Stochastic spreading processes on networks

by

Aram Vajdi

B.S., Razi University, Iran, 2006

M.S., Kurdistan University, Iran, 2009

M.S., Kansas State University, 2015

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Abstract

Spreading processes appear in diverse natural and technological systems, such as the spread of infectious diseases and the dissemination of information. It has been demonstrated that the structure of interaction among population members can dramatically influence spreading dynamics. Therefore, researchers have focused on studying spreading processes over complex networks, where interaction among individuals could be highly heterogeneous. This dissertation aims to add to the current understanding of networked spreading processes by investigating various aspects of the Susceptible-Infected-Susceptible (SIS) model.

Our first contribution is related to the inverse problem of continuous time SIS spreading over a graph. In other words, we show the possibility of inferring the underlying network from observations on the node states through time. We formulate the inverse problem as a Bayesian inference problem and find the posterior probabilities for the existence of uncertain links.

Second, we study the SIS spreading process over time dependent networks, where the contact network's links are not permanent. To analyze the effect of link durations on the epidemic threshold of the SIS process, we develop a temporal network model. In this model, the temporal links result from the transition of nodes between two auxiliary node states, namely active and inactive. Combining the dynamics of the network and the spreading process, we derive the mean-field equations that describe SIS spreading processes over such temporal networks. The analysis of these equations reveals the effect of link durations on the epidemic threshold in the SIS process.

Third, we study the localization of epidemics in the SIS process. In general, the SIS model has an absorbing state where all individuals are healthy. However, depending on the infection rate value, this process can reach a metastable state, where the infection does not die out. In this metastable state, some parts of the network can be disproportionately infected.

We quantify the infection dispersion in the network, and formulate a convex optimization problem to find an upper bound for the dispersion of infection in the network.

Finally, we focus on the estimation of spreading data from partially available information. In general, various spreading-related functions are defined over the nodes of a network. Assuming access to the values of a function for a subset of the nodes, we use the concept of *effective resistance distance* and feed forward neural networks, to estimate the function for the remaining nodes.

Although this dissertation focuses on the SIS model, the methods we have presented and developed here are applicable to a broad range of stochastic networked spreading processes. The exact mathematical treatment of such processes is intractable due to their exponential space size, and therefore there are still various unknown aspects of their behavior that require further work. Our studies in this dissertation advance the current knowledge about networked spreading models.

Stochastic spreading processes on networks

by

Aram Vajdi

B.S., Razi University, Iran, 2006

M.S., Kurdistan University, Iran, 2009

M.S., Kansas State University, 2015

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Approved by:

Major Professor
Caterina Scoglio

Copyright

Abstract

Spreading processes appear in diverse natural and technological systems, such as the spread of infectious diseases and the dissemination of information. It has been demonstrated that the structure of interaction among population members can dramatically influence spreading dynamics. Therefore, researchers have focused on studying spreading processes over complex networks, where interaction among individuals could be highly heterogeneous. This dissertation aims to add to the current understanding of networked spreading processes by investigating various aspects of the Susceptible-Infected-Susceptible (SIS) model.

Our first contribution is related to the inverse problem of continuous time SIS spreading over a graph. In other words, we show the possibility of inferring the underlying network from observations on the node states through time. We formulate the inverse problem as a Bayesian inference problem and find the posterior probabilities for the existence of uncertain links.

Second, we study the SIS spreading process over time dependent networks, where the contact network's links are not permanent. To analyze the effect of link durations on the epidemic threshold of the SIS process, we develop a temporal network model. In this model, the temporal links result from the transition of nodes between two auxiliary node states, namely active and inactive. Combining the dynamics of the network and the spreading process, we derive the mean-field equations that describe SIS spreading processes over such temporal networks. The analysis of these equations reveals the effect of link durations on the epidemic threshold in the SIS process.

Third, we study the localization of epidemics in the SIS process. In general, the SIS model has an absorbing state where all individuals are healthy. However, depending on the infection rate value, this process can reach a metastable state, where the infection does not die out. In this metastable state, some parts of the network can be disproportionately infected.

We quantify the infection dispersion in the network, and formulate a convex optimization problem to find an upper bound for the dispersion of infection in the network.

Finally, we focus on the estimation of spreading data from partially available information. In general, various spreading-related functions are defined over the nodes of a network. Assuming access to the values of a function for a subset of the nodes, we use the concept of *effective resistance distance* and feed forward neural networks, to estimate the function for the remaining nodes.

Although this dissertation focuses on the SIS model, the methods we have presented and developed here are applicable to a broad range of stochastic networked spreading processes. The exact mathematical treatment of such processes is intractable due to their exponential space size, and therefore there are still various unknown aspects of their behavior that require further work. Our studies in this dissertation advance the current knowledge about networked spreading models.

Table of Contents

List of Figures

Acknowledgments

I would like to express my gratitude to Dr. Caterina Scoglio for all her supports during my research. I appreciate all the opportunities and helps she provided for me.

I want to thank Dr. Faryad Darabi Sahneh who helped me start doing research in the field of spreading processes. I would also like to thank my committee members, Dr. Pietro Poggi-Corradini, Dr. Nathan Albin, and Dr. Don Gruenbacher for the guidance and comments to improve this dissertation. I also thank Dr. Arslan Munir who accepted to serve as chairperson of the examining committee for my doctoral degree.

Preface

This dissertation with title “Stochastic spreading prepossess on network” is submitted for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering at Kansas State University. The research has been performed under the supervision of Prof. Caterina Scoglio. Some chapters in the dissertation are adopted from our published peer-reviewed articles:

- Vajdi, Aram, David Juher, Joan Saldaña, and Caterina Scoglio. “A multilayer temporal network model for STD spreading accounting for permanent and casual partners.” Scientific reports 10, no. 1 (2020): 1-12.

- Vajdi, Aram, and Caterina M. Scoglio. “Identification of missing links using susceptible-infected-susceptible spreading traces.” IEEE Transactions on Network Science and Engineering 6, no. 4 (2018): 917-927.

- Sahneh, Faryad Darabi, Aram Vajdi, and Caterina Scoglio. “Delocalized epidemics on graphs: A maximum entropy approach.” In 2016 American Control Conference (ACC), pp. 7346-7351. IEEE, 2016.

- Sahneh, Faryad Darabi, Aram Vajdi, Heman Shakeri, Futing Fan, and Caterina Scoglio. “GEMFsim: A stochastic simulator for the generalized epidemic modeling framework.” Journal of computational science 22 (2017): 36-44.

- Sahneh, Faryad Darabi, Aram Vajdi, Joshua Melander, and Caterina M. Scoglio.“Contact adaption during epidemics: A multilayer network formulation approach.” IEEE Transactions on Network Science and Engineering 6, no. 1 (2017): 16-30.

Chapter 1

Introduction

1.1 Stochastic Spreading Processes Over Networks

Spreading processes appear in diverse natural and technological systems. Examples of such processes are the spread of infectious diseases in biological systems, dissemination of information and ideas in human populations and social networks, and the propagation of malware and fault in technological networks. To understand, predict, and control spreading processes, researchers rely on mathematical models that aspire to describe the underlying mechanisms of their propagation[2–8]. In simple spreading models, the structure of interactions among individuals is ignored, and a population is categorized into different subpopulations that reflect the nature of the spreading process. For instance, classical epidemiological models[3;9] define states (or compartment) such as *immune*, *susceptible*, *exposed*, *infectious*, *symptomatic*, *recovered*, *dead*, *vaccinated*, and determine rules for moving individuals from one state to another, assuming the entire population is fully mixed.

During the past two decades, it has been demonstrated that the structure of interaction among population members can dramatically influence the spreading dynamics[10–15]. Therefore, researchers have focused on studying spreading processes over complex networks where interaction among individuals could be highly heterogeneous[16–22]. The motivation for such studies is rooted in the complexity of modern societies and the advent of new technologies that have created complex interconnected systems.

Many types of dynamics can be conceptualized over a complex network[23], such as random walks, diffusion[24] , synchronization[25], influence propagation[26], complex contagion[27]. Among them, stochastic spreading processes over networks[17;18] have drawn substantial attention from researchers of different backgrounds. In such a model, the network's nodes represent entities that can assume various states, and the network's links represent interactions among the nodes that induce the transition of nodes between states. Specifically, stochastic spreading processes over networks are effective when the description of the process at the node level includes some uncertainty, which can be described using such models. For instance, in a biological network, the infection transmission time from an infectious node to a healthy neighbor is a random variable whose distribution can depend on the disease and behavior of individuals. Indeed, one of the main motivations for conducting research on stochastic networked spreading process is to understand and control the collective behavior, even though there are different sources of uncertainty at the individual node level. For example, analysis of the stochastic susceptible-infected-susceptible (SIS) model over complex networks has clarified the role of the network structure in the emergence of the endemic state, which in turn provides means to control the epidemic by altering the network structure or adopting other possible measures[18]. The obvious real-world instance of stochastic spreading process is the study of infectious disease transmission. However, such a modeling framework can be applied to study viral information dissemination among users of online social networks, the propagation of malware or fault in technological networks, or any other stochastic spreading process that can happen in a networked system as a result of interactions among its agents.

Although extensive studies have been conducted over the topic of stochastic networked spreading process, there are various topics in this field that needs to be explored. In this dissertation, we aim to improve our understanding of 1) spreading processes over temporal networks, 2) identification of interaction from node's states transitions, 3) epidemic localization, and 4) interpolation of networked spreading data.

1.2 Research Questions

Some of the standard networked spreading models include SI, SIS, SIR and SEIR, where S (Susceptible), E (Exposed), I (Infectious) and R (Recovered) denote the node states. Although these models have different transition rules and node states, at a theoretical level, they can be described using a common framework[19]. In this dissertation, we study several aspects of the SIS model, yet some of our methods can be applied to the other stochastic models.

A theoretical question we aim to answer in this dissertation is the possibility of inferring unknown network links from observed Susceptible-Infected-Susceptible (SIS) temporal traces. We know that the underlying network affects the epidemic course, and the states that nodes assume through time. For this study, we assume a setting where we observe the transitions of nodes among the two states S and I through time. Using such observations, we want to find the probabilities for the existence of links among different nodes. Such links represent hidden interactions among the nodes.

Another aspect of the SIS spreading, which we explore in this dissertation, is the effect of links' duration in temporal networks over the spreading of infection. This study is particularly relevant in the context of sexually transmitted diseases (STD) spreading. Recent findings have stressed the increasing role of casual partnerships in STDs spreading and we aim to quantify the effect of such temporal links by analyzing the SIS spreading over a temporal network model that captures casual partnerships.

The third aspect of the SIS model we study in this dissertation is epidemic localization. The SIS epidemic process on complex networks can show metastability, resembling an endemic equilibrium. In a general setting, the metastable state can either involve a large portion of the network, or be localized on small subgraphs of the contact network. Here, we aim to quantify the localization of an epidemic and calculate its size for a given underlying network and transmission parameters.

In our final study we investigate the interpolation of stochastic networked spreading data. For an SIS spreading or any other type of spreading, various spreading-related functions can be defined over the network nodes. In this study, we want to explore the possibility of estimating such functions assuming access to the value of these functions over a subset of the nodes.

1.2.1 Contributions

Below is a summary of the main contributions of this dissertation:

1. We developed a software tool capable of simulating a broad range of stochastic spreading models with arbitrary transition time distribution(chapter 2).

2. We derived the likelihood of observed SIS temporal traces and inferred the probabilities for the existence of uncertain links. (chapter 3).

3. We developed a time dependent network model appropriate for studying STDs spreading and derived the epidemic threshold for the SIS spreading over such a network (chapter 4).

4. We proposed a dispersion entropy measure to quantify the localization of infections in a generic contact graph and formulated a maximum entropy problem to find an upper bound for the dispersion entropy of the possible metastable state in the SIS process (chapter 5).

5. We developed a new approach that relies on the effective resistance distance and feed-forward neural network to interpolate spreading data (chapter 6).

1.3 Dissertation Organization

The dissertation is organized as follows. In the rest of this chapter, we introduce the SIS process and discuss the exact and approximate equations that govern the process dynamics. In chapter 2, we explain the computational methods for simulating spreading processes in general. We introduce the software we developed for simulation of networked spreading processes, and explain the application of such software. In chapter 3, we first derive the likelihood of observing an SIS trace, and we proceed by formulating a Bayesian inference problem that uses the likelihood of observed traces to infer the existence of uncertain links. Later, in that chapter we perform numerical experiments to validate the proposed Bayesian method. In chapter 4, we first develop a temporal network model and discuss its implication. Later, we develop a meanfield approximation that describes the SIS spreading over such a network. By analyzing the meanfield equations, we study the effect of the temporal network link durations over the SIS spreading. In chapter 5, the phenomenon of localized epidemics in SIS processes is explored. We propose entropy as a measure of epidemic localization and find an upperbound for this measure. Chapter 6 is devoted to the interpolation methods for spreading data. In that chapter, we proposed a method that uses the effective resistance distance as a measure of similarity between the nodes to estimate the unknown spreading data.

1.4 Background: Networked SIS Spreading

In this section we explain the Susceptible-Infected-Susceptible (SIS) model of networked spreading. By discussing this model we explain important concepts for stochastic spreading processes over networks.

In the SIS model, the population is represented by a network of N nodes $G = \{V, E\}$, where V is the set of nodes and $E \subseteq V \times V$ denotes the set of edges between the nodes. An edge represents possible means of infection transmission between the nodes. For the contact network, the adjacency matrix $A = [a_{ij}] \in \mathbb{R}^{N\times N}$ is defined with the elements $a_{ij} = 1$ if and only if $(i, j) \in E$ else $a_{ij} = 0$.

SIS model assumes the state of node i at time t, denoted by $x_i(t)$, is a random variable and $x_i(t) = 0$ if node i is susceptible or $x_i(t) = 1$ if it is infected. A susceptible node becomes infected through interaction with infected neighbors in the network. Moreover, an infected node recovers by itself after some time, and becomes susceptible again. We assume the infection processes are independent. In other words, if two nodes are trying to infect a common neighbor, the two nodes act independently from each other and the susceptible node becomes infected by the first neighbor that successfully transmits the infection. In general, the transition time from the infected state to the susceptible state is a random variable that can have any distribution. In the same manner, the infectious time is a random variable.

In the Markovian SIS model, the recovery time for an infected node is exponentially distributed with a curing rate $\delta \in \mathbb{R}^+$. In other words, if a node is infected the probability for that node to stay infected, exponentially decreases with time. Similarly, if a susceptible node is in contact with an infected node, the probability to stay susceptible exponentially decreases with time by the infection rate $\beta \in \mathbb{R}^+$. If a susceptible node is in contact with more than one infected neighbor, the infection occurs at rate $\beta y_i(t)$, where $y_i(t) \triangleq \sum_{j=1}^{N} a_{ij} x_j(t)$ is the number of infected neighbors. The exact mathematical treatment of the Markovian SIS process requires considering the joint state of all the nodes $\mathbf{X} \triangleq [x_1, ..., x_N]$, in other words, the network state. Indeed, the network state defines a continuous-time Markov process over a space consisting of 2^N possible states. Figure 1.1, shows the Markov process and the transitions for a network of two nodes. In this figure, we can see the absorbing state of the Markov process is the state where both nodes are susceptible. For a network with a small number of nodes, it is possible to write the Kolmogorov equations for the Markov process resulting from the SIS process. However, when the number of nodes grows, the number of possible states for the Markov process grows exponentially, i.e., 2^N. This hinders the exact mathematical treatment of the SIS process over networks.

To derive the N-intertwined[18] approximation for the Markovian SIS process, we can use the following equation obtained from the node-level description of the SIS process

$$\begin{aligned} \frac{d}{dt}E[x_i] &= \beta \sum a_{ij} E[(1-x_i)x_j] - \delta E[x_i] \qquad (1.1) \\ &= \beta \sum a_{ij} E[x_j] - \beta \sum a_{ij} E[x_i x_j] - \delta E[x_i] \end{aligned}$$

for $i \in \{1, ..., N\}$. In this equation $E(x_i)$ is the expected value for the node state random variable x_i. Indeed, $E(x_i)$ is probability of finding node i infected and similarly, $E[(1-x_i)x_j]$ is the joint probability that node i is susceptible and node j is infected. Equation (1.1) is

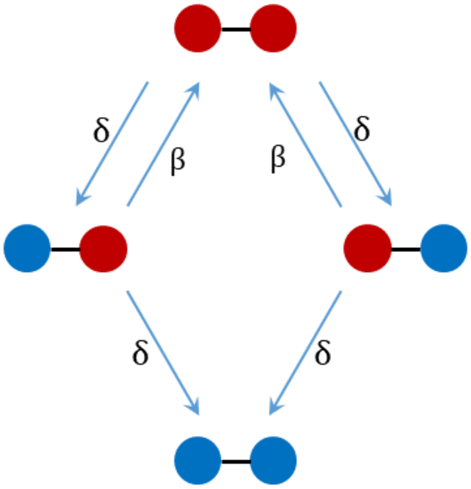


Figure 1.1: *The Markov process for the network state in the SIS spreading process for a network of two nodes. Red and blue color represent infectious and susceptible state, respectively. β and δ are infection and recovery rates.*

not a closed system as the evolution of $E[x_i]$ depends on the joint probabilities of the pairs $x_i x_j$. Furthermore, if we proceed to derive the time derivative of $E[x_i x_j]$, it turns out the time derivative depends on the higher order terms $E[x_i x_j x_k]$ which are the expected values of the triplets. The procedure goes on until we reach a closed system of $2^N - 1$ equations involving $E[x_i ... x_N]$. Such exponentially enormous state space of the exact model challenges the feasibility of any analytical investigation of the exact SIS process.

