matrix $\text{rank}(V, \epsilon)$ is defined as the number of singular values of V that are greater than the threshold ϵ. We set the threshold ϵ to the default value of the Matlab command `rank`. If $\text{rank}(V, \epsilon) < N$, then the viral state matrix V is regarded as numerical rank deficient.

4.2.2 Numerical Evaluation on the Numerical Rank

To solve the unconstrained linear system (4.5) uniquely for the adjacency matrix A, the viral state matrix V must be of full row rank, i.e.,

$$\text{rank}(V) = N. \tag{4.11}$$

However, by simulating multiple NIMFA processes, it is found that $\text{rank}(V, \epsilon)$ $< N$ almost always occurs. In fact, in most cases, the $\text{rank}(V, \epsilon)$ is much smaller than the number of nodes N. To discuss possible factors that may have an influence on the numerical rank of the viral state matrix V, the following numerical evaluations are performed.

Distribution of the Numerical Rank

We generate 100 random Erdős-Renyi networks with $N = 100$ nodes and with the link probability $p = 0.5$. For each network, a discrete-time NIMFA process is simulated with an effective infection rate $\frac{\beta}{\delta} = 0.8$ and $\delta = 1$, which satisfies $\frac{\beta}{\delta} > \frac{1}{\lambda_1(A)}$, where $\lambda_1(A)$ is the largest eigenvalue of the network's adjacency matrix A. The initial states $v_i[0]$ of the NIMFA process are uniformly and independently distributed random numbers in the interval $[0, 1]$. A NIMFA series $v[1], v[2], ..., v[n]$ is obtained, where n is the observation length. The numerical rank $\text{rank}(V, \epsilon)$ is computed for each network and a distribution of the rank is obtained. According to the simulations, with a frequency of 0.8, we have $\text{rank}(V, \epsilon) = 15$. Otherwise, $\text{rank}(V, \epsilon) = 16$ holds with a frequency of 0.2. The largest numerical rank is $\text{rank}(V, \epsilon) = 16$, which is much smaller than the number of nodes $N = 100$.

Link probability p

To figure out if the link probability p will influence the distribution of $\text{rank}(V, \epsilon)$, in this simulation, the link probability is set to different values $p = 0.1, ..., 0.9$. For each value of p, we generate 100 random Erdős-Renyi networks with $N = 100$ nodes. For each network, a discrete-time NIMFA process with an effective infection rate $\frac{\beta}{\delta} = 0.8$ and $\delta = 1$ is simulated. The numerical rank $\text{rank}(V, \epsilon)$ is computed for each network and the resulting distribution of $\text{rank}(V, \epsilon)$ with respect to the link probability p is shown in Figure 4.1.

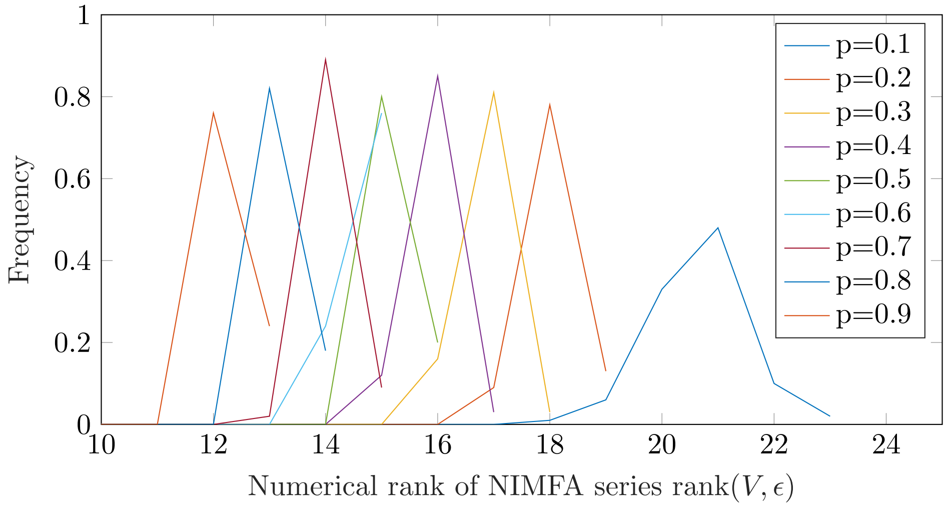

Fig. 4.1. The distribution of the numerical rank rank(V,ϵ) for different values of the link probability p when the number of nodes in the network is $N = 100$.

By calculating the average over the 100 networks for each value of the link probability, we obtain a curve that shows the relationship of the link probability p and the numerical rank rank(V,ϵ). Figure 4.2 shows that rank(V,ϵ) almost has a linear relationship with the link probability p. As the link probability p increases, the numerical rank rank(V,ϵ) decreases. Hence, the sparser the network is, the more likely the network can be identified. Despite this, as long as rank$(V) < N$, the network cannot be identified from solving the unconstrained linear system (4.5). The largest numerical rank is rank$(V,\epsilon) = 23$, which is far from the number of nodes $N = 100$.

Number of nodes N

We are interested in the relationship of the number of nodes N and the distribution of the numerical rank rank(V,ϵ). The number of nodes in the network is set to $N = 25, ..., 100$. For all the values of N, the link probability is set to $p = 0.5$. For each network, a discrete-time NIMFA process with an effective infection rate $\frac{\beta}{\delta} = 0.8$ and $\delta = 1$ is simulated, which satisfies $\frac{\beta}{\delta} > \frac{1}{\lambda_1(A)}$. The normalized rank$(V,\epsilon)/N$ is computed for each situation and a distribution of rank$(V,\epsilon)/N$ with respect to the link probability p is shown in Figure 4.3.

By calculating the average over the 100 networks for each value of N, we obtain a curve that shows the relationship of the network size N and rank$(V,\epsilon)/N$ in Figure 4.4. It can be seen that rank$(V,\epsilon)/N$ has a nonlin-

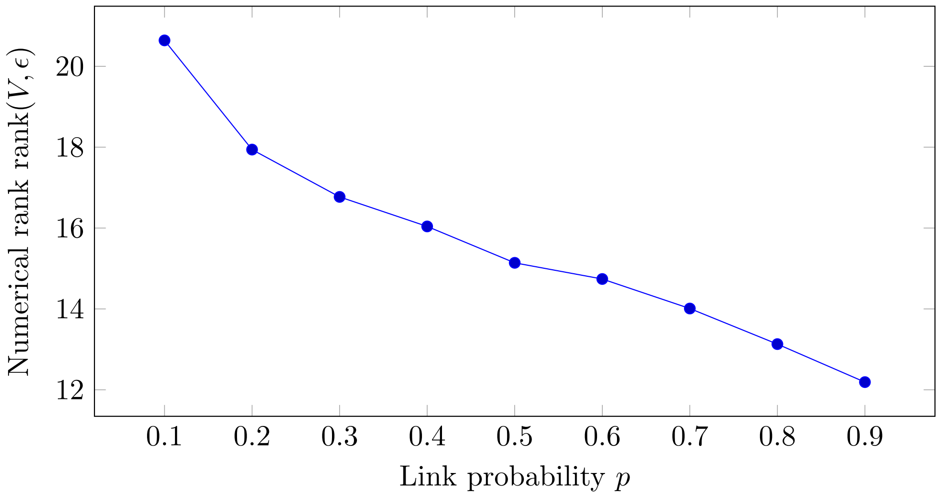

Fig. 4.2. The relationship between the numerical rank rank(V, ϵ) and the link probability p.