To derive approximated results, researchers use the mean-field approximation where the term $E[x_i x_j]$ in equation (1.1) is approximated by the multiplication of marginal probabilities $E[x_i]E[x_j]$[18;28]. By applying this approximation equation (1.1) can be rewritten as

$$\frac{d}{dt}E[x_i] = \beta \sum a_{ij} E[(1 - x_i)]E[x_j] - \delta E[x_i] \tag{1.2}$$

Equation (1.2) and similar types of approximate equations have been the starting points in analyzing networked spreading processes[18;29;30]. For example, by analyzing equation (1.2), it is shown for $\beta/\delta < \lambda_{max}^{-1}(A)$, where $\lambda_{max}(A)$ is the largest eigenvalue of the adjacency matrix

A, the prevalence of infection $\sum_{i=1}^{N} E[x_i]$ in the SIS process dies out exponentially fast[18]. This result has motivated research papers on the optimization of the largest eigenvalue of adjacency matrices to control the SIS spreading process[31;32]. Finally we want to emphasize two important facts about the N-intertwined mean-field equations. First, these equations are only relevant to Markovian processes. Second, they provide an upper-bound for the nodal infection probabilities[28].

Chapter 2

Simulation of Stochastic Spreading Processes [1]

2.1 Introduction

In general, the exact mathematical treatment of stochastic networked spreading process is not tractable, even for the simplest spreading models and a small number of nodes in the network. This problem stems from the fact that in the exact analysis we need to consider the network state, in other words, the state of all the nodes concurrently, instead of each node's state independently. Therefore, we have developed a software tool that can simulate the exact stochastic process for a broad range of networked spreading models.

In this chapter, we introduce two computational tools that can numerically simulate a broad range of spreading processes over complex networks. These two tools are based on the Gillespie algorithm[34;35], and a modified Gillespie algorithm[36]. The Gillespie algorithm generates statistically correct trajectories of continuous-time Markov processes while the modified Gillespie algorithm is an approximate method for simulating non-Markovian processes.

Indeed, the number of possible spreading models that can be defined is limitless because the possible node states and node state transitions are not restricted. However, most net-

[1] Parts of this chapter are extracted and adapted from our published article[33], Copyright © 2017, Elsevier.

worked spreading processes share a common fundamental assumption: nodes influence each other through independent pairwise interactions. *Independent* means that different nodes in the network influence a common neighbor through statistically independent processes. *Pairwise* indicates that no higher order interaction is permitted, i.e., joint interaction of three nodes A–B–C is fully described by A–B, B–C, and A–C interactions.

Based on the independent pairwise interaction characteristic of most spreading models, Sahneh *et al.*[19] defined the *generalized epidemic modeling framework* (GEMF) that incorporates a broad spectrum of stochastic spreading processes over complex networks. In order to make our computational tools applicable to a broad range of spreading models, we chose to implement the Gillespie algorithm and the modified Gillespie algorithm for the *generalized epidemic modeling framework.*

2.1.1 Generalized Epidemic Modeling Framework

GEMF describes a general epidemic model over a network composed of one set of nodes and several layers of contact. We represent the network by $G(V, E_1, \cdots, E_L)$, where L is the number of contact layers, V is a set of N nodes, and E_l is a set of links between the nodes in layer l. The incorporation of multilayer typologies in GEMF makes it a flexible framework for studying epidemic processes.

Similar to the SIS model, state of node n at time t is a random variable denoted by $x_n(t)$ and each node can assume a node state among M possible states, which are labeled with an integer from 1 to M, i.e., $x_n(t) \in \{1, \cdots, M\}$. In GEMF, transitions of x_n over the node states are classified into two categories.

1. Nodal transitions of a node are similar to the curing process in the SIS model and they are independent of the neighbors' state.

2. Edge-based transitions of a node are analogous to the infecting process in the SIS model. These transitions are caused by the interaction of a node with its neighbors in the network, and they depend on states of the neighbors. In GEMF each network layer has its own influencer state. If a node is in an influencer state of a layer it will induce some transitions on its neighbors in that network layer. For instance, the influencer state in the SIS model is the infected state. The network layer provides contacts for a node in the influencer state to induce and propagate certain transitions over neighboring nodes.

2.2 Simulation of Markovian Processes

The Gillespie algorithm is a method for sampling the earliest event among a set of independent events assuming the occurring time for each event is exponentially distributed. To understand the Gillespie algorithm, consider a set of k independent nodes where each node, such as node n, will make a transition from state i to state j at a random time $T_n \sim \exp(r_n)$. This random time T_n has an exponential distribution with rate r_n. In this case, since the transition of a node does not affect the transition of other nodes, we can generate the transition time for each node by drawing a random value from its corresponding distribution. If we arrange these transition times in increasing order we get a sequence of events. The Gillespie algorithm is another method that generates such sequences of events. It starts with all ongoing processes and samples the time for the earliest event and the node that makes the transition. Next, the algorithm advances the time and repeat the same procedure for the remaining processes. Although the Gillespie algorithm can be applied to the case of k independent nodes we described above, it is more applicable in simulating the Markovian dynamics of a complex system where the occurrence of an event can affect other ongoing processes in the system. For instance, consider an infection process of a node by two infected neighbors. In this case the infected neighbors can infect the target node through independent processes. However, when an infection event happens the competing infection process is assumed to be terminated. In order to sample the time for the infection of the target node, we can generate two random times and accept the shortest time as the infection time. When we use the Gillespie algorithm we can generate the infection time directly. Indeed, the Gillespie algorithm relies on the fact that the minimum of exponentially distributed independent random variables has an exponential distribution with a rate equal to the sum of the individual rates[37]. Hence, we only need to generate one random infection time from an exponential distribution whose rate is the sum of the competing infection processes' rates.

2.2.1 Algorithm

Considering the node-level description of transitions in GEMF, the node transition $x_n \to j$ may be viable through different possible processes. In other words, node n may undergo a transition from its current state x_n to any state j by interacting with neighbors or through a nodal transition. In such a case, the processes are assumed to be competing independent processes that try to induce the transition $x_n \to j$. Thus, the actual transition time of node n, $T_{x_n \to j}$, is the minimum of transition times for the competing processes, $T_{x_n \to j} = \min\{T_1, \cdots, T_p\}$. If the transition times in all the independent processes are distributed exponentially, $T_1 \sim \exp(r_1), \cdots, T_k \sim \exp(r_p)$, distribution of the transition time $T_{x_n \to j}$ is exponential with a rate which is sum of all rates for the possible processes, i.e., $T_{x_n \to j} \sim \exp(\lambda_n(x_n \to j))$ where $\lambda_n(x_n \to j) = \sum_p r_p$.

To proceed with the simulation algorithm, we define two arrays:

Nodal transition matrix, A_δ, where the element $A_\delta(i, j)$ is the transition rate of a node from state i to state j via a nodal transition.

Edge based transition array, A_β, where the element $A_\beta(i, j; l)$ is the transition rate for the transition of a target node from state i to j through an interaction with a neighbor in layer l while the neighbor is in state $q(l)$. State $q(l)$ is the influencer state for layer l.

In fact, the elements of A_δ and A_β define the rates for the exponential distribution of transition times corresponding to the possible processes allowed in GEMF. If a rate is zero the corresponding transition never happens.

Assuming the joint state of the network at time t is $X(t) = [x_1, \cdots, x_N]$, we can calculate all node-level transition rates $\lambda_n(x_n \to j)$, for any node n, using the nodal transition matrix A_δ, edge-based transition array A_β and the contact network $G(V, E_1, \cdots, E_L)$, where $\lambda_n(x_n \to j)$ is the transition rate of node n from its current state x_n to the state j.

However, the occurrence of any node-level transition can affect other ongoing processes in the network. Hence, we will follow the Gillespie algorithm and sample the earliest transition.

If we define $\mathcal{S}$ as the set of all node-level transition times

$$\mathcal{S} = \{T_n(x_n \to j) | n \in \{1, \cdots, N\}, j \in \{1, \cdots, M\}\},$$

the elements of $\mathcal{S}$ are independent exponentially distributed random variables. Using the theorem concerning the minimum of independent exponential distributions[37], then the probability that $T_n(x_n \to j)$ would be the minimum of $\mathcal{S}$ is

$$\Pr\left(T_n\left(x_n \to j\right) = \min(S)\right) = \frac{\lambda_n(x_n \to j)}{\lambda_{tot}},$$

where $\lambda_{tot} \triangleq \sum_n \sum_j \lambda_n(x_n \to j)$. Using this probability distribution, we can sample one of the node-level transitions. We also must sample the time at which this transition occurs. Because elements of $\mathcal{S}$ have exponential distributions, if we define $T = \min(\mathcal{S})$, then T is exponentially distributed with a rate equal to λ_{tot}. Thus, using the distribution of T, we sample a time for the network state transition. The memoryless property of Markov processes allows the entire described procedure to be repeated after the network state is updated. Particularly, we can directly update the transition rates by the adjustment required due to the change in the state of node n that made the transition, including updating the transition rates of node n and neighbors that can be affected by node n. The other rates remain constant.

The described simulation method is summarized in Algorithm 1, where we assume that network links can be directed and weighted. If a link is directed from node m to node n, node m can induce edge-based transitions on node n, but node n cannot induce edge-based transitions on node m. Moreover, we can assign a weight to each link in order to quantify the effect of neighbors on edge-based transitions of a node. The rate of an edge-based transition induced by a link is multiplied by the weight of the link. In Algorithm 1, $W(m, n; l)$ is the weight of the link directed from node m to node n in layer l of the network, and $W(m, n; l) = 0$ indicates that such a link does not exist. However, implementation of Algorithm 1 requires to only store the nonzero weights. In Algorithm 1, input $q(l)$ is the influencer node state for

Algorithm 1 GEMFsim algorithm

```
Input $A_\delta$, $A_\beta$, $W$, $X_0$, $q$, $Stop\ condition$
Output: $event$
$X \leftarrow X_0$
for $n = 1$ to $N$ do
    for $l = 1$ to $L$ do
        $wq(n,l) \leftarrow \sum_{m=1}^{N} W(m,n;l)\delta_{x_m,q(l)}$
    end for
    $\lambda_n \leftarrow \sum_{j=1}^{M} A_\delta(x_n,j) + \sum_{l=1}^{L} wq(n,l)A_\beta(x_n,j;l)$
end for
$\lambda_{tot} \leftarrow \sum_{n=1}^{N} \lambda_n$
$k = 0$
while $Stop\ condition$=FALSE do
    $\alpha \sim \mathrm{Unif}(0,1)$                                   ▷ generate $\alpha$ from Unif$(0,1)$
    $\delta t_k \leftarrow -\log(\alpha)/\lambda_{tot}$                 ▷ time period to the next event
    $P_1(n) \leftarrow \lambda_n/\lambda_{tot}$
    $n_k \sim P_1$                              ▷ sample $n_k$ from probability distribution $P_1$
    $i_k \leftarrow x_{n_k}$
    for $j = 1$ to $M$ do
        $\lambda_{n_k}(i_k \to j) \leftarrow A_\delta(i_k,j) + \sum_{l=1}^{L} wq(n_k,l)A_\beta(i_k,j;l)$
    end for
    $P_2(j) \leftarrow \lambda_{n_k}(i_k \to j)/\lambda_{n_k}$
    $f_k \sim P_2$                                  ▷ sample $f_k$ from distribution $P_2$
    $event(k) \leftarrow (\delta t_k, n_k, f_k, i_k)$
    $x_{n_k} \leftarrow f_k$                                         ▷ update network state
    for $l \mid (q(l) = f_k$ or $q(l) = i_k)$ do                        ▷ Update Rates
        $\Delta \leftarrow \delta_{q(l),f_k} - \delta_{q(l),i_k}$
        for $n \mid W(n_k,n;l) \neq 0$ do
            $wq(n,l) \leftarrow wq(n,l) + \Delta \times W(n_k,n;l)$
            $\lambda_n \leftarrow \lambda_n + \Delta \times \sum_{j=1}^{M} W(n_k,n;l)A_\beta(x_n,j;l)$
        end for
    end for
    $\lambda_{n_k} \leftarrow \sum_{j=1}^{M} A_\delta(f_k,j) + \sum_{l=1}^{L} wq(n_k,l)A_\beta(f_k,j;l)$
    $\lambda_{tot} \leftarrow \sum_{n=1}^{N} \lambda_n$
    $Update\ Stop\ condition$
    $k \leftarrow k+1$
end while
```

layer l, A_δ and A_β are nodal transition rates and edge-based transition rates, respectively, and X_0 is the initial network state. Assuming the current network state $X = [x_1, \cdots, x_n]$, node-level transition rates are calculated as

$$\begin{aligned}\lambda_n(x_n \to j) = {} & A_\delta(x_n, j) \\ & + \sum_{l=1}^{L} A_\beta(x_n, j; l) \sum_{m=1}^{N} W(m, n; l)\delta_{x_m, q(l)},\end{aligned}$$

where $\delta_{s,t}$ is Kronecker delta. In Algorithm 1, we sample a node-level transition in three steps. First, we generate one sample for a random variable δ_t which is exponentially distributed with the rate λ_{tot}. In fact, δ_t is the time period between the network state events, and $\lambda_{tot} = \sum_{n=1}^{N} \lambda_n$, where $\lambda_n = \sum_{j=1}^{M} \lambda_n(x_n \to j)$. Generating a sample for δ_t is done by generating a sample α from the uniform distribution over the interval $(0, 1)$, and inserting α into the equation $\delta_t = -\log(\alpha)/\lambda_{tot}$. The second step is to select a node according to the probability distribution $\Pr(n) = \lambda_n/\lambda_{tot}$. This is the node that will make the transition. After the node is picked, we select a new node state j for the node according to the probability distribution $\Pr(j \mid n) = \lambda_n(x_n \to j)/\lambda_n$. The event-based algorithm explained above is an adaptation of the Gillespie algorithm[34;35], to GEMF-based processes. Implementation of the algorithm is available online[38].

2.2.2 Simulations

A broad spectrum of epidemic models can be formulated in the GEMF framework. Hence, the algorithm we described above is a flexible platform capable of simulating various stochastic spreading models.Here, we show how GEMFsim (implementation of Algorithm 1) can be applied to study various compartment models that fit the description of GEMF processes. GEMFsim provides realizations of Markov processes over a space consisting of network states. In theory, GEMFsim can be used to generate enough samples to extract statistics of interest. In fact, any statistics defined in terms of marginal distributions of the Markov processes can be estimated using samples generated by the GEMFsim tool.

Comparison to Exact Kolmogorov Equations

The sampled network state trajectories generated using algorithm 1 follow a distribution which is the solution of the Kolmogorov equations for the Markov process governing the network state evolution. To experimentally test the distribution of the generated samples, we compared results of the simulation with the exact solution of the Kolmogorov equations for the SIS model. The Kolmogorov equations for the SIS process over a network of N nodes is a linear system of 2^N coupled equations and the size of linear system becomes gigantic, even for moderate values of N. Therefore, we considered a small network of $N = 10$ nodes with SIS parameters of $\delta = 1$ and $\beta = 2$. Assuming an initial condition in which only one node was infected, we solved the Kolmogorov equations, $\dot{P}(t) = -Q^T P(t)$[19], where P is a probability distribution over a space consisting of $2^{10} = 1,024$ network states and Q is the infinitesimal generator matrix. We then extracted the infection probability of each node, $p_i^{exact}(t)$, as a marginal distribution of $P(t)$. Next, using Algorithm 1, we generated n realizations of the SIS process and obtained an estimation for the infection probability of node i, $\hat{p}_i^{[n]}(t)$, as the fraction of realizations that node i was infected at time t. Our objective was to observe if the difference between $\hat{p}_i^{[n]}(t)$ and $p_i^{exact}(t)$ decreases as the number of realization n increases. Therefore, we defined two measures of error as

$$\textit{Total Error}^{[n]} \triangleq \max_i \max_t |\hat{p}_i^{[n]}(t) - p_i^{exact}(t)|, \tag{2.1}$$

$$\textit{Mean Error}^{[n]} \triangleq \max_t \frac{1}{N} |\sum_{i=1}^{N} \hat{p}_i^{[n]}(t) - p_i^{exact}(t)|. \tag{2.2}$$

Fig. (2.1a) shows how the defined measures decreased when the number of realization n increased.

Simulation of the SIS Model

The SIS model is one of the simplest models that can be simulated using GEMFsim. In this model each node is either susceptible (S) or infected (I), as represented by the integers 1 or 2, respectively. If a node is infected, it transmits infection to the susceptible neighbors at a rate

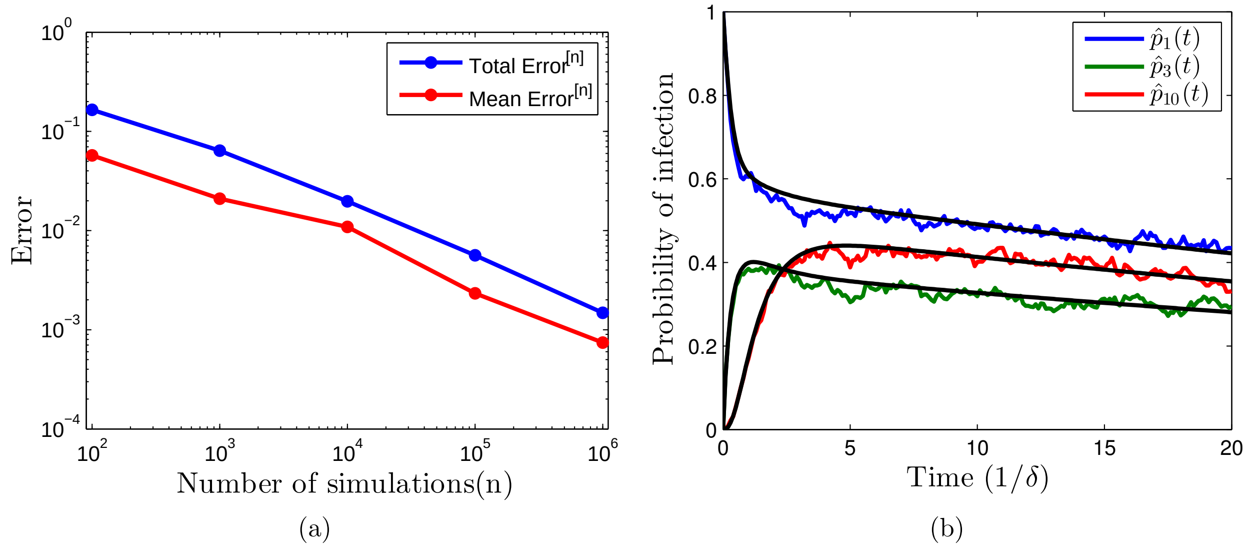
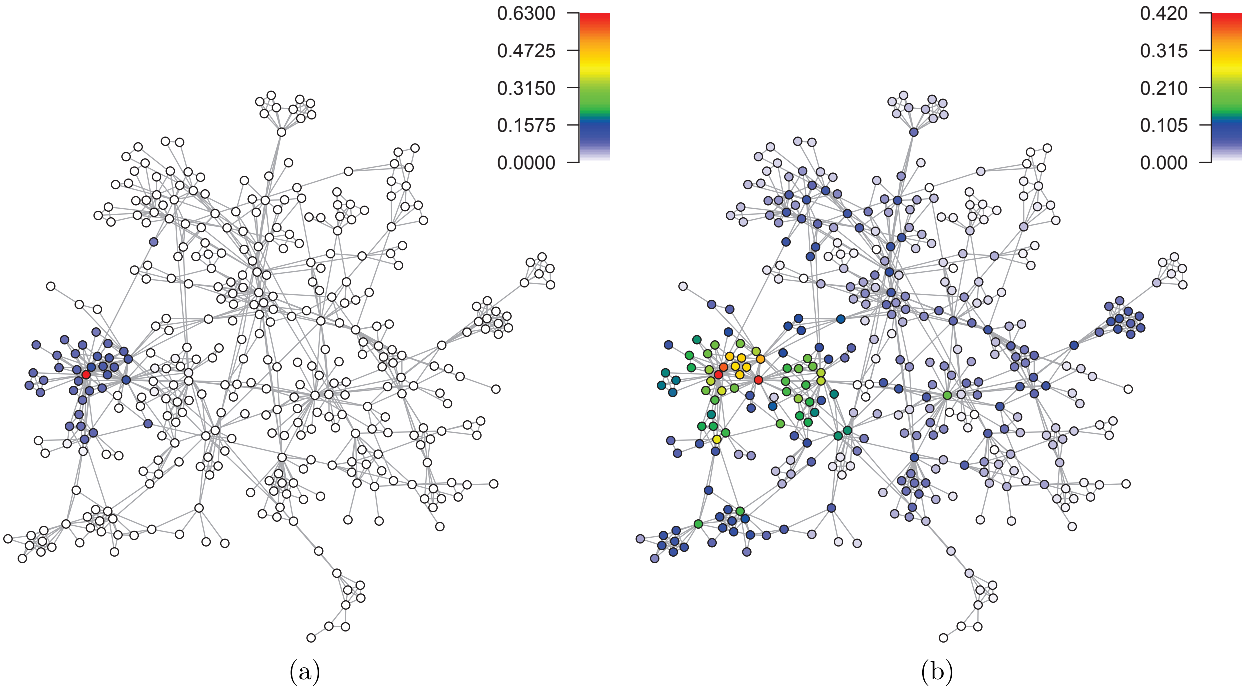


Figure 2.1: *The infection probability for each node in a toy network of ten nodes estimated using simulation in comparison to the exact probability obtained by solving the Kolmogorov equations for the SIS model: (a) total error and mean error defined in Eqs. (2.1), (2.2), (b) estimation of infection probability for some nodes obtained by averaging over 1000 simulations. The black (smooth) curves are exact probabilities obtained by solving the Kolmogorov equations.*

β, and the infected node recovers with the rate δ. We simulated SIS spreading over a contact network consisting of one layer of contact $G(V, E)$. The network we used was the largest component of the coauthorship network presented in[39]. We assumed that the links were undirected and had identical weight. Based on description of the nodal transition matrix, A_δ, and the edge-based transition array, A_β, the nonzero elements of them in the SIS model are $A_\delta(2, 1) = \delta$ and $A_\beta(1, 2; 1) = \beta$. Moreover, the influencer node state for this model is the infected state, i.e., $q(1) = 2$. Using the implementation of GEMFsim algorithm in R[38] we generated 8000 realizations of SIS spreading. We used the results of these simulations to estimate the probability of being infected for each node in the network at various time points. The probability of being infected was estimated as the fraction of SIS realizations in which the node was infected at the given time point. Results for two time points are plotted in Fig. (2.2). We assumed $\beta = 0.23$ and $\delta = 1$. The only node that was initially infected in all realizations was the node with the highest degree.



(a) (b)

Figure 2.2: *Result form simulation of SIS spreading over a network. Color of each node represents probability of being infected for the node. (a) probability of being infected at time point $t = 0.5$ $(1/\delta)$, (b) probability at time point $t = 90$ $(1/\delta)$. At $t = 0$, only the node with the highest degree was infected. These graphs show evolution of infection in the network*

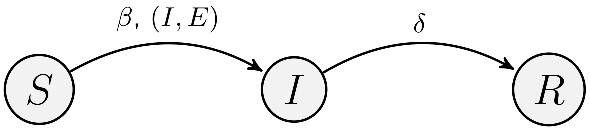


Figure 2.3: *Schematic of node-level transitions in the SIR model*

Simulation of SIR Model

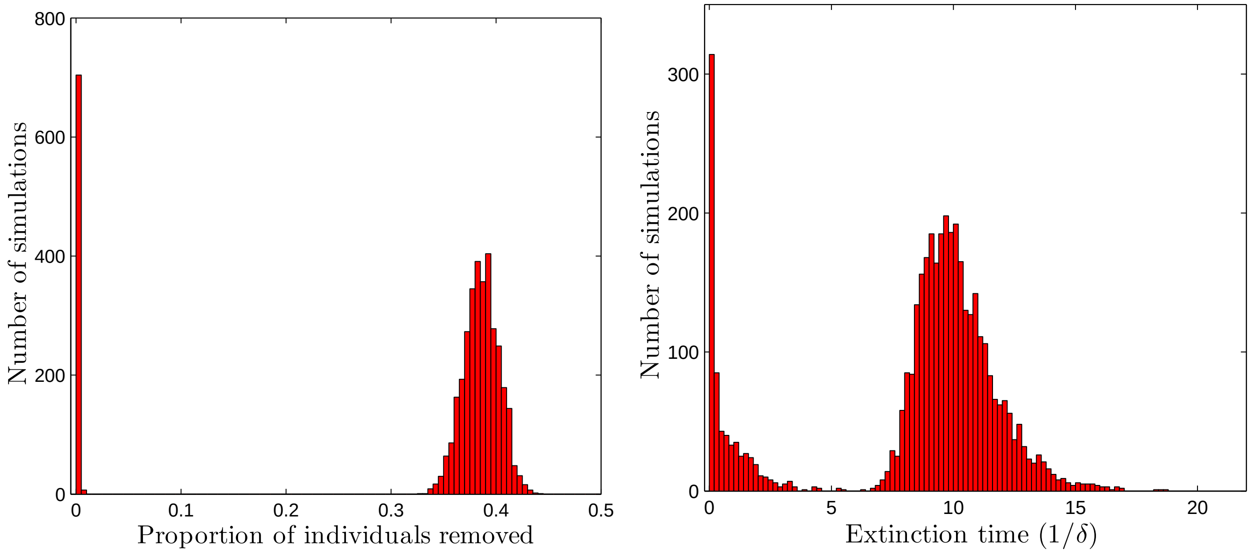
Here we show how GEMFsim can be used to estimate certain statistics which are beyond the scope of mean-field-type approximations. In fact, GEMFsim can be used to generate several realizations of a spreading process and estimate probability distribution for the epidemic measure of interest. We considered an SIR epidemic model in which a susceptible node becomes infected with the rate β as a consequence of interacting with infected neighbors. Moreover, an infected individual makes a transition to a removed state that may represent the recovered immune state. This transition occurs independently of state of neighbors;

in Fig.(2.3), the transition rate is shown by δ. In the SIR model, a removed node does not affect its neighbors or undergo any transition, and the network eventually reaches an absorbing state in which all individuals are susceptible or removed. Although the time at which the network falls into the absorbing state is not a deterministic variable, the simulation can be used to estimate the probability distribution for the extinction time. The final number of removed individuals is an important measure in epidemiology because it shows the size of outbreak. Similar to extinction time, we can use simulation to estimate probability distribution of the total number of individuals removed.

We used GEMFsim in MATLAB to generate 4000 realizations of SIR spreading over a directed and weighted network composed of 1899 nodes and 20296 edges. This network has been studied in reference[40], and its dataset is available online with the name of Facebook-like social network[40]. We assumed initially only one node, which is labeled by integer 1 in the dataset, was infected and the rest of nodes in the network were susceptible. We used transition rates $\beta = 0.05$ and $\delta = 1$ for the simulation. Node states in the SIR model are susceptible, infected and removed as labeled by the integers 1, 2, and 3, respectively. The network has one layer of contact with a set of directed and weighted links, and the influencer state is the infected state, represented by integer 2, i.e., $q(1) = 2$. The only nonzero elements of the nodal transition matrix and the edge based transition array are $A_\delta(2,3) = \delta$ and $A_\beta(1,2;1) = \beta$. Using simulation we were able to generate a histogram of the extinction time and the total fraction of removed individuals in the defined SIR spreading. Fig. (2.4) shows the total number of affected individuals and extinction time as they follow bimodal distributions.

Simulation of SAIS Model

The Susceptible-Alert-Infected-Susceptible (SAIS) model was developed to incorporate individual reactions to the spread of a virus[41;42]. In the SAIS spreading model, each node (individual) is either susceptible (S), infected (I), or susceptible-alert (A). A susceptible node gets infected with a rate β through interaction with an infected node, and an infected



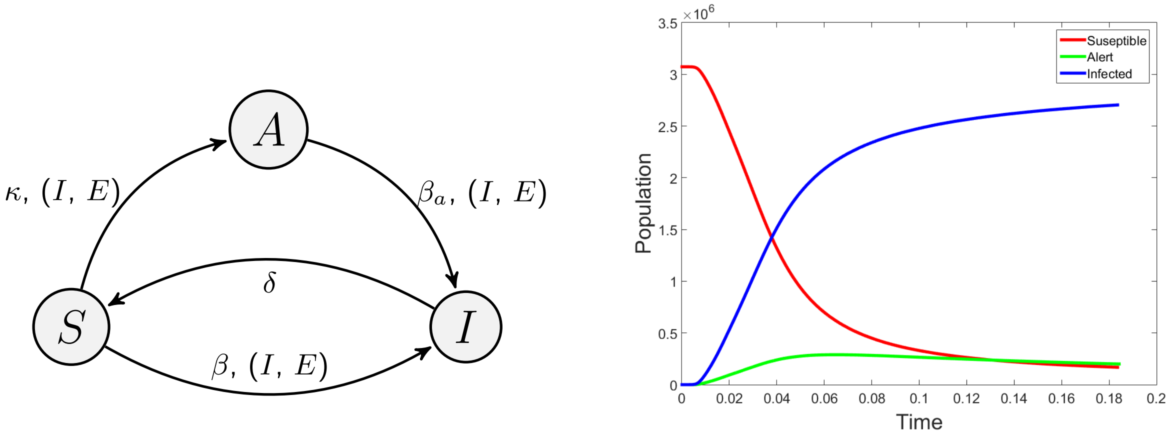
(a)

(b)

Figure 2.4: *Results from 4000 realizations of SIR spreading over a network: (a) histogram of the fraction of removed individuals (b) histogram of extinction time defined as the time when the last infected node in the network is removed*

node recovers with a rate δ. The SAIS model also accounts for another possibility that a susceptible node can become alert with a rate κ if it senses an infected node in its neighborhood. An alert node can also become infected by a process similar to the infection process of a susceptible node. However, the infection rate for an alert node, denoted by β_a, is lower due to the adoption of preventative behaviors. In order to simulate a realization of the SAIS process, we set up a problem according to the GEMF framework in which three node states (S, A, I) were denoted by integers $1, 2, 3$, respectively. The network had one layer of contact, $G(V, E)$, where E represents a set of links that could be generally directed and weighted. The influencer state in this model was infected state as represented by integer 3, i.e., $q(1) = 3$. The only nonzero element of the nodal transition matrix in the SAIS model is $A_\delta(3, 1) = \delta$. The nonzero elements of the edge-based transition array are $A_\beta(1, 3; 1) = \beta$, $A_\beta(2, 3; 1) = \beta_a$, and $A_\beta(1, 2; 1) = \kappa$. Schematic of node-level transitions in the SAIS model is shown in Fig. (2.5a)

Using the implementation of GEMFsim algorithm in C language[38] we generated one realization of the SAIS model over a network[43] of 3,072,441 nodes that were connected



(a) (b)

Figure 2.5: *(a) Schematic of node-level transitions in the SAIS model (b) Simulation of SAIS spreading over a large-scale network. Plot represents the population of each node state in the network over time.*

through 11,7185,083 links. Network links were undirected and had identical weights. The simulation result is shown in Fig. (2.5b). The simulation initially began with 20 infected nodes and 20 nodes in the alert state; the other nodes in the network were initially susceptible. Transition rates for the simulation were $\delta = 1$, $\beta = 2$, $\beta_a = 0.4$, and $\kappa = 0.2$.

Simulation of SI$_1$SI$_2$S Model

The SI$_1$SI$_2$S model is an extension of the SIS model in which two types of infection can attack a susceptible node[29]. However, we assumed a competitive scenario in which the two viruses were exclusive, or a node did not harbor both types of infection simultaneously. Therefore, in this model, each node is either susceptible (S), infected by virus one (I$_1$), or infected by virus two (I$_2$). Similar to the SIS model, infected nodes recover with a rate δ_1 or δ_2 depending on the infection. In general, different infections can be transmitted to a susceptible node through different contacts. In order to account for different means of spreading for I_1 and I_2, the assumption was made that they spread through different layers of contact such that a susceptible node undergoes a transition to infected state I_1 (I_2) with a rate β_1 (β_2) if it is in contact with an I_1 (I_2) node through layer E_1 (E_2). Fig. (2.6) depicts the SI$_1$SI$_2$S model of spreading. The SI$_1$SI$_2$S model can be described in the GEMF framework, by three node

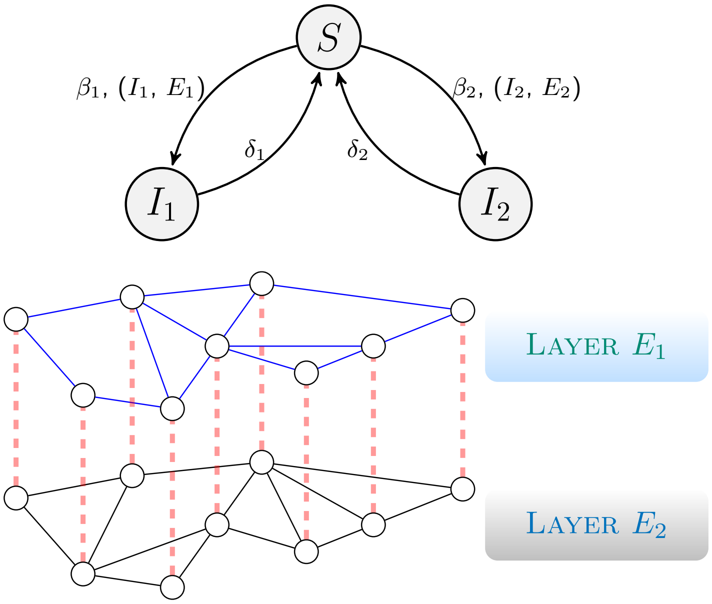


Figure 2.6: *Node-level transitions in the SI_1SI_2S spreading model over a two-layer network. Layers E_1 and E_2 define two types of contact over the same set of nodes.*

states (S, I_1, I_2) represented by integers $1, 2, 3$, respectively. The network consists of two layers, $G(V, E_1, E_2)$, where the first layer spreads I_1, and the second layer, I_2. The influencer node state for layer one is I_1 and the influencer node state for the second layer is I_2. The only nonzero elements of nodal transition matrix are $A_\delta(2,1) = \delta_1$ and $A_\delta(3,1) = \delta_2$. Nonzero elements of edge-based transition array are $A_\beta(1,2;1) = \beta_1$ and $A_\beta(1,3;2) = \beta_2$. In general, E_1, E_2 could be two different sets of links between nodes. However, if both types of infection use the same kind of contacts to spread, E_1 and E_2 are similar.