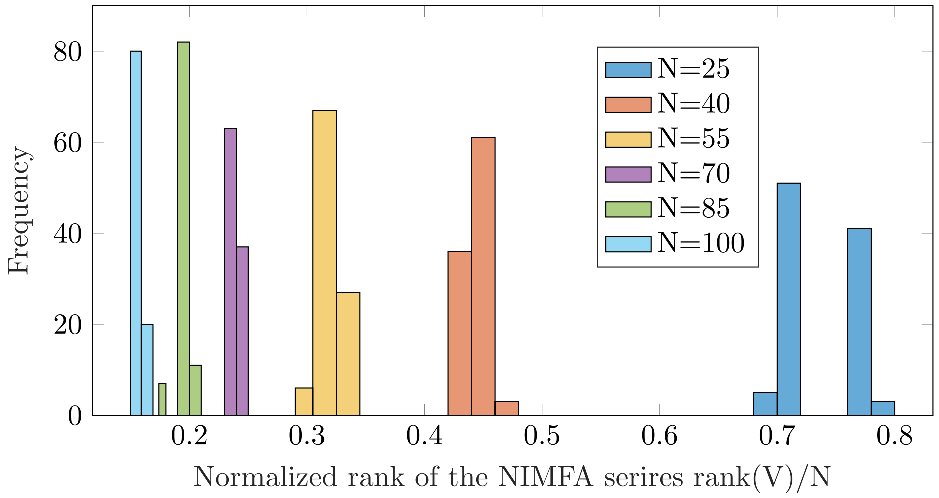

Fig. 4.3. The distribution of the normalized rank(V) with respect to different values of the network size N when the link probability is $p = 0.5$.

ear relationship with the network size N, and as N increases, rank$(V, \epsilon)/N$ decreases. It means more nodes cannot guarantee that more information can be obtained form the viral state observations. Despite this, as long as rank$(V) < N$, the network cannot be identified from solving the unconstrained linear system (4.5).

The above numberical analysis shows clearly that the numerical rank of the viral state for a single epidemic outbreak is always too small to allow for an

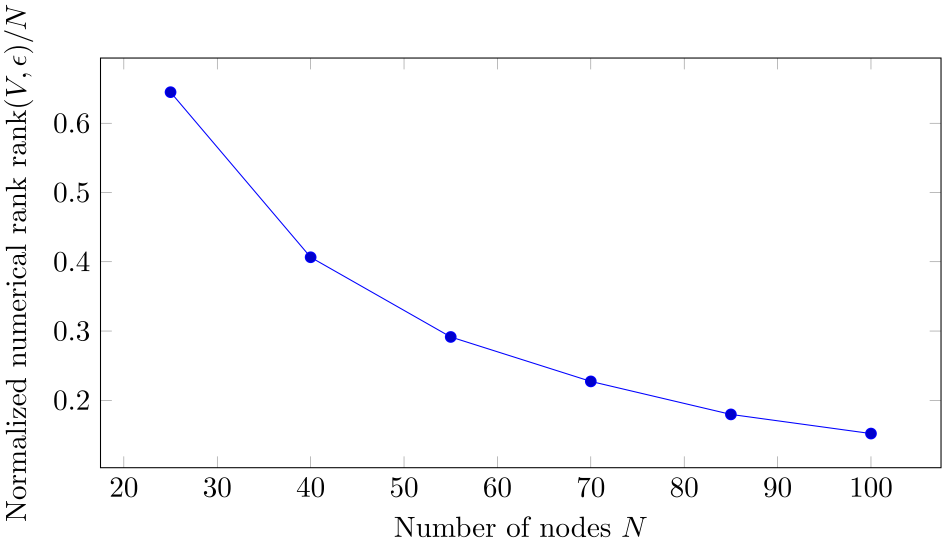

Fig. 4.4. The relationship between the normalized numerical rank rank$(V,\epsilon)/N$ and the number of nodes N.

accurate network reconstruction.

4.3 Network Reconstruction Algorithm

As the simulation results above shows, it is hard to find a network which results in a viral state matrix V of full row rank. One solution to this problem is to consider multiple epidemic outbreaks with different initial states $v[0]$ for each process. The whole viral state matrix after cascading K processes with each process is denoted as the $N \times nK$ matrix: $V_{\text{mult}} = [V_1, ..., V_K]$, where V_i is the viral state matrix of the i-th epdemic outbreak with $i = 1, ..., K$. By considering multiple outbreaks, a viral state matrix V_{mult} with $\text{rank}(V_{\text{mult}}) = N$ can be obtained. As it is shown in Figure 4.5, the numerical rank $\text{rank}(V_{\text{mult}}, \epsilon)$ increases linearly as the number of epidemic outbreaks K increases until it reaches to the full row rank, i.e., $\text{rank}(V_{\text{mult}}, \epsilon) = N$.

Similarly, we define $M_{\text{mult}} = [M_1, ..., M_K]$, where $M_i = \frac{1}{\beta_T}(C_i - \delta_T B_i)$ for the i-th NIMFA process with $i = 1, ..., K$. Therefore, similar to equation (4.8), the estimate for the adjacency matrix $\hat{A}$ becomes:

$$\hat{A} = M_{\text{mult}} V_{\text{mult}}^{\dagger} \tag{4.12}$$

The pseudo code of this algorithm is shown in Algorithm 1.

Algorithm 1 Network Reconstruction from Multiple Epidemic Outbreaks

```
Input: Multiple viral state matrices $V_1, ..., V_K$
Output: Estimated adjacency matrix $\hat{A}$
Obtain $M_1$ from $V_1$ by (4.7)
$V_{\text{mult}} \leftarrow V_1$
$M_{\text{mult}} \leftarrow M_1$
$i \leftarrow 2$
while $i \leq K$ do
    $V_{\text{mult}} \leftarrow [V_{\text{mult}}, V_i]$
    Obtain $M_i$ from $V_i$ by (4.7)
    $M_{\text{mult}} \leftarrow [M_{\text{mult}}, M_i]$
    $i \leftarrow i + 1$
end while
$\hat{A} \leftarrow$ Least-squares solution to $AV_{\text{mult}} = M_{\text{mult}}$ by QR-decomposition (4.12)
for $i = 1, ..., N$ do
    for $j = 1, ..., N$ do
        if $\hat{a}_{ij} + \hat{a}_{ij} >= 1$ then
            $\hat{a}_{ij} \leftarrow 1$
        else
            $\hat{a}_{ij} \leftarrow 0$
        end if
    end for
end for
```

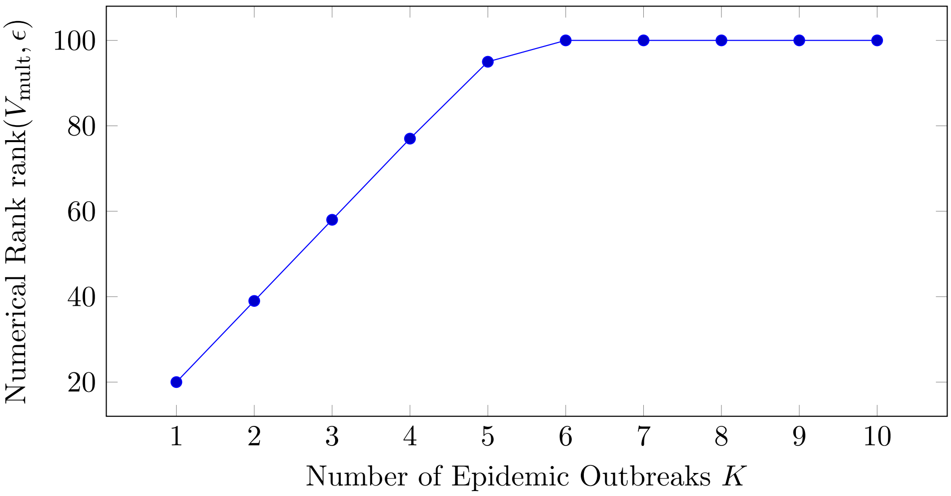

Fig. 4.5. The relationship between the numerical rank rank$(V_{\text{mult}}, \epsilon)$ and the number of epidemic outbreaks K for an Erdős-Renyi network with $N = 100$ nodes and the link probability $p = 0.1$. The effective infection rate $\frac{\beta}{\delta} = 1.1\tau_c^{(1)}$.

Reduced-size Linear Equations by Truncated Singular Value Decomposition

Considering multiple epidemic outbreaks overcomes the problem that the numerical rank rank(V) is smaller than the number of nodes N, but it still has its limitations. For large scale networks with hundreds or thousands of nodes, the numerical rank of the viral state matrix V is often does not exceed twenty, which means that there are just few linear independent rows in a matrix. In other words, a viral state matrix V with thousands of rows, only about twenty rows contain meaningful information. Therefore, if the whole matrix V is considered in the calculation, most of the memory is wasted and the computation time is greatly increased. Hence, Algorithm 1 seems to be suitable only in small-scale networks.