The SI_1SI_2S model described above, exemplifies a competitive spreading scenario in which two types of infection try to invade a network. However, mean-field-type approximation shows, in a network with two different layers of contact, I_1 and I_2 can coexist depending on their infection rates[29]. We used the implementation of GEMFsim in Python[38] to show this coexistence via simulation. We adopted a network of 500 nodes with two different contact layers, E_1 and E_2. We assumed I_1 spreads through contact layer E_1, which is a scale-free network[44] of 2,475 edges and that I_2 uses a geometric network, E_2, of 3,560 edges to invade the nodes. Assuming B (A) is the adjacency matrix for contact layer E_2 (E_1), we used value

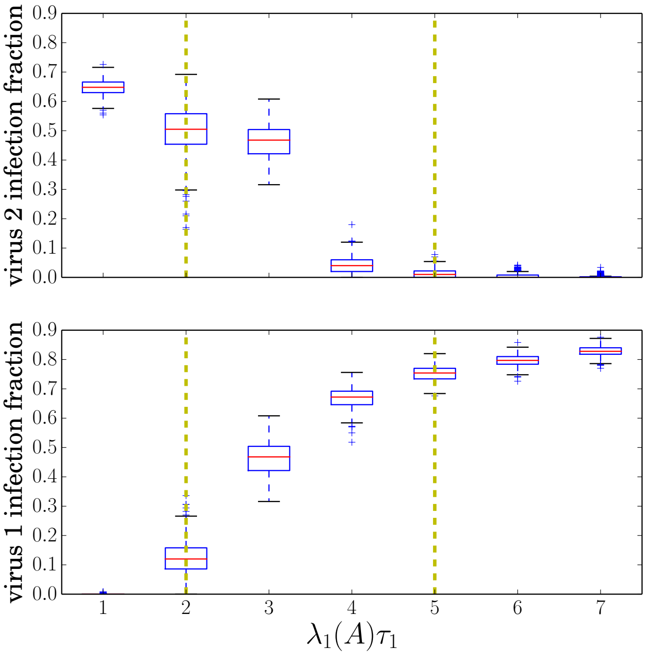


Figure 2.7: *Fraction of nodes infected by virus type* 2 *(above) and virus type* 1 *(below) in the* SI_1SI_2S *competitive spreading model. The infection strength of* I_2, τ_2, *was* $5/\lambda_1(B)$, *while the infection strength of* I_1, τ_1, *varied. If* $2/\lambda_1(A) \leq \tau_1 \leq 5/\lambda_1(A)$, *viruses coexist; only one virus survives outside this region.*

of $5/\lambda_1(B)$ as the infection strength $\tau_2 = \beta_2/\delta_2$ where $\lambda_1(B)$ is the largest eigenvalue of adjacency matrix B. However, for infection strength $\tau_1 = \beta_1/\delta_1$ we used seven values from $\tau_1 = 1/\lambda_1(A)$ to $\tau_1 = 7/\lambda_1(A)$; for each value of τ_1 we generated 500 realizations of SI_1SI_2S processes. For all simulations we assumed that each virus had initially infected 2% of the nodes. Fig. (2.7) shows metastable state population sizes extracted from simulations for values of τ_1. As shown in the figure, either one of the viruses prevails or the viruses coexist depending on the value of τ_1.

2.3 Simulation of non-Markovian Stochastic Processes

In this section, we discuss a general method that can be used for simulation of independent non-Markovian processes[36]. Since this method uses the concept of the earliest event, it can

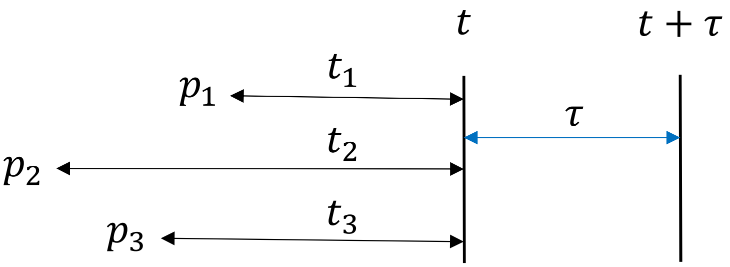


Figure 2.8: *Three different processes, p_1, p_2, p_3, have been initiated but they have not occurred up to time t. We want to calculate the probability densities that they occur later at some time denoted by $t + \tau$.*

be considered a generalized Gillespie algorithm.

Consider a set of N statistically independent processes, each with an inter-event time distribution $\psi_i(\tau)$; $i \in \{1, \cdots, N\}$. Suppose that, for a given process i, we know t_i, which is the time interval between the process initiation and the latest observation of the process performed at the current time t (see figure 2.8 for illustration). First, we want to know what is the probability density that the event for process i will happen in the time interval $[\tau, \tau + d\tau]$ where τ is measured from current moment t. This probability density can be written as

$$\psi_i(\tau|t_i) = \frac{\psi_i(\tau + t_i)}{\Psi_i(t_i)},$$

where $\Psi_i(t_i) = \int_{t_i}^{\infty} \psi_i(s)ds$ is the survival function of process i and gives the probability that the event for process i does not happen in the time interval $[0, t_i]$ after the process initiation. Indeed, ψ_i is the truncated distribution for the inter-event time and reflects our observation that the process has survived up to t_i from its initiation.

To generate a statistically correct sequence of events for the set of all ongoing processes, in the next step, we should calculate the probability density that the next event corresponds to the process i among all the processes, and it will occur at time $\tau + t_i$. This probability is

$$\phi(\tau, i|\{t_k\}) = \psi_i(\tau|t_i) \prod_{k \neq i} \Psi_k(\tau|t_k) = \frac{\psi_i(\tau + t_i)}{\Psi_i(\tau + t_i)} \prod_{k=1}^{N} \Psi_k(\tau|t_k), \tag{2.3}$$

where $\Psi_i(\tau|t_i)$ is the conditional survival probability,

$$\Psi_i(\tau|t_i) = \int_\tau^\infty \psi_i(s|t_i)ds = \frac{\Psi_i(\tau + t_i)}{\Psi_i(t_i)},$$

which is the probability that the event for the process i occurs after $t+\tau$, assuming we know it did not happened before t. To sample a time for the occurrence of next event, we evaluate the joint conditional survival function, which is the probability that no event happens before $t + \tau$ and it can be written as

$$\Phi(\tau|\{t_k\}) = \prod_{k=1}^{N} \Psi_k(\tau|t_k) = \prod_{k=1}^{N} \frac{\Psi_k(\tau + t_k)}{\Psi_k(t_k)}. \tag{2.4}$$

Now we can express the algorithm for the generation of a statistically correct sequence of events as follows:

1. draw u a random number in the interval $(0, 1)$ and find the random time for the next event τ by solving $u = \Phi(\tau|\{t_k\})$.

2. choose a process that corresponds to the next event by sampling the following discrete distribution
$$p(i) = \frac{\phi(\tau, i|\{t_k\})}{\sum_j \phi(\tau, j|\{t_k\})} = \frac{\lambda_i(\tau + t_i)}{\sum_j \lambda_j(\tau + t_j)},$$
where $\lambda_i(s) \equiv \psi_i(s)/\Psi_i(s)$ is the instantaneous rate of the process.

3. update the elapsed time for the processes
$$t_k \to t_k + \tau, \forall k \neq i \quad \text{and} \quad t_i = 0;$$

4. update the processes. Terminate some processes or activate new processes if needed. Go to step 1.

To understand the connection between the algorithm above and the original Gillespie algorithm, assume that all the processes in the algorithm are Markovian, i.e., the transition times are exponentially distributed, $\psi_k(t) = r_k \exp(-r_k t)$ for $k \in \{1, \cdots, N\}$. With this assumption, the survival function for any process k becomes $\exp(-r_k t)$. Therefore, the joint conditional survival function in the equation 2.4 can be written as

$$\Phi(\tau|\{t_k\}) = \prod_{k=1}^{N} \Psi_k(\tau|t_k) = \prod_{k=1}^{N} \exp(-r_k \tau) = \exp(-\tau \sum_k r_k). \tag{2.5}$$

Using this equation, we can see the distribution of the earliest transition time, τ, is exponential with the rate equal $\sum_k r_k$. This is the result of the Markovian nature of exponential distributions, which make the conditional survival functions, $\Psi_k(\tau|t_k)$, independent of the age of processes, t_k. Hence, step one in the algorithm for the simulation of non-Markovian processes is the generalization of the step in the Gillespie algorithm where we sample the earliest event for a set of independent Markovian processes. Moreover, the instantaneous rates defined in step 2 of the non-Markovian algorithm reduce to the rates of exponential distributions, $\lambda_i(s) = r_i$, which are constant. Therefore, the non-Markovian algorithm follows the same logic as the Gillespie algorithm except that the processes have variable ages, and the rates are not constant.

Although the non-Markovian algorithm is statistically exact, it can be computationally expensive. In step one of the algorithm, we need to solve the equation $u = \Phi(\tau|\{t_k\})$, to find the time to the next event, τ. This step can be computationally expensive. However, if we use the following approximation

$$\Psi_k(\tau + t_k) \approx \Psi_k(t_k) + \tau\Psi'_k(t_k) = \Psi_k(t_k) - \tau\psi_k(t_k) = \Psi_k(t_k)(1 - \tau\lambda_k(t_k)), \tag{2.6}$$

where we have assumed $\tau \sim 0$, the joint conditional distribution can be approximated as

$$\Phi(\tau|\{t_k\}) \approx \prod_{k=1}^{N}(1 - \tau\lambda_k(t_k)) \approx 1 - \tau\sum_{k=1}^{N} \lambda_k(t_k). \tag{2.7}$$

Now solving $u = \Phi(\tau|\{t_k\})$ for τ, in the first step of the algorithm, is a simple task and that leads to

$$\tau = -\frac{\ln(u)}{\sum_{k=1}^{N} \lambda_k(t_k)} \tag{2.8}$$

for the earliest event time. As it is explained in the reference[36], when the number of processes, N, is large the total rate, $\sum_{k=1}^{N} \lambda_k(t_k)$, is large and with a high probability, τ is small. Hence, the approximation is valid. However, in some cases it is possible that the total rate is not always large through the course of the simulation, for instance at the start of an epidemic. In order to always keep τ small, we can include an independent auxiliary Markovian process in the set of active processes such that it has a large constant rate of λ_0. Adding this process to the pool of other active processes changes the denominator on r.h.s of equation 2.8 to the larger rate of $\lambda_0 + \sum_{k=1}^{N} \lambda_k(t_k)$. Therefore, we always obtain a small value for τ.

To understand this approximated algorithm, assume a case where an independent node repeatedly is transitioning between two states while the transition times have a truncated normal distribution,

$$\psi(t) = \frac{2}{1 + erf(\frac{\mu}{\sigma\sqrt{2}})} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} \qquad \text{for } t \geq 0. \tag{2.9}$$

In this distribution, erf stands for the error function and μ, σ are the mean and standard deviation for the normal distribution before truncation. In figures 2.9a and 2.9b, we have plotted the distribution of equation 2.9 and its instantaneous rates for two different sets of σ and μ. The instantaneous rates are calculated as $\lambda(t) = \psi(t)/\int_t^{\infty} \psi(s)\ ds$. From figure 2.9b we can see the initial rate for the distribution with $\mu = 5$ is close to zero. If we try to find the transition time by plugging the initial rate in equation 2.8, with a high probability we will get a very large τ, which is not acceptable. Therefore, we add another independent node that makes similar transitions except that the transition times are exponentially distributed with the rate $\lambda_0 = 20$. Now, if we consider both nodes together and sample the earliest transition time using equation 2.8 and the total rate of $\lambda_0 + \lambda(t)$, the average of transition time ,τ, is smaller than $1/\lambda_0 = 0.05$. Moreover, the probability that $\tau > 0.25$ is smaller than $e^{-5} = 0.0068$. After sampling τ, we can follow the remaining steps

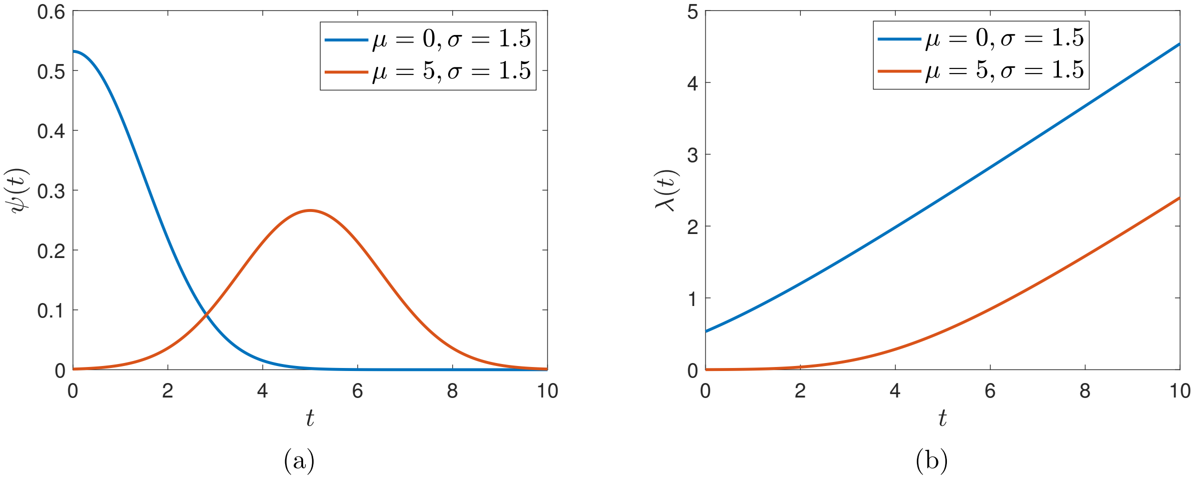


Figure 2.9: *Panel (a) shows distribution in the equation 2.9 for two different sets of the distribution parameters, and panel(b)shows the corresponding instantaneous rates*

in the non-Markovain algorithm of page 25, and sample the node that makes the transition, which can be either one of the two independent nodes. If we sample the auxiliary node, in that iteration of the algorithm, we only advance time and update the instantaneous rate for our initial target node. Figure 2.10 shows the distribution of transition times for the target node in the two simulations performed using the approximated non-Markovian algorithm. In the first simulation, we assumed the distribution of transition time is truncated normal with $\mu = 0,\ \sigma = 1.5$, and for the second simulation we used $\mu = 5,\ \sigma = 1.5$. We can see in both cases the density of transition time obtained from the simulation follows the theoretical curve.

2.3.1 Algorithm

In this section we discuss an implementation of the approximated version of the non-Markovian algorithm on page 25 for GEMF processes. As we described in section 2.1.1, GEMF is a general framework to model spreading processes. In this framework each node, depending on its current state, can have several active nodal processes, independent from the state of the other nodes in the network. Each active process represents a possible transition from the current to other possible node states. In addition to the nodal processes, each

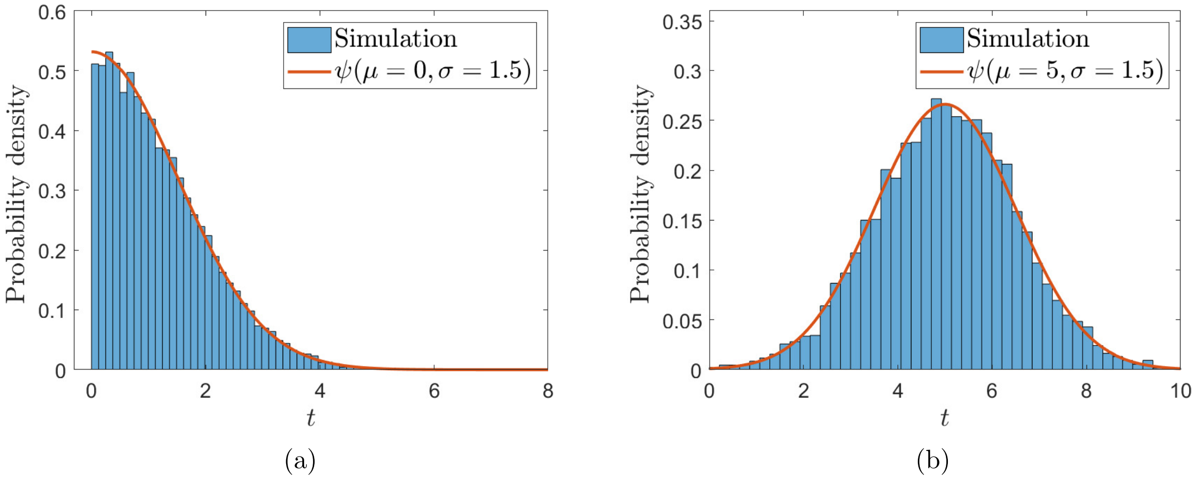


Figure 2.10: *panels (a) and (b) compare the distribution of transition time in the simulation performed using the approximate non-Markovian algorithm with their exact truncated normal distribution.*

directed link in the multilayer network can activate several processes, where each process represents a possible transition for the node at the head of the link. These processes depend on the states of the nodes at the head and the tail of the links. In GEMF, each layer represents a set of links that have the same influencer states. Figure 2.11 shows possible processes in an instance of a spreading process. Each node can be in one of three states. In the figure, the blue node is in state 2 and can go to the states 1 or 3 through the nodal processes $p1$ and p_2, respectively. Moreover, the blue nodes can go to state 1 through an edge based transition, p_3, induced by the red node. Red node is in state 1 and can go to state 3 through the nodal process p_4. In this example, the influencer state for the link is state 1, which is the reason that the node at the tail of the link, the red node, can induce some edge based transitions on the blue node. If the blue node was not in the influencer state for the link, it could not use the link to induce the edge based transitions.

The main difference between the approximate non-Markovian algorithm and the Gillespie algorithm is that the distribution of transition times for the non-Markovian processes in the system are not exponential, so their rates change with the age of the processes. Therefore, we need to keep track of the age of each non-Markovian process, and update the age and the instantaneous rate after each event accordingly.