We propose an improvement of Algorithm 1, such that the network adjacency matrix A can also be reconstructed for large scale networks. The basic idea is to replace the viral state matrix V by its truncated singular value decomposition (TSVD) [27], which we introduce in the following.

As discussed in Section 4.2.1, the singular value decomposition for the viral state matrix V_i of the i-th outbreak is given by

$$V_i = U_i \Sigma_i Q_i^T \tag{4.13}$$

The truncated singular value decomposition of the matrix V_i is obtained from equation (4.13) by considering only the largest $\mathrm{rank}(V_i, \epsilon)$ singular values of V_i and setting the other singular values to zero, which yields:

$$V_i \simeq \widetilde{U}_i \widetilde{\Sigma}_i \widetilde{Q}_i^T, \tag{4.14}$$

where $\widetilde{U}_i$ is an $N \times \mathrm{rank}(V_i, \epsilon)$ matrix, $\widetilde{\Sigma}_i$ is a $\mathrm{rank}(V_i, \epsilon) \times \mathrm{rank}(V_i, \epsilon)$ matrix, and $\widetilde{Q}_i$ is an $n \times \mathrm{rank}(V_i, \epsilon)$ matrix. We introduce the matrix $S_i = \widetilde{\Sigma}_i \widetilde{Q}_i^T$. Then, comparing with equation (4.7), we define a new matrix M_{TSVD}:

$$M_{\mathrm{TSVD},i} = S_i^{\dagger} M_i \tag{4.15}$$

Similar to Algorithm 1, instead of concatenating the viral states V_i, the singular vectors U_i are concatenated for $i = 1, ..., K$ in this improved algorithm. So we have $U_{\mathrm{mult}} = [\widetilde{U}_1, ..., \widetilde{U}_K]$ and $M_{\mathrm{TSVD,mult}} = [M_{\mathrm{TSVD},1}, ..., M_{\mathrm{TSVD},K}]$. Finally, the adjacency matrix is estimated as

$$\hat{A} = M_{\mathrm{TSVD,mult}} U_{\mathrm{mult}}^{\dagger} \tag{4.16}$$

The pseudo code of this algorithm is shown in Algorithm 2.

Constrained Solution for $\hat{A}$

In Algorithm 1 and Algorithm 2, the estimated adjacency matrix $\hat{A}$ is computed by the QR factorization [25]. The solutions (4.12) and (4.16) are not constrained to be in the interval $[0, 1]$. As the true elements a_{ij} of the adjacency matrix A are in $\{0, 1\}$, a solution that is constrained in the interval $[0, 1]$ will yield more accuracy.

We pose a constrained linear least-squares problem based on the equation (4.16):

$$\min_A ||AU_{\mathrm{mult}} - M_{\mathrm{svd,mult}}||_2^2, \quad \text{s.t.} \quad a_{ij} \in [0, 1] \tag{4.17}$$

to ensure that every element in the estimated adjacency matrix $\hat{A}$ lies in the interval of $[0, 1]$. To solve equation (4.17), we use the command `lsqlin` in Matlab, which implements the Trust Region Refelctive algorithm which is based on the interior-reflective Newton method described in [28]. The pseudo code of this algorithm is shown in Algorithm 3.

4.4 Numerical Evaluation

4.4.1 Error Metrics

We introduce two kinds of error metrics to quantify the accuracy of the network reconstruction method introduced in Section 4.3.

Algorithm 2 Network Reconstruction from Multiple Epidemic Outbreaks by TSVD

```
 1: Input: Multiple viral state matrices $V_1, ..., V_K$
 2: Output: Estimated adjacency matrix $\hat{A}$
 3: Obtain the TSVD $\widetilde{U}_1 \widetilde{\Sigma}_1 \widetilde{Q}_1^T$ of $V_1$
 4: $S_1 \leftarrow \widetilde{\Sigma}_1 \widetilde{Q}_1^T$
 5: Obtain $M_1$ from $V_1$ by (4.7)
 6: Obtain $M_{\text{TSVD},1}$ from $M_1$ and $S_1$ by (4.16)
 7: $U_{\text{mult}} \leftarrow \widetilde{U}_1$
 8: $M_{\text{TSVD,mult}} \leftarrow M_{\text{TSVD},1}$
 9: $i \leftarrow 2$
10: while $i \leq K$ do
11:     Repeat step 3-6 for $V_i$ to obtain $\widetilde{U}_i$ and $M_{\text{TSVD},i}$
12:     $U_{\text{mult}} \leftarrow [U_{\text{mult}}, \widetilde{U}_i]$
13:     $M_{\text{TSVD,mult}} \leftarrow [M_{\text{TSVD,mult}}, M_{\text{TSVD},i}]$
14:     $i \leftarrow i + 1$
15: end while
16: $\hat{A} \leftarrow$ solution to $AU_{\text{mult}} = M_{\text{TSVD,mult}}$ by QR-decomposition
17: for $i = 1, ..., N$ do
18:     for $j = 1, ..., N$ do
19:         if $\hat{a}_{ij} + \hat{a}_{ji} >= 1$ then
20:             $\hat{a}_{ij} \leftarrow 1$
21:         else
22:             $\hat{a}_{ij} \leftarrow 0$
23:         end if
24:     end for
25: end for
```

Algorithm 3 Network Reconstruction from Multiple Epidemic Outbreaks by TSVD and constrained solutions

```
 1: Input: Multiple viral state matrices $V_1, ..., V_K$
 2: Output: Estimated adjacency matrix $\hat{A}$
 3: Obtain the TSVD $\widetilde{U}_1 \widetilde{\Sigma}_1 \widetilde{Q}_1^T$ of $V_1$
 4: $S_1 \leftarrow \widetilde{\Sigma}_1 \widetilde{Q}_1^T$
 5: Obtain $M_1$ from $V_1$ by (4.7)
 6: Obtain $M_{\text{TSVD},1}$ from $M_1$ and $S_1$ by (4.16)
 7: $U_{\text{mult}} \leftarrow \widetilde{U}_1$
 8: $M_{\text{TSVD,mult}} \leftarrow M_{\text{TSVD},1}$
 9: $i \leftarrow 2$
10: while $i \leq K$ do
11:     Repeat step 3-6 for $V_i$ to obtain $\widetilde{U}_i$ and $M_{\text{TSVD},i}$
12:     $U_{\text{mult}} \leftarrow [U_{\text{mult}}, \widetilde{U}_i]$
13:     $M_{\text{TSVD,mult}} \leftarrow [M_{\text{TSVD,mult}}, M_{\text{TSVD},i}]$
14:     $i \leftarrow i + 1$
15: end while
16: $\hat{A} \leftarrow$ solution to $\min_{A} ||AU_{\text{mult}} - M_{\text{svd,mult}}||_2^2, \quad \text{s.t.} \quad a_{ij} \in [0, 1]$
17: for $i = 1, ..., N$ do
18:     for $j = 1, ..., N$ do
19:         if $\hat{a}_{ij} + \hat{a}_{ji} >= 1$ then
20:             $\hat{a}_{ij} \leftarrow 1$
21:         else
22:             $\hat{a}_{ij} \leftarrow 0$
23:         end if
24:     end for
25: end for
```

Scaled One-norm Deviation ε_A

$$\varepsilon_A = \frac{1}{2L_{\max}}||A - \hat{A}||_1, \tag{4.18}$$

where $L_{\max} = N(N-1)/2$ is the maximal number of links for a network of N nodes.

TPR-FPR

In a two-class estimation problem, the outcomes are defined as positive (P) and negative (N). There are four possible outcomes. If the estimation is $\hat{a}_{ij} = 1$ and the true value is also $a_{ij} = 1$, then the result is true positive (TP); however if the true value is $a_{ij} = 0$ but $\hat{a}_{ij} = 1$, then it is false positive (FP). On the other hand, when both the estimation and the true value are $\hat{a}_{ij} = a_{ij} = 0$, the result is true negative (TN) and false negative (FN) is when the estimation is $\hat{a}_{ij} = 0$ while the true value is $a_{ij} = 1$. These four results can be drawn as a 2 × 2 table of confusion[29] in Table 4.1:

Table 4.1: Table of Confusion

		True Value	
		$a_{ij} = 1$	$a_{ij} = 0$
Estimated Value	$\hat{a}_{ij} = 1$	TP	FP
	$\hat{a}_{ij} = 0$	FN	TN

The true positive rate (TPR) defines the ratio of correct positive results in all positive samples, while false positive rate (FPR) defines the ratio of incorrect positive results in all negative samples.

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \tag{4.19}$$

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) \tag{4.20}$$

The estimation is accurate when $TPR = 1$ and $FPR = 0$.

4.4.2 Numerical Evaluation of the Accuracy

To study the accuracy of Algorithm 3, we apply it to Erdős-Renyi random networks and Watts-Strogatz small-world networks. We generate a Erdős-Renyi network with $N = 200$ nodes and with a link probability $p = 0.2$. Furthermore, we generate a small-world network with $N = 200$ nodes of average degree $E[d] = 0.2N$ and with a rewiring probability $p = 0.5$. by setting the parameters in this way, both networks have the same network size $N = 200$ and the same average degree $E[d] = 40$. The effective infection rate for both networks are set to be $\frac{\beta}{\delta} = 1.1\tau_c^{(1)}$, where $\tau_c^{(1)}$ is the epidemic threshold. Figure 4.6 and Figure 4.7 shows the relationship of the accuracy and the number of outbreaks. The estimation error ε_A roughly seems to decrease linearly with the increase of the number of epidemic outbreaks K until the estimation error ε_A reduces to 0 when the number of epidemic outbreaks equals $K = 5$. Compared to the small-world network, the Erdős-Renyi network has a higher estimation error before ε_A reduces to 0. Considering the TPR-FPR error metrics, we obtain Figure 4.7.

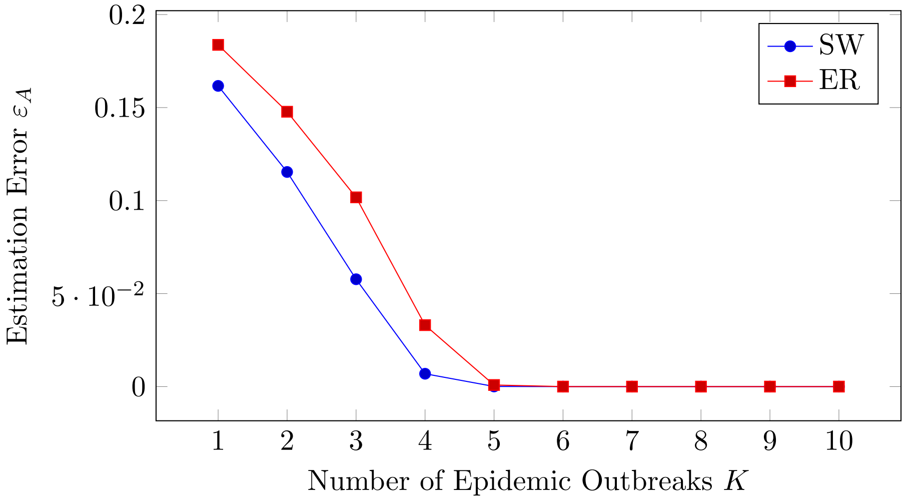

Fig. 4.6. The estimation error ε_A versus the number of epidemic outbreaks K for the Erdős-Renyi network and the small-world network.

Number of nodes N

To figure out how many outbreaks are required for an accurate estimation for networks of different sizes N, the following simulations are performed. Both the Erdős-Renyi network and the Small-world network are considered. The network size N ranges from $[100, .., 700]$. For each value of N, we generate 1000 Erdős-Renyi networks with link probability $p = 0.1$ and 1000 Small-world networks with average degree $E[d] = 0.2N$ and with a rewiring probability $p_r = 0.5$. The effective infection rate for both networks is set to $\frac{\beta}{\delta} = 1.1\tau_c^{(1)}$, where $\tau_c^{(1)}$ is the epidemic threshold. The average network reconstruction error over 1000 networks $\bar{\varepsilon}_A$ can be calculated as:

$$\bar{\varepsilon}_A = \frac{1}{1000}\sum_{l=1}^{1000} \varepsilon_{A,l}, \tag{4.21}$$

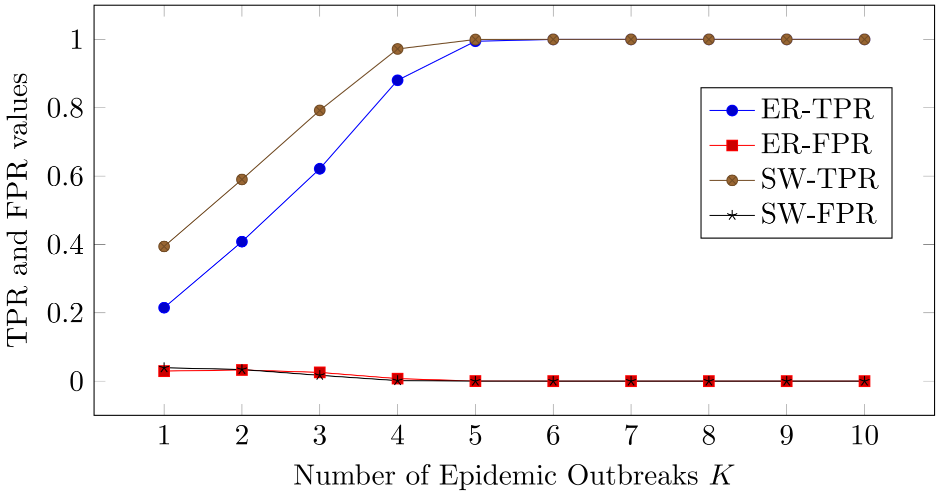

Fig. 4.7. The TPR and FPR value of the Erdős-Renyi and small-world network reconstruction results, indicating the relationship of the accuracy and the number of epidemic outbreaks K.

where $\varepsilon_{A,l}$ is the reconstruction error of the l-th network. The relationship of the averaged error $\bar{\varepsilon}_A$ and the number of outbreaks K is shown in Figure 4.8 and Figure 4.9 for the Erdős-Renyi network and the small-world network respectively. For both network models, the larger the size of the network N is, the more epidemic outbreaks are required to obtain an accurate estimation of the adjacency matrix $\hat{A}$. The averaged estmation error $\bar{\varepsilon}_A$ roughly seems to decrease linearly with the number of epidemic outbreaks K. Figure 4.10 and Figure 4.11 present the TPR and FPR values for the Erdős-Renyi network. Figure 4.12 and Figure 4.13 present the TPR and FPR values for the small-world network. As the network size increases for both epidemic models, more epidemic outbreaks are required so that TPR value reaches to 1 and FPR value reaches to 0, where an accurate estimation is obtained.