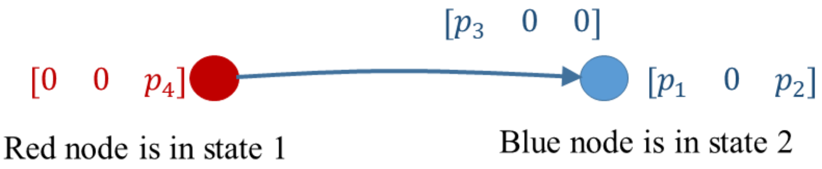


Figure 2.11: *An example of non-Markovian networked spreading discussed in section 2.3.1*

Now we can express the approximated algorithm for the generation of a network state trajectory with non-exponential distributions of the transition times as follows:

1. Identify all the possible active processes in the network. This depends on the state of the nodes and the network links. Initialize the age for all the non-exponential transition times and calculate the instantaneous rates accordingly. Choose a large value for λ_0, which determines the preferred maximum advance in time. We denote the set of all active processes in the network by P.

2. Draw u, a random number in the interval $(0, 1)$ and find the random time for the next event τ as

$$\tau = -\frac{\ln(u)}{\lambda_0 + \sum_{p \in P} \lambda_p(t_p)}.$$

   In the equation above the summation is over all the active processes in the network and t_p is the process age. If some of the active processes in the network are Markovian their rates are constant and do not depend on the age, $\lambda_p(t_p) = \lambda_p$.

3. Choose a process i, that corresponds to the next event by sampling the following discrete distribution

$$\Pr(i) = \frac{\lambda_i}{\sum_{j \in \{0\} \cup P} \lambda_j}, \quad i \in \{0\} \cup P,$$

   where λ_j is the instantaneous rate of the non-Markovian processes process.

4. Advance time by τ and update the age of non-Markovian processes in P accordingly,

$$t_p \to t_p + \tau,\ \forall p \in P - \{i\}.$$

5. Terminate process i if $i \neq 0$, and activate new processes if needed. Go to step 2.

We have implemented the algorithm above in MATLAB and the related code is available online[38].

2.3.2 Simulations

To see the application of the non-Markovian algorithm in studying spreading processes we performed simulations of an SEIR spreading model in a network of 60000 individuals. For the network, we assumed each node is connected to 60 other nodes randomly. In the SEIR model, each node can be found in one the four states: susceptible (S), exposed (E), infectious (I), or removed (R). In this model if a node is susceptible, it becomes exposed (infected) via contact with an infectious neighbor in the network. We assume the infecting process is a Markovian process with a constant rate. When a node becomes exposed, after going through the incubation period it becomes infectious and start infecting other neighbors. However, an infectious node does not stay infectious for a long time. At some random time the infectious node is removed from the network. In general the incubation period and the infectious period can be non-exponential. For instance, recent papers[1;45] analyzing data from the Wuhan outbreak of COVID-19, highlighted non-exponential distributions for some critical transition times, such as the infectious period (figure 2.12.a) and the incubation period.

To study the effect of the distribution for the incubation and infectious periods on the outcome of a spreading model for COVID-19 in Wuhan, China, we performed two sets of simulations. For the first set, we used the lognormal distribution of mean 5.5 days and median 5 days for the incubation period. For the infectious period, we used a lognormal distribution of mean 5.5 days and median 5 days, up to day number 57, and after day 57, we used a lognormal distribution of mean 4 days and median 1.5 days. We changed the

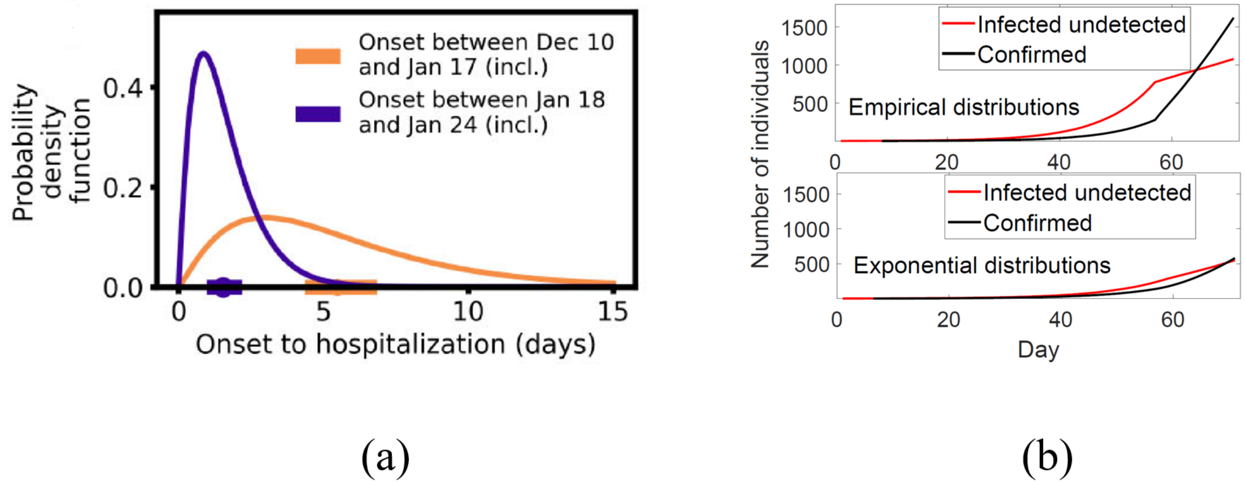
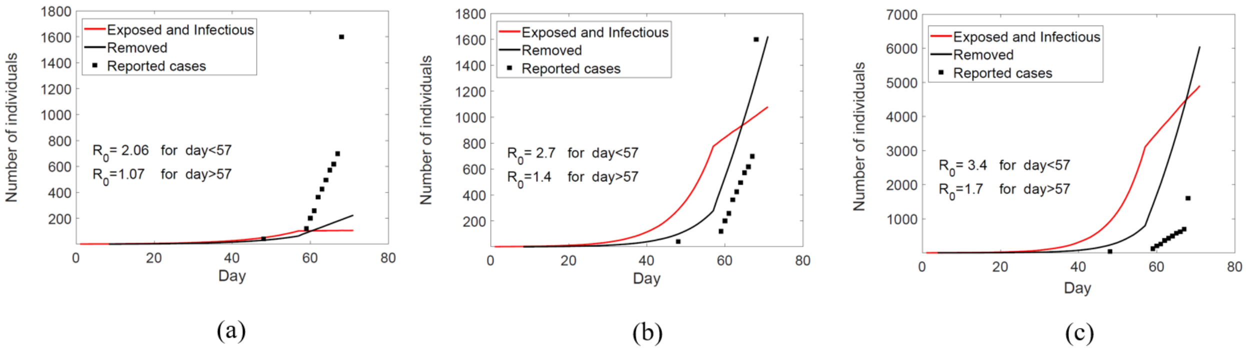


Figure 2.12: *(a) Empirical distributions of the infectious period from reference*[1]*. These curves are clearly non-exponential; (b) Our simulations of the spreading process obtained using the empirical distributions of figure 2.12.a (upper panel) and using exponential distributions (lower panel) with similar means as the empirical distributions. Fig. 2.12.b shows two sets of epidemic curves, infected undetected (exposed and infectious) and confirmed (removed) for the empirical and exponential distributions. Even though the two distributions share the same mean, they have different quantitative behaviors.*

distribution of the infectious period to follow the spreading process according to the Wuhan data (figure 2.12.a). We chose the distribution of incubation period based on Wuhan data as well.

For the second set of simulations, we used the same setting as the first set, except we changed the infectious period distributions to exponential distributions with similar means as the distributions in the first set of simulations. In all the simulations, we assumed the infectious transmission rate was 0.4/60. In figure 2.12.b, we have shown the result of the simulations. We can see how different distributions for the duration of the infectious period, even though having the same mean, lead to different epidemic curves.

In another experiment, we estimate the reproductive number of COVID-19 in Wuhan. To this end, we performed multiple sets of simulations with different infection rates to observe which value of the infection transmission rate produces the closest epidemic curve with respect to the reported number of cases in the early phase of the outbreak in Wuhan (see figure 2.13). In these simulations, based on some estimations we assumed the spreading



(a) (b) (c)

Figure 2.13: *Non-Markovian simulation of the SEIR model over a population of 60,000 nodes for three different values of the infection transmission rates. For the infectious period we used the distributions in figure 2.12.a and we switched to the distribution with the smaller mean at day 57, while keeping the infection transmission rate constant. We can see that panel (b) shows a better fitting of the number of reported cases in Wuhan, China, compared with panels a and c.*

processes started around November 20, and we changed the distribution of the infectious period after January 18, to account for the specific policies implemented in Wuhan that decreased the mean detection time (see figure 2.12.a). For the incubation period, we used a lognormal distribution of mean 5.5 and median 5.

In epidemiology, the basic reproduction number R_0 of an infection can be thought of as the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. Using these simulations, we estimated R_0 at 2.7 before January 18, and at 1.4 after January 18 (see figure 2.13.b). Although we heuristically explored a limited number of values for the transmission rate, our estimation of R_0 is close to the average of the estimated values of R_0 for the COVID-19[46].

Chapter 3

Identification of Missing Links Using SIS Spreading Traces[1]

3.1 Introduction

Recently, the effect of the network structure on the epidemic has been an active line of research[10;48–52]. Because the network structure leaves its imprint on the epidemic data, we expect the possibility of recovering some information about the network using the observed epidemic data. This inverse problem can be of particular interest when we have only partial information about the network structure that may render control of spreading impossible. To have an intuitive picture of how the epidemic data can be applied in the network structure inference, consider the simple graph in Fig. (3.1) where the links bc and ac are uncertain. Let's assume we have observed an SI process such that the infection times for the nodes a, b, c are $T_a = 0,\ T_b = \alpha$ and $T_c = 2\alpha$, respectively. If the transmission time through any network link is an exponential random variable with the expected value α, the probability of link bc to exist is higher than that of the link ac. This is due to the higher value of the argument of exponential function for the link ac, $\alpha^{-1}(T_c - T_a)$, than that of the link bc.

Here, we address the problem of recovering network structure from the traces of contin-

[1] This chapter is a slightly modified version of our published article[47], Copyright © 2019, IEEE.

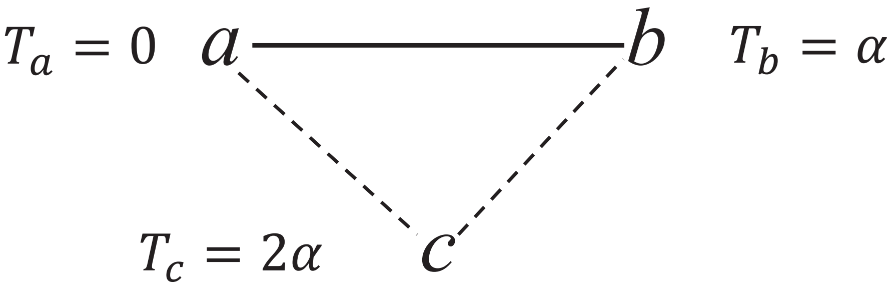


Figure 3.1: *In this network, we know the link ab exists, but the links bc and ac are uncertain. Also, we know the expected time for transmission of infection through any link is α. Since the difference between the infection time of nodes b and c equals the expected transmission time α and the same difference calculated for the nodes a and c is 2α, link bc is expected to be present in the network with a higher probability than link ac.*

uous time SIS spreading processes. We assume a setting where we observe the state of all the individual nodes through time. In section (3.2) we review some of the related works. In section (3.3) we derive the likelihood of the observed SIS traces in term of transmission rates as model parameters. When the domain of the transmission rates is a convex set, maximum likelihood estimation (MLE) of the transmission rates is a convex optimization problem. Since the transmission rates are disease-specific parameters, there are cases where we have prior knowledge about the transmission rate over an existing link. In such cases, the network-links recovery process could be cast as an MLE problem where the transmission rates are either zero or specific values. Instead of trying to solve such a binary optimization problem, in section (3.4) we formulate the binary network reconstruction within a Bayesian framework. Comparing to MLE approach, Bayesian inference can incorporate our prior belief regarding the existence of the links in the inference problem. This property of Bayesian inference is particularly useful when there are other layers of information concerning the presence of the links, independently of the SIS traces. Moreover, using Bayesian inference, unlike the maximum likelihood method which is a point estimator, we obtain posterior probabilities for the existence of the links. Thus, we have a measure for confidence of the estimation. In section (3.5), we perform numerical experiments on synthetic data, and we use Gibbs sampling to find the posterior probabilities of the links.

3.2 Related Works

Network structure inference using spreading data has been an active research area in data mining[53–64]. One of the earlier works in this field uses sequences of cascading events to recover the links among the nodes of a network[53]. In their model, the cascades spread as directed trees over an underlying graph, such that all the possible trees can happen with the same probability. But for a fixed tree, the cascades with different timings occur with different probabilities. Although this model may resemble the conventional SI model, there is a clear difference in the node-level description. Indeed, it is straightforward to see that different spreading trees have different probabilities in the SI model. Therefore, the algorithm proposed in[53] does not reconstruct the binary network of SI cascades.

In[56], the authors consider the spreading process that follows SI model in the node-level. They formulate the network reconstruction problem as a maximum likelihood estimation (MLE) of transmission rates among the nodes of the network. They assume that the domain of transmission rates is a convex set, which turns the MLE problem into a convex optimization problem. Moreover, another assumption is made that the transmission rates between any two nodes could be asymmetric. This assumption decouples the global maximum likelihood problem into N local problems, where N is the number of nodes in the network. In[56], it is shown that by adding the L_1 norm of transmission rates to the likelihood function the transmission rates can be recovered with a finite sample of SI traces. The authors of[54] assume a discrete time SIR model and use the maximum likelihood estimator to make inferences about the network structure. Moreover, they address the question concerning the number of required samples for the graph recovery. Finally, the network structure inference is not limited to link inference. In the existing literature, authors find the community structure among the network nodes using epidemic data[65;66]. In a different setting, the researchers' goal was to identify the source of infection from some observation of epidemic data[63;67–71].

After deriving the likelihood of SIS traces–equation (3.5)–it turns out we can use a similar approach as[56] to recover the network links from the observed SIS traces. In fact, if we assume the domain for the transmission rates is a convex set, the MLE problem is a convex optimization problem and decouples into local problems, if the rates between the nodes are asymmetric. Although the consistency of the MLE approach guarantees the recovery of correct transmission rates, we expect that including the known information in the likelihood function decreases the amount of data required for an accurate estimation of the network. For instance, if the transmission rate of the existing links is the specific value β_0, we can use $\{0, \beta_0\}$ as the domain of the rates. In section (3.5.2), we experimentally show how this extra information about the links' transmission rate affect the inference.

Our original contribution in this work includes (1) deriving the exact likelihood of continuous-time SIS traces, assuming complete nodes' state observation, and (2) formulating the binary network reconstruction in a Bayesian framework which enables us to use Gibbs sampling approach to find the exact probability that an uncertain link exists or not.

3.3 Likelihood of SIS Traces

In the susceptible-infected-susceptible (SIS) stochastic epidemic model, the individuals are either susceptible or infected. Here, we assume if a node is infected the probability to stay infected decreases with a constant rate δ. Thus, after the node gets infected at time t_0, it becomes susceptible again at time $t + t_0$ where t is a random variable with the exponential probability density function, $f(t) = \delta \exp(-\delta t)$. In this model, the transition from susceptible to infected state is caused by the interaction with an infected neighbor in the network. We assume, for a susceptible node that has one infected neighbor, the probability to remain susceptible up to time t is $\exp(-\lambda t)$, where t is the duration of contact and λ is a constant. In other words, the transmission time for the infection is a random variable that is exponentially distributed with the rate λ. In general, exponential distribution can be understood as limit of a discrete time process. For example consider a discrete infection process where in each time step the infected neighbor succeeds to transmit the infection with probability $\lambda\ dt$ and fails with probability $1 - \lambda\ dt$. In this scenario the probability that the susceptible node survives for k steps and gets the infection at step $k + 1$ is $(1 - \lambda\ dt)^k\ \lambda\ dt$. If dt is small we recover the the exponential distribution for the infection time,

$$(1-\lambda\ dt)^k \lambda\ dt \xrightarrow{dt\to 0} \exp(-\lambda t)\ \lambda dt, \tag{3.1}$$

where $k\ dt = t$. Considering this discrete time limit, we deduce that in the exponential distribution, $\lambda \exp(-\lambda t)$, the exponential function is the probability that the system survives up to time t and λ is a time-independent density.

In practice, a susceptible node may have several infected neighbors trying to infect it. In this case, the infection processes by different neighbors are independent, and the susceptible node gets the infection from the first neighbor that transmits it. If at $t = 0$ node a is susceptible and the infected neighbors set is $\mathcal{N}_{\mathcal{I}}$, the probability density function for the infection time, assuming all the infected neighbors remain infected, is

$$\begin{aligned} f(t) &= \prod_{n'\in\mathcal{N}_{\mathcal{I}}} \exp(-\lambda_{n',a} t) \sum_{n\in\mathcal{N}_{\mathcal{I}}} \lambda_{n,a} \\ &= \lambda_0 \exp(-\lambda_0 t) \end{aligned} \tag{3.2}$$

where $\lambda_0 = \sum_{n\in\mathcal{N}_{\mathcal{I}}} \lambda_{n,a}$. In the first line of the equation above, the product term is the probability that the susceptible node a survives up to time instant t and the summation term is the probability that at least one of the infected neighbors transmits the infection in the interval $(t, t+dt)$. The second line in the equation indicates that the minimum of several exponentially distributed random variables has an exponential distribution with a rate equal to the sum of all the rates for the independent variables.