Average degree $E[d]$

Besides the number of nodes N, the average degree $E[d]$, or the sparsity of the network may also have influence on the network reconstruction accuracy and the number of required epidemic outbreaks. To find the relationship among the accuracy, the number of epidemic outbreaks and the average degree $E[d]$, the following simulations are performed. Both the Erdős-Renyi network and the small-world network are considered. The number of nodes in the network N is set to 200. The average degree $E[d]$ ranges in the interval of $[40, 160]$. The effective infection rate for both networks are set to be $\frac{\beta}{\delta} = 1.1\tau_c^{(1)}$, where $\tau_c^{(1)}$ is the epidemic threshold. For each value of the average degree $E[d]$, 1000 networks are generated. The averaged accuracy

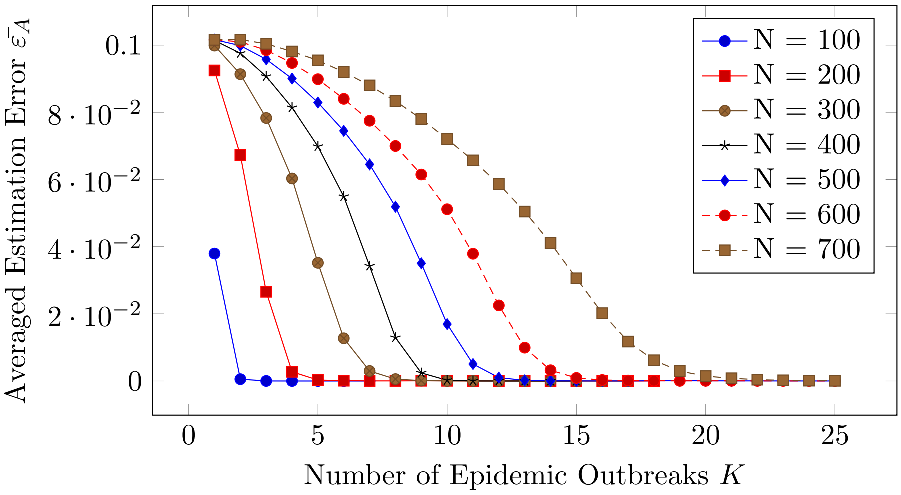

Fig. 4.8. The averaged estimation error $\bar{\varepsilon}_A$ versus the number of epidemic outbreaks K for the Erdős-Renyi networks.

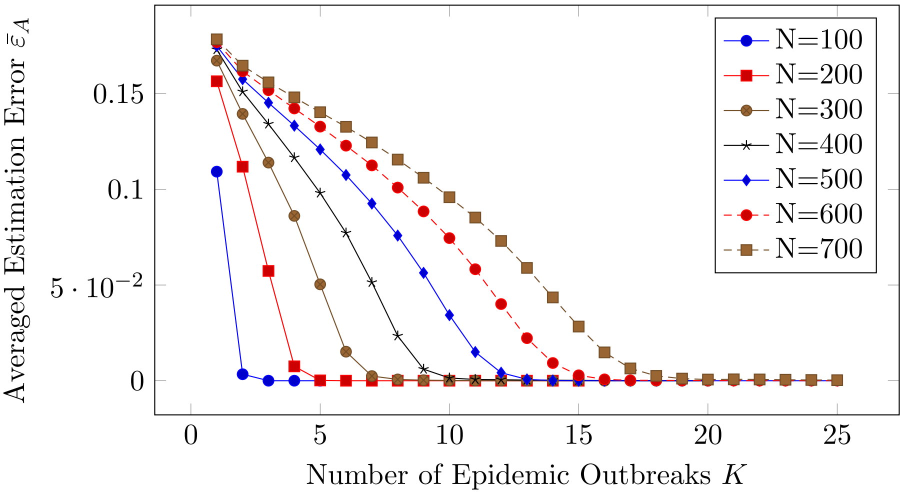

Fig. 4.9. The averaged estimation error $\bar{\varepsilon}_A$ versus the number of epidemic outbreaks K for the small-world networks.

over these 1000 networks for each value of $E[d]$ is considered.

Figure 4.14, Figure 4.15 and Figure 4.16 show the averaged estimation error $\bar{\varepsilon}_A$ and the TPR, FPR values versus the number of epidemic outbreaks K for the Erdős-Renyi networks with different average degree $E[d]$. They show that as the average degree $E[d]$ increases, the number of epidemic outbreaks required K to obtain an accurate estimation increases. The same conclusion can be drawn in the small-world network as Figure 4.17, Figure 4.18 and

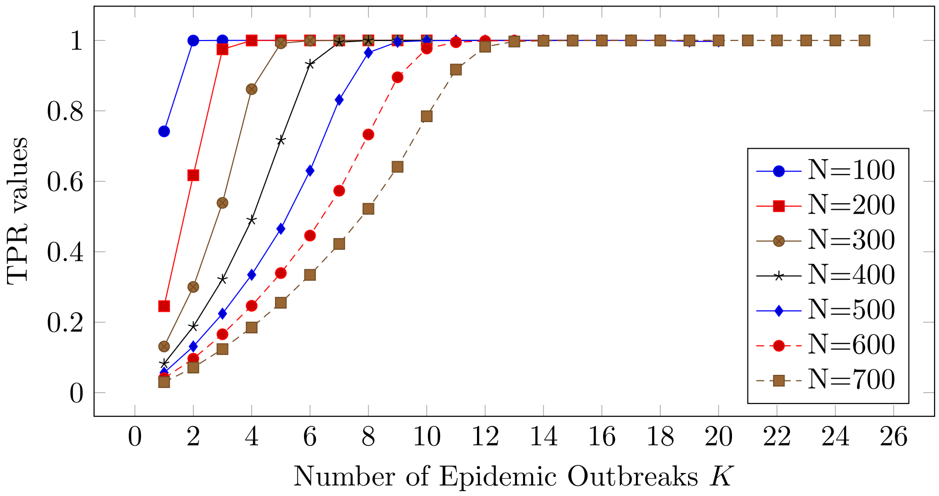

Fig. 4.10. The TPR values of the Erdős-Renyi network reconstruction results with different network sizes with respect to the number of epidemic outbreaks K.

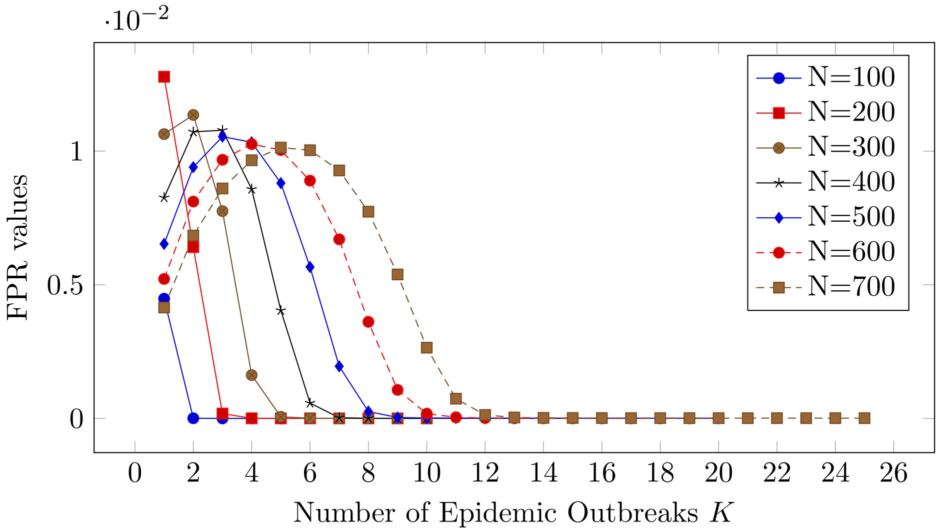

Fig. 4.11. The FPR values of the Erdős-Renyi network reconstruction results with different network sizes with respect to the number of epidemic outbreaks K.

4.19 show.

4.4.3 Network Reconstruction for Real Networks

In this section, the network reconstrution Algorithm 3 is applied to several real networks. All of them are undirected and unweighted graphs. Firstly,

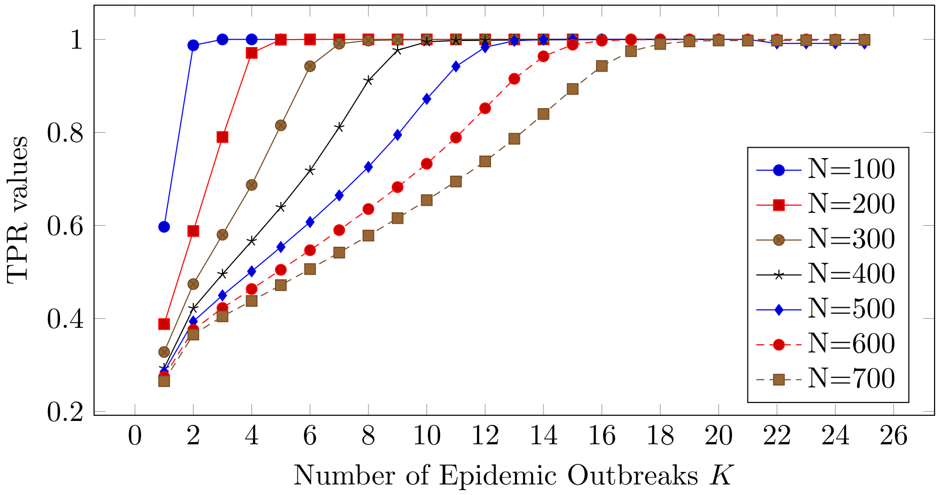

Fig. 4.12. The TPR values of the small-world network reconstruction results with different network sizes with respect to the number of epidemic outbreaks K.

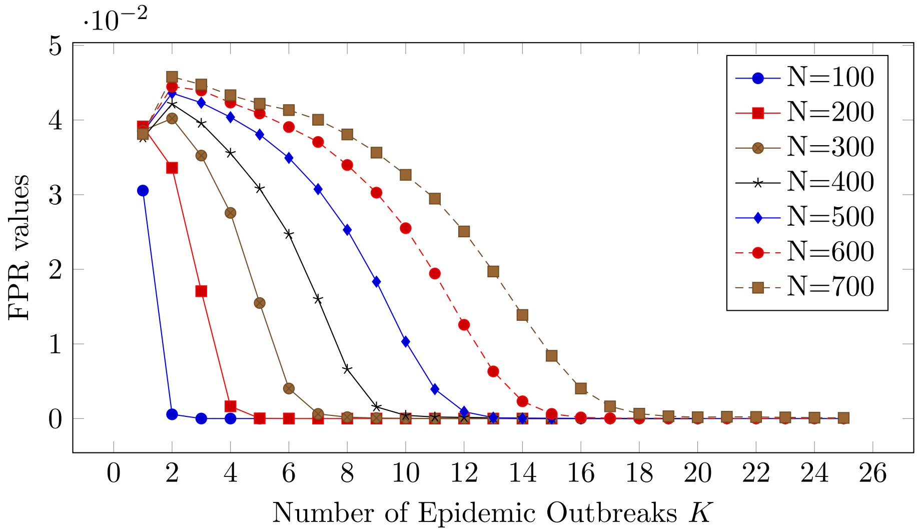

Fig. 4.13. The FPR values of the small-world network reconstruction results with different network sizes, with respect to the number of epidemic outbreaks K.

we simulate NIMFA processes on the real-world networks for 10 different initial consitions $v[0]$, which were generated randomly. Secondly, we use the generated virual states matrices V to reconstruct the adjacency matrix $\hat{A}$ by applying Algorithm 3. We consider the following real-world networks.

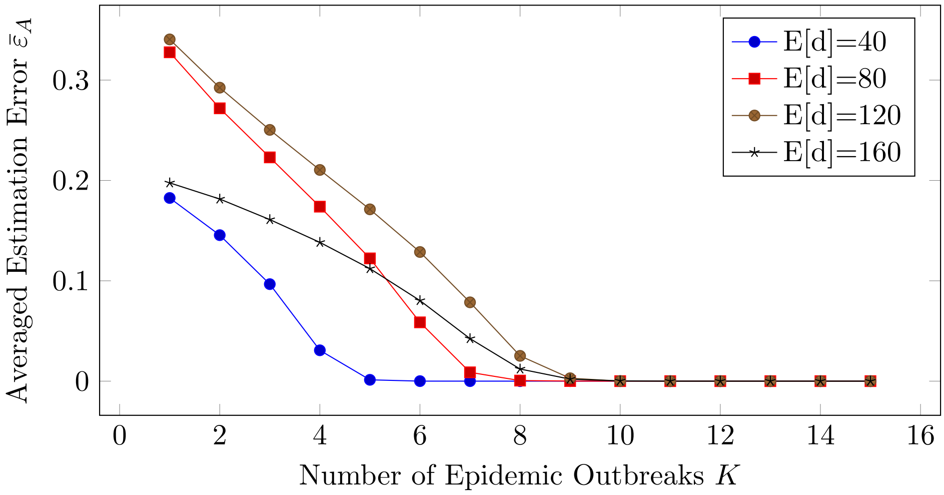

Fig. 4.14. The averaged estimation error $\bar{\varepsilon}_A$ versus the number of epidemic outbreaks K for the Erdős-Renyi networks of size $N = 200$ with different average degree $E[d]$.

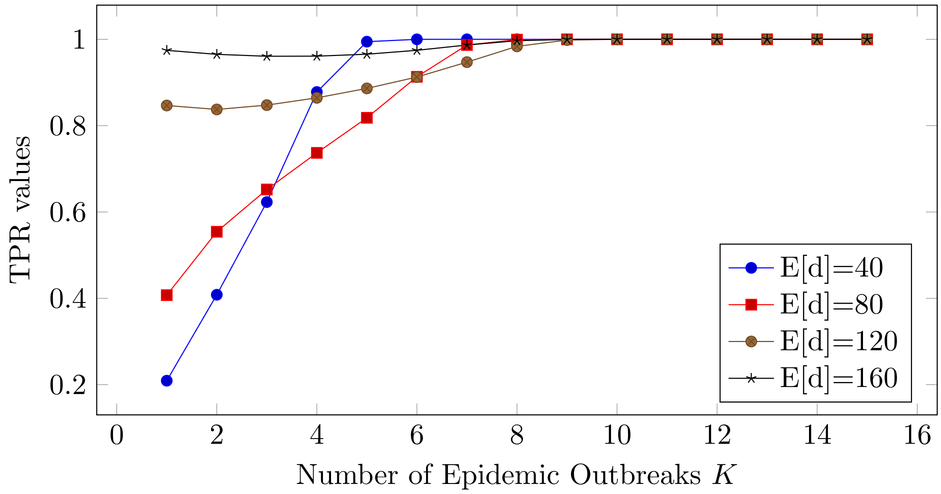

Fig. 4.15. The TPR values of the Erdős-Renyi network reconstruction results with different average degrees, with respect to the number of epidemic outbreaks K.

Contiguous USA

This network includes the 48 contiguous US-American states, which are connected by land with the other states and the district of Columbia of the United States of America [30]. Each state is denoted by a node in the network. If two states share a border, then there exists a link between these two nodes. In this dataset, the network has 49 nodes (states) and 107 links

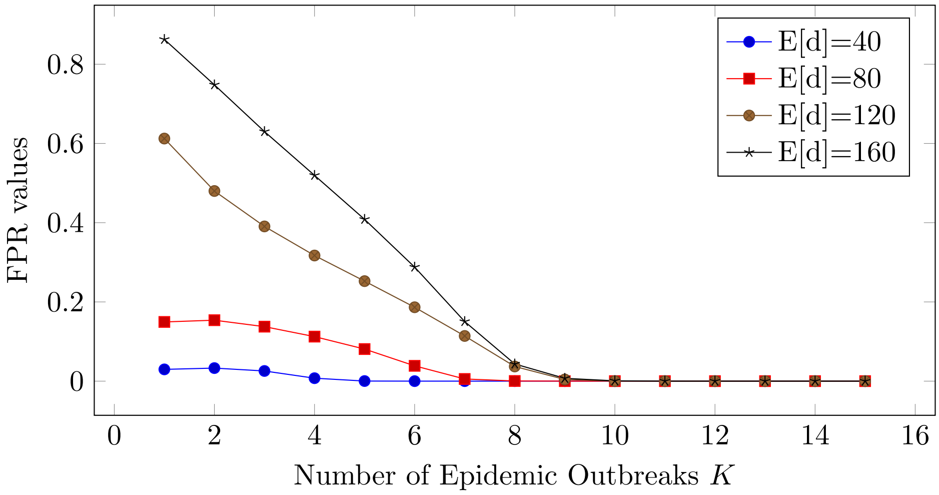

Fig. 4.16. The FPR values of the Erdős-Renyi network reconstruction results with different average degrees, with respect to the number of epidemic outbreaks K.