Based on the node level description of the SIS process, clearly the transition of a node from susceptible state to infectious state depends on the state of its neighbors; this implies the dynamics of a node can not be decoupled from those of its neighbors. Instead, the joint state of all the N nodes in the network denoted as $\mathbb{S} = [s_1, s_2, \cdots, s_N]$, where $s_i = 1$ ($s_i = 0$) if n_i is infectious (susceptible) , is a continuous-time Markov chain over a space consisting of 2^N possible network states. To derive the likelihood of a network state trace, we first obtain the likelihood of events happening in the network. Assuming at $t = t_0$ the network state is

$\mathbb{S}(t_0)$, by an event we mean the first time that a node makes a transition after t_0. We specify an event by the ordered pair $e = (n(e), t(e))$ where $n(e)$ is the node that makes the transition and $t(e)$ is the time at which the event happens. Based on the node level description of the SIS process, we can write the probability density function of the event as

$$\begin{aligned} & f(t(e) \mid \mathbb{S}(t_0), \Lambda, \delta) = \\ & \Big(s_{n(e)}(t_0)\delta_{n(e)} + (1 - s_{n(e)}(t_0)) \sum_{q \in \mathcal{N}} s_q(t_0)\lambda_{q,n(e)} \Big) \\ & \times \exp\Big(-\Delta\Big(\mathbb{S}(t_0)\delta + \mathbb{S}(t_0)\Lambda\big(\mathbb{1} - \mathbb{S}^{\dagger}(t_0)\big)\Big)\Big), \end{aligned} \tag{3.3}$$

where $\Delta = t(e) - t_0$. In Eq.(3.3), Λ is the matrix of interaction rates among the nodes of the network $\Lambda_{q,p} = \lambda_{q,p}$, δ is a column vector of recovery rates for different nodes and $\mathbb{1}$ is a column vector with all the elements equal one.

The second line on the r.h.s of the of Eq.(3.3) is the probability that the network state $\mathbb{S}$ stays constant in the time interval $(t_0, t_0 + \Delta)$, while the first line gives the density for the transition of the node $n(e)$.

3.4 Bayesian Inference of Missing Links

To use the Bayesian method of inference, we need to calculate the likelihood of observed data conditioned on the parameters. In our problem statement, we assumed we have access to the complete history of nodes' traces for a period of time from $t = 0$ to $t = T$. Although in theory it is possible to observe the state of each node, $s_i(t)$, on the continuous timeline, in practice we may observe the node's state only at some points on the continuous timeline. For such kind of observation we assume the observation window is much smaller than the inverse of transition rate of the nodes. With this assumption the probability that two different events happen in the same time window goes to zero as we make the time window smaller. Moreover, the likelihood of the events tends to the one for the continuous observation in Eq.(3.3).

Using the nodes' traces, we can extract all the network state events that occur for that period of time. In other words, the observed data is the sequence of network state events. Assuming $\mathcal{C} = \{e_1, e_{2,\dots}\}$ is the observed sequence of events ordered by occurrence in time, $t(e_1) < t(e_2) < \cdots$, the likelihood of the sequence $\mathcal{C}$ is

$$f(\mathcal{C} \mid \Lambda, \delta) = \prod_{i=1} f(t(e_i) \mid \mathbb{S}(t(e_{i-1})), \Lambda, \delta) \tag{3.4}$$

where the terms in the product are calculated from Eq.(3.3) Since we want to make an inference about the interaction rates, in the expression for the likelihood of sequence $\mathcal{C}$, we are only interested in those terms that are a function of λ. After inserting the density function of events from Eq.(3.3) into Eq.(3.4) and absorbing the terms that are a function of δ into the variable $K(\delta)$, the probability density function of sequence $\mathcal{C}$ simplifies as

$$\begin{aligned} f(\mathcal{C} \mid \Lambda, \delta) = K(\delta) \times \exp\Big(- \sum_{q,p \in \mathcal{N}} T_{q,p}\ \lambda_{q,p} \Big) \\ \times \prod_{e_i \in \mathcal{C}_I} \sum_{q \in \mathcal{N}} s_q(t(e_{i-1}))\ \lambda_{q,n(e_i)} \end{aligned} \tag{3.5}$$

In this expression, $\mathcal{C}_I$ refers to the set of all the events e_i in the sequence $\mathcal{C}$ that are infecting events, $s_{n(e_i)}(t(e_{i-1})) = 0$ and $s_{n(e_i)}(t(e_i)) = 1$. Here, $s_q(t(e_{i-1}))$ is the state of node q just before the event e_i happens. The constant parameter $T_{q,p}$ in Eq.(3.5) is the total period of time that node q had the possibility to infect node p, in other words

$$T_{q,p} = \int_0^T (1 - s_p(t)) s_q(t)\ dt.$$

The likelihood of sequence $\mathcal{C}$ presented in Eq.(3.5) is valid for both directed and undirected networks. When the network is directed $\lambda_{q,p}$ is assumed to be a parameter different from $\lambda_{p,q}$. Instead, when the network is undirected, $\lambda_{q,p}$ and $\lambda_{p,q}$ refer to the same parameter. Although in deriving the likelihood of the SIS trace we assumed a link between any two nodes p, q with a corresponding rate $\lambda_{p,q}$, the expression in Eq.(3.5) is also valid when some of these links are absent. For nonexisting links, we can simply apply $\lambda_{p,q} = 0$ and arrive at the correct expression for the likelihood.

Now that we have an expression for the likelihood of observed data, we can use Bayes' theorem to find the posterior distribution for the uncertain links. If we use Λ^u to indicate the set of transmission rates for the uncertain links, Bayes' theorem gives the joint posterior distribution of the transmission rates as

$$\begin{aligned} f(\Lambda^u \mid \mathcal{C}) = \kappa \times \exp\Big(- \sum_{\lambda \in \Lambda^u} T_\lambda \; \lambda \Big) \\ \times \prod_{e_i \in \mathcal{C}_I} \Big(\beta_{e_i} + \sum_{\lambda \in \Lambda^u_{e_i}} \lambda \Big) \times \mathfrak{F}(\Lambda^u). \end{aligned} \tag{3.6}$$

In this equation, κ is a normalization factor, and $\mathfrak{F}(\Lambda^u)$ is the prior distribution of the transmission rates for the uncertain links. Here we have used $\Lambda^u_{e_i}$ to refer to the set of transmission rates for those uncertain links that were active in the event e_i. The link $(q, n(e_i))$ is active in the infecting event e_i if and only if the node q is infectious at the time when the event happens. Moreover, in Eq.(3.6), β_{e_i} is the sum of all the transmission rates for the active links in the event e_i except those active links that are uncertain.

$$\Lambda^u_{e_i} = \Big\{ \lambda_{q,n(e_i)} \mid \lambda_{q,n(e_i)} \in \Lambda^u, s_q\big(t(e_{i-1})\big) = 1 \Big\}$$

$$\beta_{e_i} = \sum_{\substack{q \in \mathcal{N} \\ \lambda_{q,n(e_i)} \notin \Lambda^u}} s_q\big(t(e_{i-1})\big) \; \lambda_{q,n(e_i)}$$

As we mentioned before, if an uncertain link is undirected, both $\lambda_{q,p}$ and $\lambda_{p,q}$ refer to the same parameter $\lambda \in \Lambda^u$. In such cases $T_\lambda = T_{p,q} + T_{q,p}$. Conversely, when $\lambda \in \Lambda^u$ refers to a directed link $\lambda_{p,q}$, then we have $T_\lambda = T_{p,q}$.

The expression in Eq.(3.6) provides a joint distribution for the uncertain links. However, we are often interested in some marginal probability distribution such as the distribution of a specific link λ_0. To obtain the posterior distribution $f(\lambda_0 \mid \mathcal{C})$, we need to integrate the joint distribution over all the other uncertain links. If $\Lambda^- = \Lambda^u - \{\lambda_0\}$, we have

$$f(\lambda_0 \mid \mathcal{C}) = \int f(\Lambda \mid \mathcal{C}) \; d\Lambda^- \; . \tag{3.7}$$

Although in some cases this integration is a straightforward task, when we have a large number of uncertain links, the integration might be intractable. Nevertheless, when the prior distribution in Eq.(3.6) is a product of independent factors

$$\mathfrak{F}(\Lambda^u) = \prod_{\lambda \in \; \Lambda^u} \mathfrak{f}_\lambda(\lambda) \; , \tag{3.8}$$

the integration in Eq.(3.7) results in a marginal distribution that has a functional form of

$$\begin{aligned} f(\lambda_0 \mid \mathcal{C}) = \kappa_0 \times \exp\Big(- T_{\lambda_0}\lambda_0\Big) \; \mathfrak{f}_{\lambda_0}(\lambda_0) \\ \times \sum_{j=0}^{J} a_j \; {\lambda_0}^j . \end{aligned} \tag{3.9}$$

In this equation, J is the number of infecting events that the link λ_0 has been active in, κ_0 is a normalization factor, and the coefficients in the polynomial term can be calculated by performing the integration. However, for a general case when λ_0 is coupled with a large number of uncertain links through the factor terms in the joint distribution of Eq. (3.6), the integration is not tractable. In such cases, it is possible to apply one of the commonly used numerical methods in Bayesian inference such as the Markov chain Monte Carlo (MCMC) or Belief propagation. Here, we use the prior distribution of Eq.(3.8) where the independent factors have the functional form as

$$\mathfrak{f}_\lambda(\lambda) = \mathrm{P}_\lambda \; \delta(\lambda - r_\lambda) + (1 - \mathrm{P}_\lambda) \; \delta(\lambda) \tag{3.10}$$

Here, P_λ is the prior probability for the existence of the link, r_λ is the transmission rate assuming the link exists, and $\delta(\lambda)$ is the Dirac delta function. To find P_λ, one can use some source of information about the network other than the epidemic traces. For example, when the existence of links between the nodes stochastically depends on some kind of distance between the nodes, one can use the known distance to find the prior probability P_λ. Furthermore, when we do not have any prior information about a link, we can use $\mathrm{P}_\lambda = 1/2$ as the prior probability. In cases where the integration in Eq.(3.7) is tractable, we can find the posterior probability for existence of the link, $\widehat{\mathrm{P}}_\lambda$, from the equation below

$$\frac{\widehat{\mathrm{P}}_\lambda}{1-\widehat{\mathrm{P}}_\lambda} = \frac{\mathrm{P}_\lambda \ \exp\Big(-T_\lambda r_\lambda\Big)\ \sum_{j=0}^{J} a_j \ {r_\lambda}^j}{(1-\mathrm{P}_\lambda)\ a_0} \ . \tag{3.11}$$

In practice, for most cases where it is not possible to calculate polynomial coefficients analytically, we can use Gibbs sampling to find the posterior probability $\widehat{\mathrm{P}}_\lambda$. Gibbs sampling requires constructing a Markov chain over the parameters' space that has an equilibrium distribution similar to the joint distribution of the parameters. In our inference problem, the parameters' space is the direct sum of transmission rate space of uncertain links, and the Gibbs sampling constructs a Markov chain $X = (\lambda_1, \lambda_2, \cdots, \lambda_k)$, $\lambda_i \in \Lambda^u$, which has an equilibrium distribution similar to that in Eq.(3.6). When we use the prior distribution in Eq.(3.10) for the transmission rate λ, the corresponding component of the Markov chain can only assume values of 0 or r_λ, and using samples from the Markov chain, we can estimate $\widehat{\mathrm{P}}_\lambda$ as the fraction of samples with $\lambda = r_\lambda$.

3.4.1 Illustrative Example

To clarify the formulas discussed previously through the text, here we apply the Bayesian reconstruction method to a simple example. Consider the graph in Fig.(3.2) where the link $(1,2)$ exists and the links $(1,4)$, $(2,4)$, $(1,3)$ and $(2,3)$ are uncertain and we want to find out the probability that they exist. Moreover, we assume the links in the graph are undirected and the transmission rates over all the links are symmetric and if a link exists the transmission rate is β. For the network state trace, we use a trace of three events where in the first event, shown by e_a, node 3 recovers, in the event b node 1 gets infected and in the event c node 2 gets infected. This network state trace is shown in Eq.(3.12) where Δ over

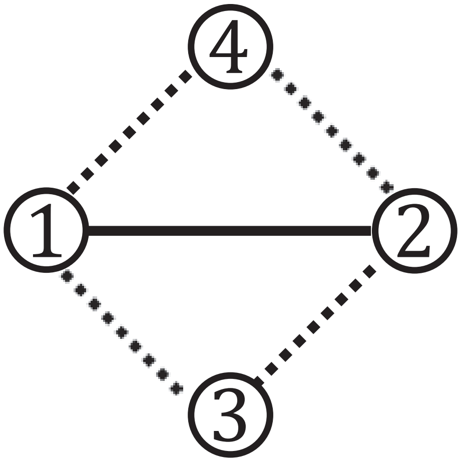


Figure 3.2: *Network used in section 3.4.1. In this network, the link* $(1,2)$ *exists, but the links* $(1,4)$, $(2,4)$, $(1,3)$ *and* $(2,3)$ *are uncertain. Moreover, we assume the transmission rates are symmetric.*

each arrow is the period of time that network state has remained constant before the event.

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \xrightarrow{\Delta_a} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \xrightarrow{\Delta_b} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \xrightarrow{\Delta_c} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{3.12}$$

Using the formula in the equation 3.6, the joint posterior distribution of the transmission rates can be written as

$$\begin{aligned} f\big(\lambda_{1,4}, \lambda_{2,4}, \lambda_{1,3}, \lambda_{2,3} \mid \mathcal{C}\big) &= \kappa \times \lambda_{1,4} \times \big(\lambda_{2,4} + \lambda_{1,2}\big) \\ &\times e^{-\big(\lambda_{1,4}(\Delta_a+\Delta_b)+\lambda_{2,4}(\Delta_a+\Delta_b+\Delta_c)+\lambda_{1,3}(\Delta_a+\Delta_c)+\lambda_{2,3}\Delta_a\big)} \\ &\times \mathfrak{f}(\lambda_{1,4}) \times \mathfrak{f}(\lambda_{2,4}) \times \mathfrak{f}(\lambda_{1,3}) \times \mathfrak{f}(\lambda_{2,3}), \end{aligned} \tag{3.13}$$

where κ is a normalization factor and we have used independent prior distribution such that each of the uncertain links could exit with a probability equals $1/2$ independent from the other links,

$$\mathfrak{f}(\lambda) = \frac{1}{2}\,\delta(\lambda - \beta) + \frac{1}{2}\,\delta(\lambda).$$

However, in a different setup we could define a prior distribution such that the existence of several uncertain links were tangled together.

In order to find the posterior distribution of a specific uncertain link we need to integrate the joint distribution in Eq.(3.13) over all the other uncertain links. For example, the posterior distribution of $\lambda_{2,4}$ is

$$\begin{aligned} f\big(\lambda_{2,4} \mid \mathcal{C}\big) &= \int f\big(\lambda_{1,4}, \lambda_{2,4}, \lambda_{1,3}, \lambda_{2,3} \mid \mathcal{C}\big)\ d\lambda_{1,4}\ d\lambda_{2,3}\ d\lambda_{1,3} \\ &= \kappa \times \big(\lambda_{2,4} + \beta\big) \times e^{-\lambda_{2,4}(\Delta_a+\Delta_b+\Delta_c)}\ \mathfrak{f}(\lambda_{2,4}), \end{aligned}$$

where κ is again a normalization factor and after normalization, the posterior distribution for $\lambda_{2,4}$ becomes

$$f(\lambda_{2,4}) = \frac{2e^{-\beta(\Delta_a+\Delta_b+\Delta_c)}\ \delta(\lambda_{2,4}-\beta) + \delta(\lambda_{2,4})}{1 + 2e^{-\beta(\Delta_a+\Delta_b+\Delta_c)}}.$$

In the equation above, the coefficient of Dirac delta function $\delta(\lambda_{2,4} - \beta)$, gives the probability that the link $(2, 4)$ exists.

3.5 Numerical Experiments

In this section, we report the result of two numerical experiments on synthetic data. In both experiments, we assume if an uncertain link exists, it is undirected, and the infection rate is known. In the first experiment, we show how the extra information incorporated in the Bayesian link inference as a prior distribution affects the posterior probabilities. In the second experiment, we compare two different approaches to the binary network reconstruction where we estimate the transmission rates using MLE and compare the result with the posterior probabilities for the existence of the links obtained using Bayesian approach. Although in these experiments we use two specific underlying networks, our network reconstruction approach is independent of the underlying network topology and only depends on the spreading model.

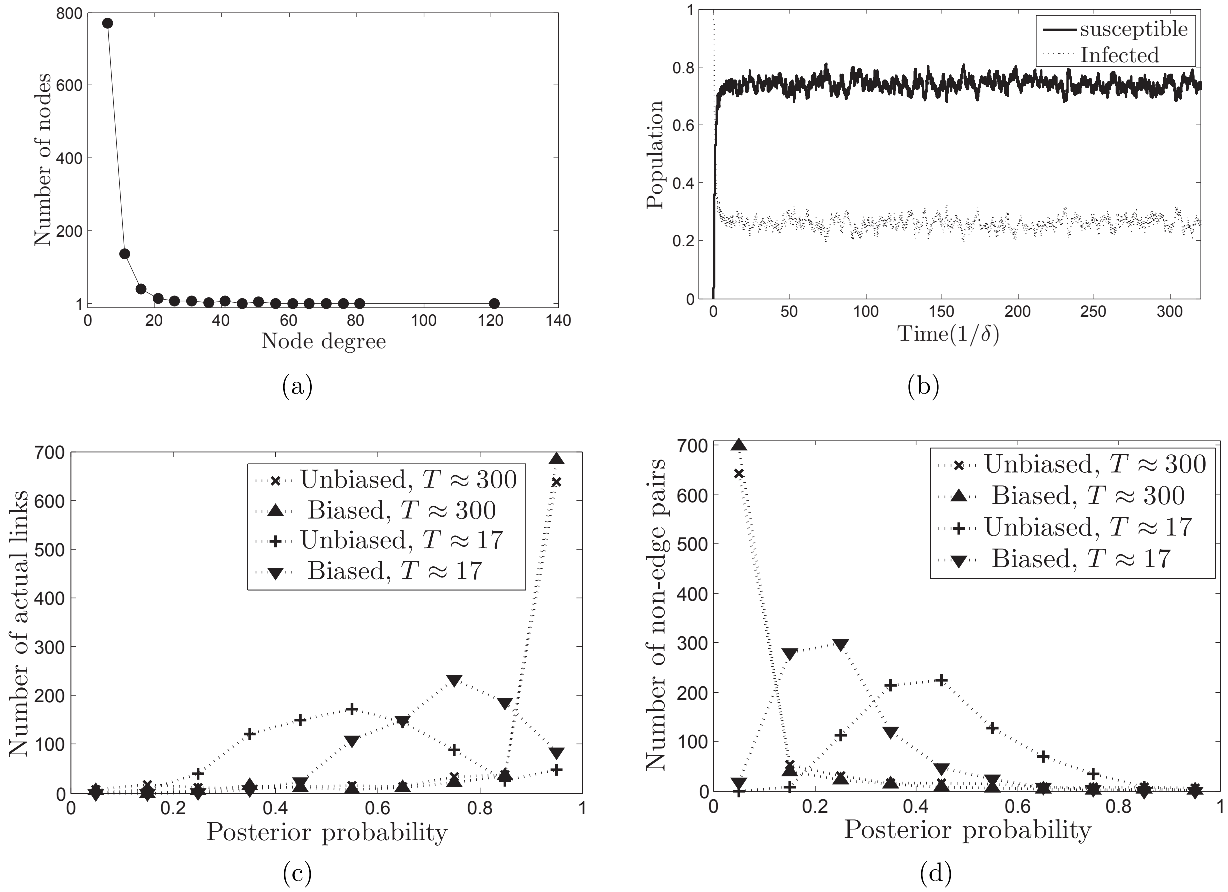


Figure 3.3: *Plots for the experiment in the section (3.5.1). (a) node degree distribution of the underlying scale-free network. (b) proportion of the infected and susceptible nodes in the SIS trace. (c) distribution of the posterior probabilities for the actual links of the network and (c) distribution of posterior probabilities for the non-edge pairs, when different prior distributions and traces of different length were used in the link inference.*

3.5.1 Experiment Using Informative Prior

In this experiment, we first generated a random scale-free network with 1000 nodes following Barabasi–Albert (BA) model[44]. The nodes' degree of the networks resulting from this model has a Power law distribution, $P(k) \sim k^{-3}$. The generated network we used in this experiment has 3980 undirected links and the average node degree is 7.96. The node degree distribution of the network is plotted in Fig(3.3a), which shows there are a few hubs in the network.

After generating the network, we simulated an SIS spreading over the network using the GEMF simulator[33;38]. We assumed all the nodes are initially infected and the infection transmission rate for all the links is $\beta = 0.15\ \delta$, where δ is the recovery rate. Fig.(3.3b) shows the infected and susceptible population for the simulated SIS process.

For this experiment, we randomly chose a set of uncertain links, and we find the posterior probabilities for the uncertain links using the information in the SIS epidemic trace. An uncertain link can be an actual link of the network or a possible link that is not realized in the network. If the network is shown by $G = (N, E)$, where N is the set of nodes and $E \subseteq N \times N$ is the set of actual links, we refer to the elements of $\bar{E} = N \times N - E$ as non-edge pairs. For the set of uncertain links we randomly chose 800 actual links and 800 non-edge pairs. Figure (3.3c) shows the distribution of posterior probabilities for those uncertain links that are the actual links and Fig.(3.3d) shows the similar distribution for the non-edge pairs. Moreover, in figures (3.3c) and (3.3d) we can see the comparison between two cases with different prior distributions. In one case we used uninformative prior such that in the equation (3.8)

$$\mathfrak{f}_\lambda(\lambda) = \frac{1}{2}\,\delta(\lambda - \beta) + \frac{1}{2}\,\delta(\lambda).$$

This prior distribution does not favor the presence or absence of an uncertain link, but it limits the transmission rate of a link to the known value of $\beta = 0.15$. In the figures those distributions denoted by "Unbiased prior" are the result of using the uninformative prior. In the second case we used a prior distribution that assigns a prior probability of 0.7 to those uncertain links that are actual links of the network and assigns 0.3 to the non-edge pairs,

$$\begin{aligned} \mathfrak{f}_\lambda(\lambda) &= 0.7\,\delta(\lambda - \beta) + 0.3\,\delta(\lambda) \quad \text{for the actual links} \\ \mathfrak{f}_\lambda(\lambda) &= 0.3\,\delta(\lambda - \beta) + 0.7\,\delta(\lambda) \quad \text{for the non-edge pairs} \end{aligned} \tag{3.14}$$

This prior distribution reflects our beliefs about the uncertain links. If these beliefs are correctly biased, as we have assumed here, we need less amount of data to infer a posterior probability that is closer to the true value, 1 for the actual links and 0 for the non-edge pairs. In figures 3.3c and 3.3d, the plots denoted by "Biased prior" are obtained starting from the informative prior in the equation (3.14).

In Fig.(3.3c) we can see the difference between distribution of posterior probabilities for the actual links when the SIS traces of different time lengths are used in the inference. When the SIS trace is long enough, regardless of the prior, the distributions are close to the true distribution, which is nonzero only for the posterior probability equals 1. However, when the SIS trace does not contain enough information (plots with $T \approx 17$), using the biased priors leads to the posterior probabilities that are closer to the true values. In Fig.(3.3d) we can see a similar trend in the inferred posterior probabilities for the non-edge pairs. The only difference is that the true value of the posterior probabilities for the non-edge pairs is zero.

3.5.2 Experiment to Compare MLE and Bayesian Approaches

In the next experiment, we considered inferring all the links in a network of 100 nodes using an SIS spreading trace. We assumed the nodes are positioned on the vertices of a square grid, and we constructed a synthetic network by connecting any two nodes a, b with a probability that decreases with their distance as

$$p(a, b) = \frac{0.3}{d(a, b)}, \tag{3.15}$$

$d(a, b)$ is the Euclidean distance of nodes a, b in a length unit defined by the grid's edges. This resulted in a network with 390 edges. For this network, we assumed the transmission rates over the links are $\beta = 0.21\delta$ where δ is the recovery rate of the nodes. We simulated an SIS epidemic initiated from 10 randomly infected nodes using the GEMF simulator[38]. In the simulation the population of infected nodes on average was about 1/3 of the nodes.

In this experiment, the set of uncertain links includes all possible links between the nodes and we infer them using two different methods and compare the results. For the first method, we use Maximum Likelihood Estimation (MLE) and for the second method we adopt Bayesian approach with an uninformative prior.

If we assume the network is directed, the likelihood of SIS trace in Eq.(3.5) decouples into a set of independent functions,

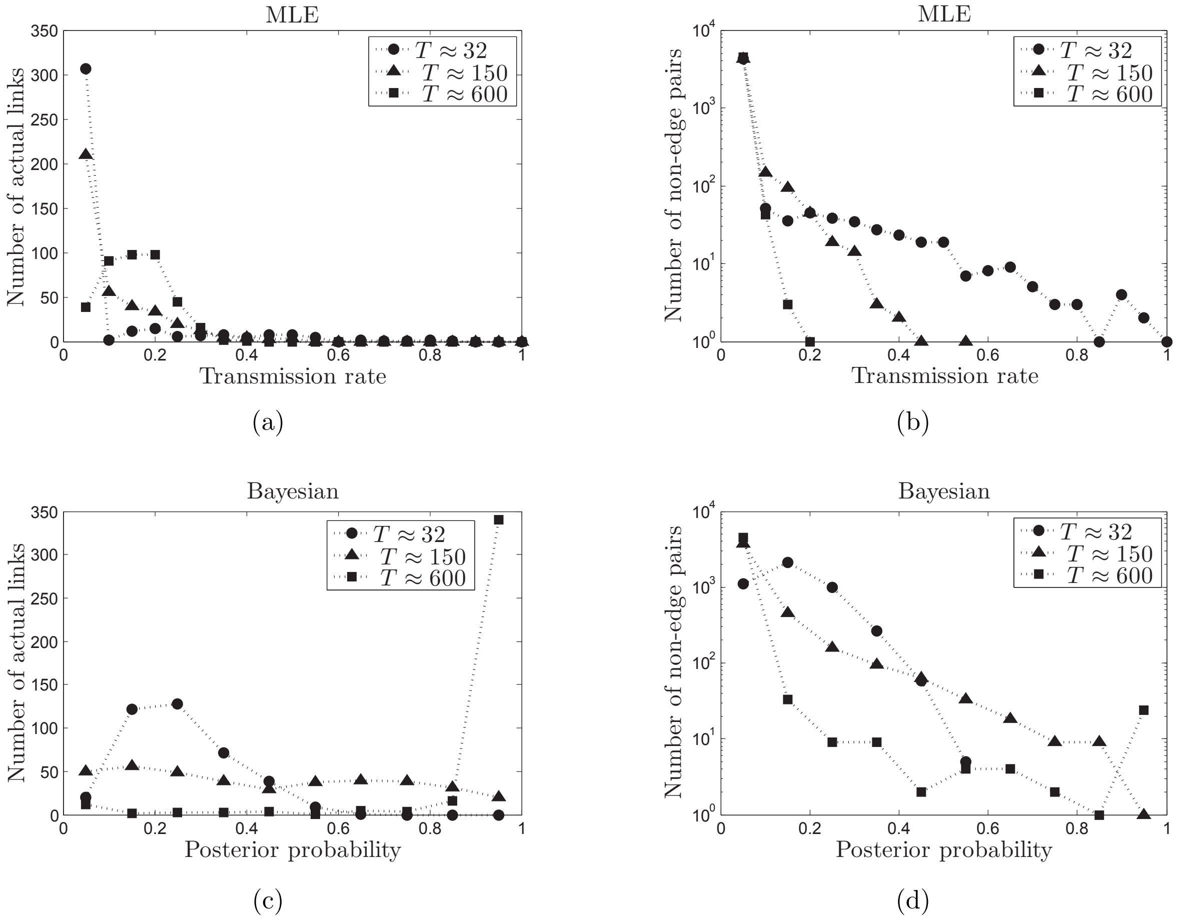


Figure 3.4: *Results of the experiment in section 3.5.2. (a) and (b) show distributions of the maximum likelihood estimation of the transmission rates for the actual links and the non-edge pairs respectively. (c) and (d) show distributions of the posterior probabilities for the actual links and the non-edge pairs. In the plots, the curves with different T are obtained using SIS traces with different length.*

$$f(\mathcal{C} \mid \Lambda, \delta) = K(\delta) \times \prod_{p \in \mathcal{N}} g(\mathcal{C}_I^p \mid \Lambda_p), \tag{3.16}$$

where

$$\begin{aligned} g(\mathcal{C}_I^p \mid \Lambda_p) = \exp\Big(- \sum_{q \in \mathcal{N}} T_{q,p}\ \lambda_{q,p} \Big) \\ \times \prod_{e_i \in \mathcal{C}_I^p} \sum_{q \in \mathcal{N}} s_q(t(e_{i-1}))\ \lambda_{q,p}. \end{aligned} \tag{3.17}$$

In the equation above, $g(\mathcal{C}_I^p \mid \Lambda_p)$ is the likelihood of the set events that node p got infected, denoted by $\mathcal{C}_I^p$, and $\Lambda_p = \{\lambda_{q,p} \mid q \in \mathcal{N}\}$ is the set of transmission rates from all the other nodes to the node p. In[56], authors have used a similar decoupling to reduce the size of optimization problem in the MLE estimation of links using SI cascades. They have assumed the transmission rates are continuous variables and they maximize the objective function of each node p separately. If the domain of transmission variables is convex, the resulting optimization is convex. Although this approach leads to a simple convex problem, in some cases it is not preferable. Assume we know the value of transmission rate if a link exists. If we incorporate this information in the MLE estimation, instead of a convex problem, we need to solve an integer-programming problem which is not linear. One may claim if the SIS trace contains enough information, there is no need to incorporate the known value of transmission rates in the MLE. However, when the observed data is limited, considering all the known facts about the process results in a more accurate estimation. In order to investigate the effect of neglecting such prior information we perform the MLE estimation of the links and compare the result with the true values. Moreover, we followed[56] and add a constraint on the L1-norm of transmission rates such that

$$\sum_{q \in \mathcal{N}} \lambda_{q,p} \leq d \times \beta = 3.4,$$

where d is the largest degree of the nodes in the network and $\beta = 0.21$. This constraint encourages sparsity and it is shown that by applying such type of constraints, the network can be recovered with finite samples of SI epidemic model,[56]. Since the likelihood function of SIS model in Eq.(3.17) is similar to the likelihood obtained from the traces of SI model, we applied the L1-norm constraint to guarantee the recovery of network with a finite length SIS trace. Figures 3.4a and 3.4b show the distribution of estimated transmission rates for the actual links and the non-edges pairs correspondingly. Since in the MLE we assumed the links are directed, there are two estimated values per link. In the plots we have used the average of the two estimated values.

In addition to the MLE, we performed Bayesian inference with an uninformative prior,

$$\mathfrak{f}_\lambda(\lambda) = \frac{1}{2}\,\delta(\lambda - \beta) + \frac{1}{2}\,\delta(\lambda),$$

where $\beta = 0.21$. Figures 3.4c and 3.4d show the distribution of posterior probabilities for the actual links of the network and the non-edge pairs correspondingly. From the figures it is obvious when the SIS trace is long, $T \approx 600$, the posterior probabilities for almost all the links are close to the true probabilities, 1 for the actual links, and 0 for the non-edge pairs. In practice one can assume a threshold for the transmission rates such that if the maximum likelihood estimation of transmission rate for a link is higher than the threshold, the link is considered as an actual link of the network,[56]. Since the distribution of transmission rates for the actual links and the non-pair edges overlap (compare figures 3.4a and 3.4b), this thresholding procedure identifies some non-edge pairs as the actual links of the network. In figure 3.5 the curves denoted by "MLE" show the number of non-edge pairs and the number of the actual links that are recovered as the actual link of the network when the threshold value varies. Similarly, we can define a threshold for the posterior probability such that if the inferred posterior probability of a link is higher than the threshold, the link is considered as an actual link of the network. In figure 3.5 the curves denoted by "Bayesian" show the number of actual links that are recovered, respect to the number of non-edge pairs that are wrongly identified as the actual links if the posterior threshold changes. In the figures 3.5 we can see when the SIS trace is short Bayesian inference has a better performance than the MLE approach since it can recover more actual links if the number of falsely identified non-edge pairs is fixed.

3.5.3 Experiment on Different Underlying Network

To study influence of the underlying network topology on network recovery, we perform an experiment using three different networks. The first one is the largest component of the co-authorship network[39]. This network has 379 nodes and 913 links. The second network is a Barabasi-Albert network with 350 nodes and 1035 links and the third network is a random network with 350 nodes and 1035 links. Although these networks almost have the same number of nodes and links, their structures are extremely different. Figure 3.6a shows the node degree distribution of the three networks. We can see the Barabasi-Albert (scale-free) network has several hubs while the co-authorship (real network) has fewer hubs and the random network contains no nodes with high degrees. Figure 3.6b shows the population of infected nodes through time in the simulation of SIS processes over these underlying networks. We adjusted the transmission rate, β, such that the populations of infected nodes in the steady state are almost similar. Thus, we expect the accuracy of network recovery mainly would be influenced by the networks' structure. In this experiment, we assumed a set of uncertain links composed of all the actual network links and an equal number of randomly chosen non-edge pairs. Moreover, we used the uninformative prior probabilities for the uncertain links where each uncertain link is an actual link of the network with probability 1/2. Finally, in order to measure the accuracy of network recovery, we used average error that we define as

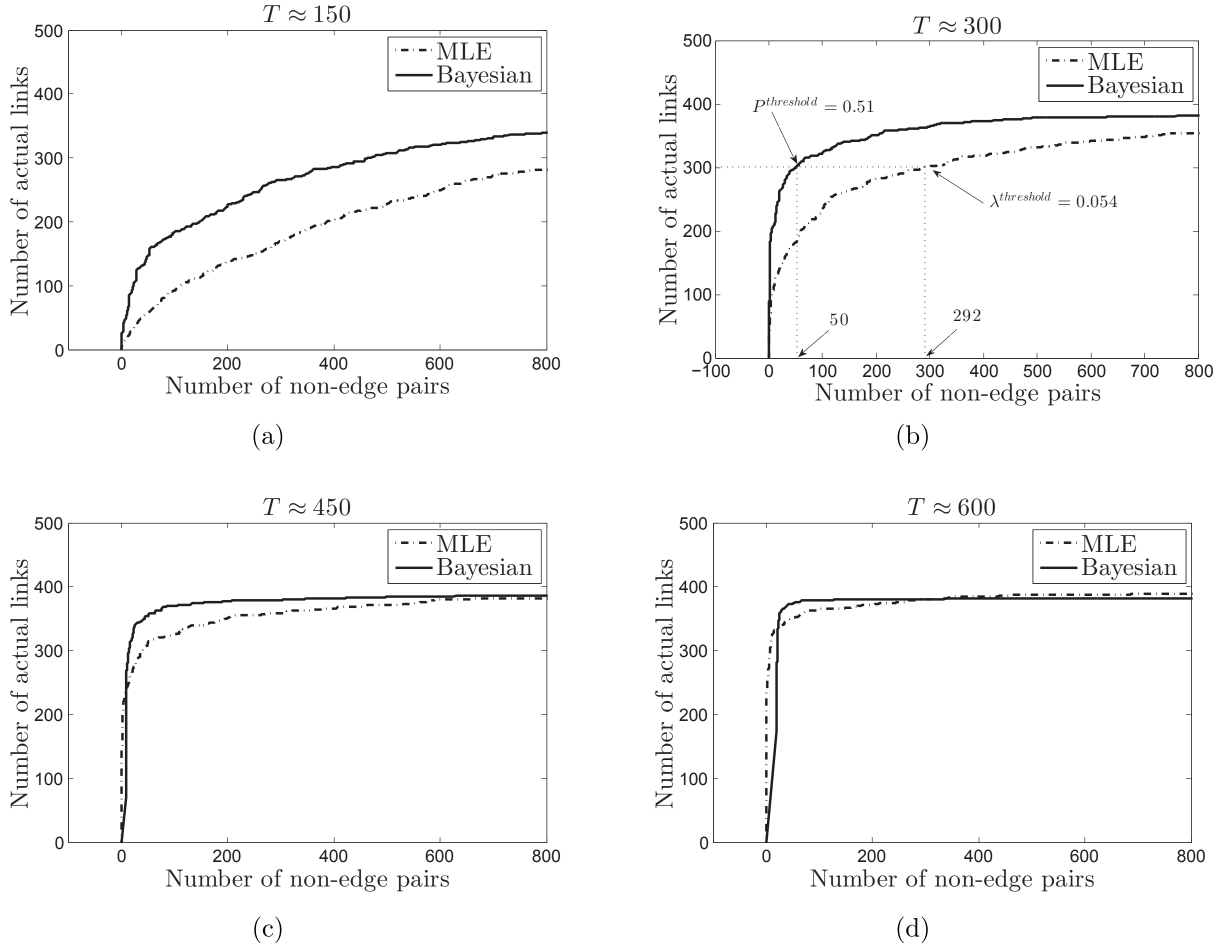
$$\overline{e} = \frac{\sum_{i=1}^{n} |\gamma_i - \widehat{\mathrm{P}}_i|}{n}. \tag{3.18}$$

In this equation, $\widehat{\mathrm{P}}_i$ is the posterior probability of the uncertain link i. If the uncertain link is an actual link in the network, $\gamma_i = 1$, otherwise, $\gamma_i = 0$.