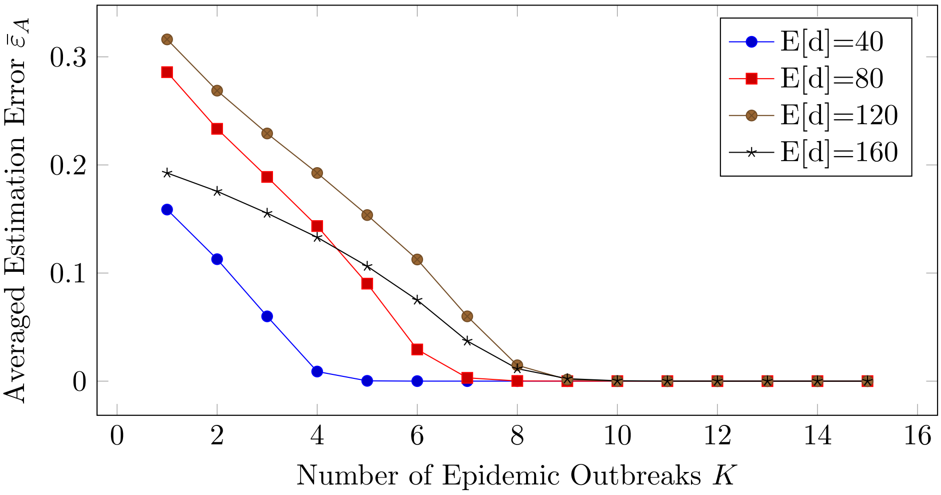

Fig. 4.17. The averaged estimation error $\bar{\varepsilon}_A$ versus the number of epidemic outbreaks K for the small-world networksof size $N = 200$ with different average degree $E[d]$.

(borders) with an average degree of 4.3673 links per node.

Euroroad

This network is an international road network, which represent roads in European countries [31]. Each node corresponds to a European city and the link between two nodes represents that there is a direct road connecting

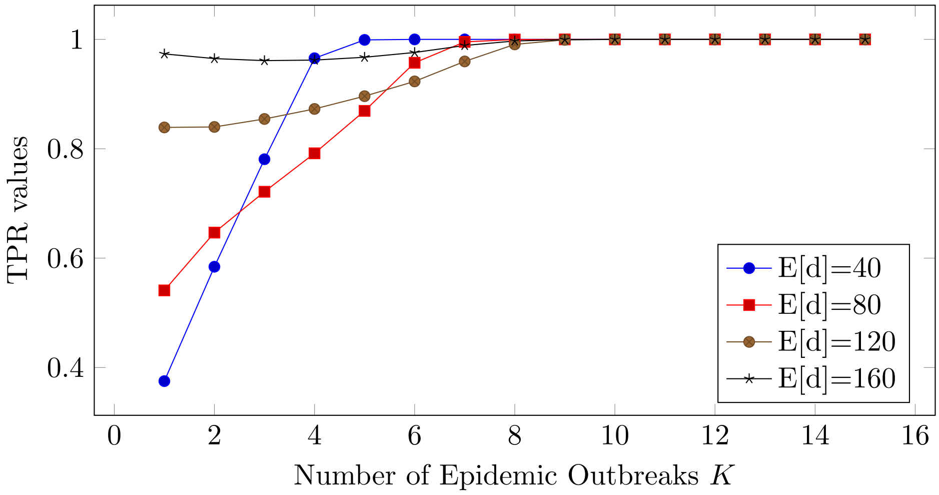

Fig. 4.18. The TPR values of the small-world network reconstruction results with different average degrees, with respect to the number of epidemic outbreaks K.

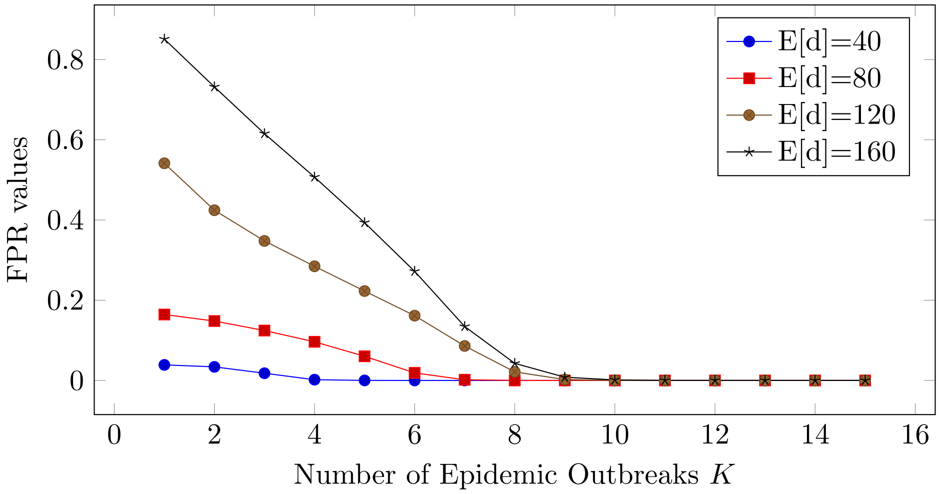

Fig. 4.19. The FPR values of the small-world network reconstruction results with different average degrees, with respect to the number of epidemic outbreaks K.

two cities. The network is neither a scale-free nor a small-world network [3]. The network has 1,174 nodes (cities) and 1,417 links (roads), whose layout is shown in Figrue 4.20 with an average degree of 2.4140 links per node.

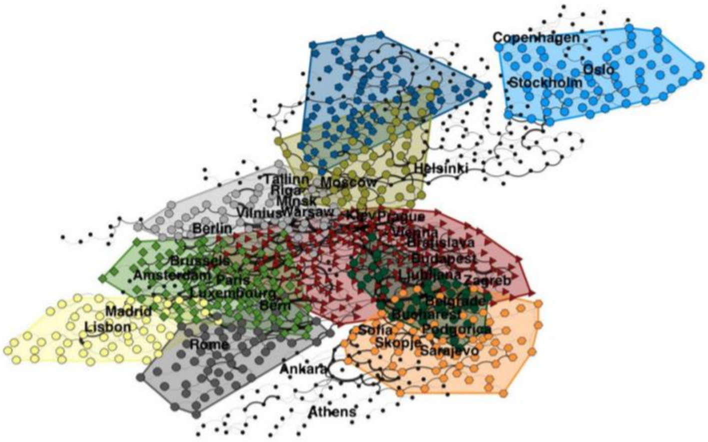

Fig. 4.20. European road network with N = 1174 nodes and 1417 edges [3]

Hamsterster

This social network contains friendships among users of the website hamsterster.com [32]. Each node represents a user. If two users are friends, then there is a link between these two nodes. This network has 1,858 nodes (users) and 12534 links (friendships), which leads to an average degree of 13.492 links per node.

Facebook

This connected network represents the friendships among a small subgroup of users of Facebook [33]. Each node represents a user of Facebook. If two users are friends, then there is a link between these two nodes. The Facebook network has 2,888 nodes (users) and 2,981 links (friendships), which leads to an average degree of 2.0644 links per node.

Table 4.2 gives an overview of these networks. Figure 4.21 show the estimation error $\bar{\varepsilon}_A$ averaged over 10 different initial viral state sequences and the TPR-FPR metrics with respect to the number of epidemic outbreaks K for these real networks. In the contiguous USA network, as the number of nodes in the network N is small, the network can be accurately reconstructed by only one epidemic outbreak. So the estimation error for the contiguous USA network equals to zero for all K, thus cannot be displayed in this semilog figure. For the Facebook network, for two of the ten viral state sequences the reconstruction error $\bar{\varepsilon}_A$ did not converge to zero, even after considering 30 epidemic outbreaks. For the other Facebook networks, the network can be accurately reconstructed by less than five epidemic outbreaks, and the low average degree of the network may be an explaination of an accurate reconstruction by only five outbreaks K on average. For the Hamsterer friendship network, which has a large number of nodes and the highest average degree, the most number of epidemic outbreaks are required to obtain an accurate estimation among these four real networks.

Table 4.2: Network Information

Network	Number of nodes N	Number of links L	Average degree $E[d]$
Contiguous USA	49	107	4.3673
Euroroad	1174	1417	2.4140
Hamsterster friendships	1858	6594	13.492
Facebook	2888	2981	2.0644

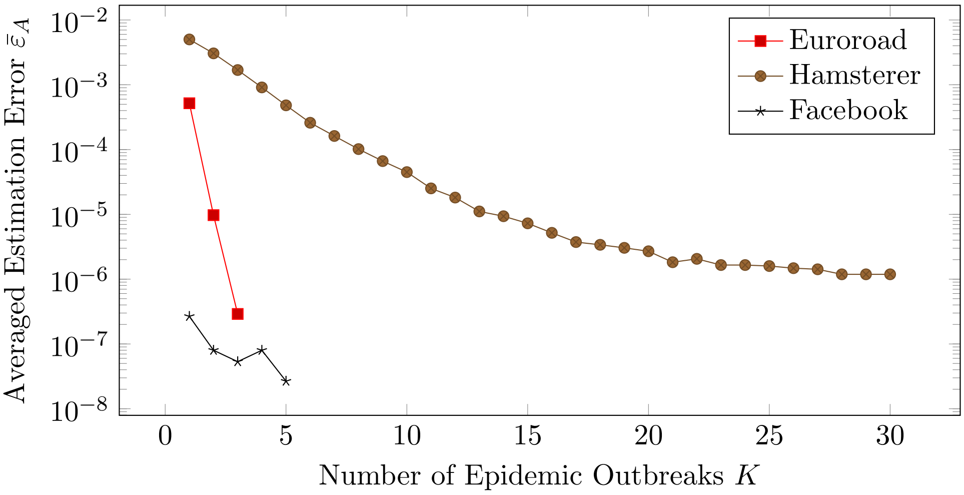

Fig. 4.21. The averaged estimation error $\bar{\varepsilon}_A$ versus the number of epidemic outbreaks K for real-world networks. Discontinued curves refer to zero error. The contiguou USA network has been exactly reconstructed with only one outbreak ($\bar{\varepsilon}_A = 0$) for all realizations. For the Facebook network, only the estimation error of the eight (of ten total) realizations which converged to $\bar{\varepsilon}_A = 0$ are plotted.

Chapter 5

Conclusions and Outlook

5.1 Conclusions

This thesis focuses on reconstructing the network topology from observing the viral state series of the NIMFA epidemic process. In order to achieve this purpose, our contribution is as follows.

Firstly, we numerically evaluate the performance of the approximated models: the sampled-time SIS model and the discrete-time NIMFA model with respect to the continuous-time SIS model, and studying the feasibility of network reconstruction. The discrete-time NIMFA model is finally decided to be applied in the following network reconstruction process.

Secondly, the basic data model, which is derived from the discrete-time NIMFA equation is studied. We found that it is not possible to uniquely estimate the adjacency matrix A from the viral state series of a single epidemic outbreak for the majority of networks, since the rank of the viral state series is much smaller than the number of nodes N.

Thirdly, to solve the problem of the rank deficiency, we implement a network reconstruction algorithm that estimates the adjacency matrix A from the viral state series of multiple epidemic outbreaks. To improve the estimation accuracy and solve the network reconstruction problem more efficiently, we resort to the truncated singular value decomposition and a constrained linear least-squares formulation of the network reconstruction problem. The larger the network is and the larger the average degree is, the more number of epidemic outbreaks are required to obtain an exact estimation of the network.

5.2 Outlook

The network reconstruction algorithm can be perfectly applied in small to medium size networks. In larger networks, more epidemic outbreaks are required and the computation time grows accordingly. Therefore, how to improve the algorithm so that it can solve the network reconstruction problem faster is a topic for future work. Then, more real networks with larger size can be estimated by the network reconstruction algorithm.

Bibliography

[1] P. Van Mieghem, *Performance analysis of complex networks and systems*. Cambridge University Press, 2014.

[2] K. Anderson, S. Lee, and C. Menassa, “Effect of social network type on building occupant energy use,” in *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. ACM, 2012, pp. 17–24.

[3] L. Šubelj and M. Bajec, “Robust network community detection using balanced propagation,” *The European Physical Journal B*, vol. 81, no. 3, pp. 353–362, 2011.

[4] F. Brauer, C. Castillo-Chavez, and C. Castillo-Chavez, *Mathematical models in population biology and epidemiology*. Springer, 2012, vol. 40.

[5] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, “Revealing strengths and weaknesses of methods for gene network inference,” *Proceedings of the national academy of sciences*, 2010.

[6] R. J. Wilson, “An eulerian trail through Königsberg,” *Journal of graph theory*, vol. 10, no. 3, pp. 265–275, 1986.

[7] P. Erdős and A. Renyi, “On random graphs I,” *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[8] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks,” *Nature*, vol. 393, no. 6684, p. 440, 1998.

[9] I. de Sola Pool and M. Kochen, “Contacts and influence,” *Social networks*, vol. 1, no. 1, pp. 5–51, 1978.

[10] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[11] K. Choromański, M. Matuszak, and J. Mikisz, “Scale-free graph with preferential attachment and evolving internal vertex structure,” *Journal of Statistical Physics*, vol. 151, no. 6, pp. 1175–1183, 2013.

[12] R. Albert, "R. albert and a.-l. barabási, rev. mod. phys. 74, 47 (2002)." *Rev. Mod. Phys.*, vol. 74, p. 47, 2002.

[13] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of modern physics*, vol. 87, no. 3, p. 925, 2015.

[14] B. Prasse and P. Van Mieghem, "Maximum-likelihood network reconstruction for SIS processes is NP-hard," *arXiv preprint arXiv:1807.08630*, 2018.

[15] P. Van Mieghem, J. Omic, and R. Kooij, "Virus spread in networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 1, pp. 1–14, 2009.

[16] P. Van Mieghem, "Approximate formula and bounds for the time-varying susceptible-infected-susceptible prevalence in networks," *Physical Review E*, vol. 93, no. 5, p. 052312, 2016.

[17] Q. Liu and P. Van Mieghem, "Die-out probability in sis epidemic processes on networks," in *International Workshop on Complex Networks and their Applications.* Springer, 2016, pp. 511–521.

[18] B. Prasse and P. Van Mieghem, "Exact network reconstruction from complete sis nodal state infection information seems infeasible," *IEEE Transactions on Network Science and Engineering*, 2018.

[19] R. Pastor-Satorras and A. Vespignani, "Epidemic dynamics and endemic states in complex networks," *Physical Review E*, vol. 63, no. 6, p. 066117, 2001.

[20] K. Devriendt and P. Van Mieghem, "Unified mean-field framework for susceptible-infected-susceptible epidemics on networks, based on graph partitioning and the isoperimetric inequality," *Physical Review E*, vol. 96, no. 5, p. 052314, 2017.

[21] C. Li, R. van de Bovenkamp, and P. Van Mieghem, "Susceptible-infected-susceptible model: A comparison of N-intertwined and heterogeneous mean-field approximations," *Physical Review E*, vol. 86, no. 2, p. 026116, 2012.

[22] K. E. Atkinson, *An introduction to numerical analysis.* John Wiley & Sons, 2008.

[23] P. E. Pare, J. Liu, C. L. Beck, B. E. Kirwan, and T. Basar, "Analysis, identification, and validation of discrete-time epidemic processes," *arXiv preprint arXiv:1710.11149*, 2017.

[24] B. Prasse and P. Van Mieghem, "Network reconstruction and prediction of epidemic outbreaks for NIMFA processes," *TU Delft report*, 2018.

[25] S. Boyd and L. Vandenberghe, "Vectors, matrices, and least squares," *Available: stanford. edu/class/ee103/mma. pdf*, 2016.

[26] G. H. Golub and C. F. Van Loan, *Matrix computations.* JHU Press, 2012, vol. 3.

[27] P. C. Hansen, "Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank," *SIAM Journal on Scientific and Statistical Computing*, vol. 11, no. 3, pp. 503–518, 1990.

[28] T. F. Coleman and Y. Li, "A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1040–1058, 1996.

[29] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[30] http://konect.uni-koblenz.de/networks/contiguous-usa.

[31] http://konect.uni-koblenz.de/networks/subelj_euroroad.

[32] http://konect.uni-koblenz.de/networks/petster-friendships-hamster.

[33] http://konect.uni-koblenz.de/networks/ego-facebook.