Figures 3.6c and 3.6d shows the average error in the recovery of the actual inks and the non-edge pairs as a function of the SIS trace lengths used in the inference. We can see the error in the recovery of the Random network and the scale-free network are very similar even though they are quite different in their structure. In general it is not obvious how the network structure should affect the network recovery. However, we expect the error in the recovery of actual links should depend on the number of infecting events at either side of the link and the population of the infected nodes. In fact when there is a small number of infected nodes that could transmit infection it should be easier to trace the source of infecting events. In the same manner, we expect the inference of the uncertain link which is a non-edge pair becomes more accurate when the SIS trace contains incidents where one side of the link is infected and the other side is susceptible. In this case the probability for the existence of the link decreases exponentially with the total time duration of such incidents. When the population of infected nodes is high it is less possible to see such cases and the inference gets less accurate and we need to use SIS traces with longer duration. From this argument we expect the population of infected nodes in the SIS trace should play a more important role than the underlying network structure in the accuracy of network recovery. In fact figures 3.6c and 3.6d show when the infection populations are the same for the two completely different network structures, scale-free and random network, the accuracy of the network recovery are very similar. On the other hand, from figure 3.6d we see the accuracy in the inference of non-edge links in the real network is higher than that of the other two networks. We think this is due to the slow rise of the infection population that provides more events where one end of the uncertain link is infected and the other end is susceptible. In general, since the population of the infection is very influential in network recovery, we expect the SIS traces that the transition from small infection population to the steady state takes longer time contains more information about the uncertain links and this leads to a more accurate inference.

3.6 Summary

In this work, we investigated the inverse problem of continuous time SIS spreading over a graph. We derived the likelihood of SIS traces assuming we observe the states of the nodes over a period of time. Here, we formulated the binary network reconstruction as a Bayesian inference problem. Using this approach we obtain the probability that an uncertain exists. In section (3.5) we saw when the SIS traces contain enough information these probabilities were close to 1 for almost all the actual network links and close to 0 for the non-existing links. Although we only considered the links recovery using the SIS traces, the generalization to the traces of SI and SIR models of spreading is straightforward. Indeed, in the SI and SIR models, the only nodal transition that depends on the network links is the transition from susceptible state to the infected state, therefore the posterior distributions are similar to the one in the equation (3.6). Moreover, if we want to recover the binary network using the SI or SIR traces we can follow the same approach we adopted for the SIS traces and perform Gibbs sampling. Finally, the numerical comparison in section 3.5.2 shows when the transmission rates are known and the SIS traces does not contain enough information, the binary network reconstruction leads to a more accurate result than the weighted network recovery. Moreover, by employing the Bayesian approach we can obtain the posterior probabilities for the existence of the links, which provides a proper measure to evaluate the confidence of the estimation.



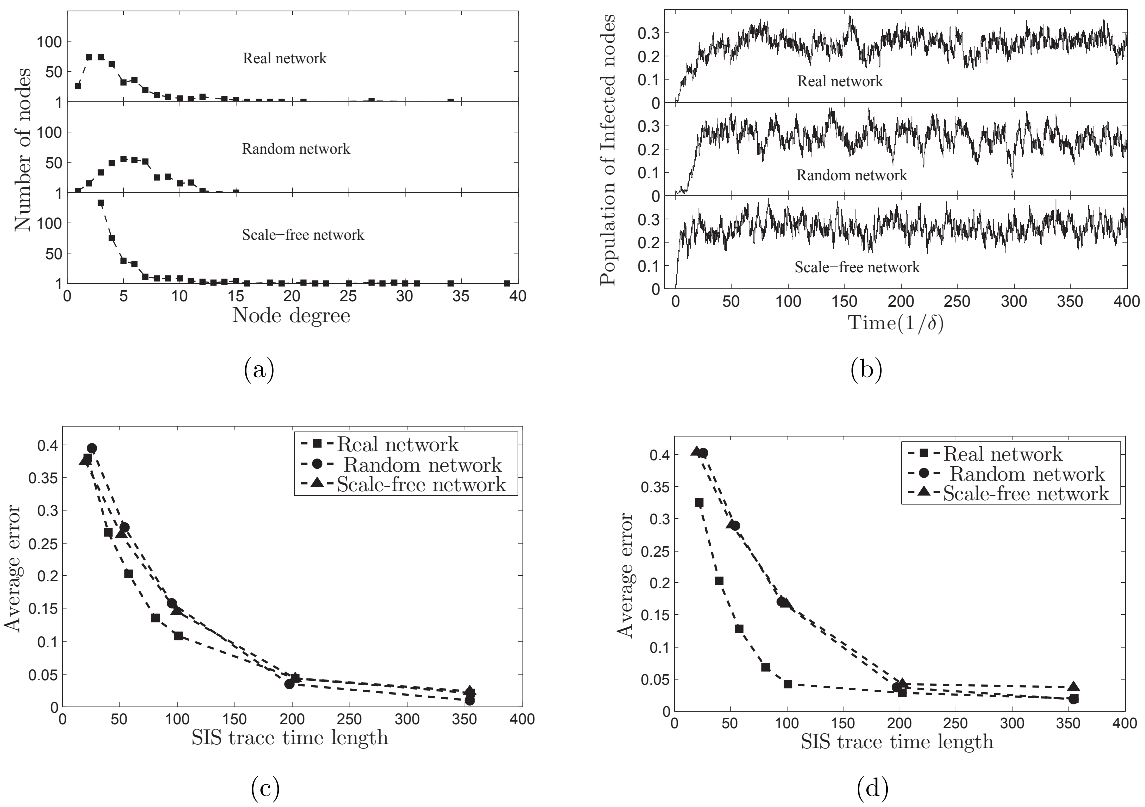
(a)

(b)

(c)

(d)

Figure 3.5: *Plots show the number of recovered actual links compared to the falsely recovered non-edge pairs when we apply the thresholding procedure explained in section (3.5.2). The curves are generated by changing the threshold value. To obtain each point on the curves denoted by "MLE", we associate a link for any pair of nodes if the MLE of transmission rate between them is higher than the threshold. For the curves denoted by "Bayesian" we apply threshold on the posterior probabilities. Plots (a), (b), (c), (d) are the result of applying SIS traces of length* $T \approx 150$, $T \approx 300$, $T \approx 450$, $T \approx 600$ *, respectively. In plot (b),* $\lambda^{threshold}$ *and* $P^{threshold}$ *are the transmission rate and the posterior probability thresholds that are applied to obtain the specified point on the curves.*



(a)

(b)

(c)

(d)

Figure 3.6: *Plots for the experiment in the section (3.5.3). (a) node degree distribution of the three different underlying network used in the experiment. (b) proportion of the infected nodes in the simulated SIS process over the underlying networks. (c) ,(d) average error, Eq.(3.18), in the inference of the actual links and the non-edge pairs, respectively.*

Chapter 4

A Multilayer Temporal Network Model for STD Spreading [1]

4.1 Introduction

The rising number of infected individuals with sexually transmitted diseases (STD) is a significant concern for public health. It is estimated there are one million new cases of curable STDs acquired each day globally[73]. Indeed, with the increasing trend in online dating, sexual networks become more complex and dynamic. For example, a recent study indicates a relationship between using an online dating application and having had five or more previous sexual partners in young adults[74]. In this chapter we study the effect of casual partnerships in the propagation of STDs. We develop a temporal network model that incorporates the effect of each individual in the sexual network on the spread of STDs. The susceptible-infected-susceptible (SIS) model over a complex network is used for describing the spread of a pathogen in a population with heterogeneous connectivity among individuals.

In the existing literature, we can find several works analyzing the SIS processes over various models of dynamic networks[75–77]. Paré *et al.* analyze the N-intertwined approximation of the SIS process when the adjacency matrix of the network is a deterministic

[1] This chapter is a slightly modified version of our published article[72]

and continuous function of time[78]. Another approach to model the temporariness of partnerships is to adopt the switching network concept. In such a model the contact network randomly switches among a set of predetermined adjacency matrices. In[79;80] the authors have studied sufficient conditions for the stability of the disease-free equilibrium in the SIS spreading model over switching networks. In[81] the authors analyze the initial phase of an SIS epidemic on a network with preventive rewiring. Another class of time-varying network that has been studied in the existing literature is the edge-Markovian networks where the edges appear and disappear following independent Markov processes[82]. In[83], the authors have used an improved effective degree compartmental modeling framework to study the SIS spreading process in the edge-Markovian networks. Ogura et al. consider a generalized version of edge-Markovian model where the inter-event time distribution for the appearance and disappearance of the links is not necessarily exponential[84]. Moreover, they provide a sufficient condition for the exponential stability of the disease-free state in the SIS process that is unfolding on such a time-varying network. A different approach to model time-varying networks is the so-called activity-driven network models that have been studied mostly in the physics literature[85–87]. In these models, the probability that an individual engages in a connection with other individuals is determined by a random variable called "activity". Typically, in a discrete-time activity-driven model (see[85] and, for multiplex networks,[87]), at each time step all the existing links are deleted and each node becomes active according to this activation probability. Then, if active, a node establishes links with randomly selected nodes (active or not) in the population. Certainly, in some practical cases most of the previous models are oversimplifications of the real scenario and it may miss some critical aspects for the infection spreading in a population.

In the next section, we develop a temporal network model and a mean-field type approximation to describe the SIS spreading process over such a temporal network. In sections 4.3, we analyze the disease-free state of the SIS spreading process using the mean-field equations and we find a condition that guarantees the exponential die out of an infection in a time-varying network that can be described via our modeling approach. Finally, using the exact simulation of the process, in section 4.4 we show how the duration of casual partnerships can affect the metastable state of the SIS spreading process.

4.2 The Model

In the following, we first introduce the notation and the assumptions of the two-layer temporal network model, and later, we develop the SIS mean field-equations on this network model.

4.2.1 Two-layer Temporal Network Model

We consider a population of N agents that are connected with two different types of links. The first network layer, $\mathbb{L}_1$, represents steady partnerships among the agents. Independently of the relationship paradigm, one wants to account for (serial monogamy, polygamy, etc.), it is reasonable to expect that $\mathbb{L}_1$ is a quite sparse disconnected network, especially when considering a time scale of interest for SIS-type infections. Besides these permanent links, we assume a second type of links that correspond to potential casual partnerships. These links become active with a probability p_0 only when the agents at both ends of the links are simultaneously seeking casual partners. This second layer of links is denoted by $\mathbb{L}_2$. In general, p_0 can be different for each pair of nodes. However, for clarity of the presentation and since it is straightforward to generalize our result to the heterogeneous case, we assume the same p_0 value for all potential links. By definition, the intersection of the sets of links in $\mathbb{L}_1$ and $\mathbb{L}_2$ is empty. In contrast to the low connectivity of $\mathbb{L}_1$, it is plausible, especially from the advent of match-making apps outside social networks, that $\mathbb{L}_2$ is highly connected. So, we will assume that $\mathbb{L}_2$ is connected (that is, for any pair $\{i, j\}$ of nodes there is a path of links in $\mathbb{L}_2$ joining i and j). While the links in layer $\mathbb{L}_1$ always can transmit infection, a link in layer $\mathbb{L}_2$ transmits the infection only when it becomes an active link. In the model the activation of a potential link in $\mathbb{L}_2$ depends on the activity state of agents at both ends of the link. Apart from the node infection state, we assume that nodes are either active or inactive at any time t. When a node becomes active, it seeks a partner among its active neighbors in

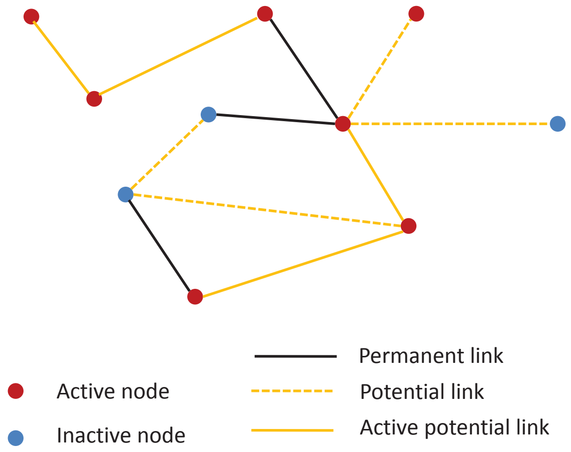


Figure 4.1: *A snapshot from a realization of the network model. At any time t the nodes are either active or inactive. A potential link is activated with probability p_0 if its both ends are active at the same time.*

$\mathbb{L}_2$ and, with a probability p_0, it is established a casual partnership. Later, when one of the two nodes goes to the inactive state, the link is inactivated. This node transition between active and inactive states introduces temporariness in the sexual network. Here, we assume node activation processes are independent Poisson processes, where node i becomes active with rate γ_1^i, and if it is active, it goes to the inactive state with rate γ_2^i. Since the inverse of the transition rate is the expected value of transition time, if node i is active, it is expected to stay active for a period of time of length $(\gamma_2^i)^{-1}$. Thus, when we want to model a node that is frequently activating occasional links, we can assign high values of γ_2 and γ_1 to that node. Moreover, if a node does not participate in casual partnerships —it never becomes active— γ_1 is set equal to zero for that node. Figure 4.1 shows a snapshot of a realization of the temporal network.

Since the inactivation time for each node has an exponential distribution, and the inactivation of a temporal link depends on its both ends, it is straightforward to see that the duration of a casual partnership has an exponential distribution. In fact, a temporal link disappears when one of its ends becomes inactive. Since the minimum of two independent random variables with exponential distributions is distributed exponentially with a rate that is summation of the rates in the independent distributions, it follows that the duration of a temporal link between nodes i and j has an exponential distribution with the rate $\gamma_2^i + \gamma_2^j$. Hence, the expected duration of a casual partnership is $T_{i,j} = (\gamma_2^i + \gamma_2^j)^{-1}$. Moreover, a straightforward calculation shows that, after some initial period of time, the probability to find node i in the active state is $p_2^i = \gamma_1^i/(\gamma_2^i + \gamma_1^i)$. Hence, in the steady state of the process, the probability for the existence of a casual link between the two nodes is $P_{i,j} = p_2^i \; p_2^j \; p_0$.

Temporal Characteristics of the Network Structure

Here, we calculate the probability that any two nodes i, j develop a link, during the period that node i is active.

Based on the description of the model, when node i becomes active, node j is active with probability p_2^j and they develop a link with probability p_0 or the link formation fails with probability $1 - p_0$. Assuming they do not develop link, node j has another possibility to develop link with i, if it becomes inactive and active again before node i becomes inactive. Assuming that the two nodes are active, the probability that node j becomes inactive sooner than node i is $f_1 = \gamma_2^j/(\gamma_2^j + \gamma_2^i)$. This stems from the fact that for any two competing exponential processes A and B, the probability that A will be the minimum is $rate_A/(rate_B + rate_A)$. Now, if we assume that node j became inactive sooner than i, the probability that node j becomes active again before node i goes to the inactive state is $f_2 = \gamma_1^j/(\gamma_1^j + \gamma_2^i)$. In summary, the probability that the two nodes develop a link in a second trial, assuming the first trial fails, is $f_1 f_2 p_0$.

Figure 4.2 depicts the process we described above. Moreover, in that figure we have accounted for the possibility that the node j might be initially inactive when node i becomes active. Hence, accounting for all the possibilities shown in figure 4.2, and allowing for a higher number of link development trials, we can obtain the probability for the establishment of a link between nodes i and j while node i is active as

$$P_{j|i} = p_2^j p_0 \sum_{r=0}^{\infty} (f_1 f_2 (1-p_0))^r + (1-p_2^j) f_2 p_0 \sum_{r=0}^{\infty} (f_1 f_2 (1-p_0))^r = \frac{p_2^j p_0}{1 - f_1 f_2 (1-p_0)} \left(1 + \frac{p_1^j}{p_2^j} f_2 \right), \tag{4.1}$$

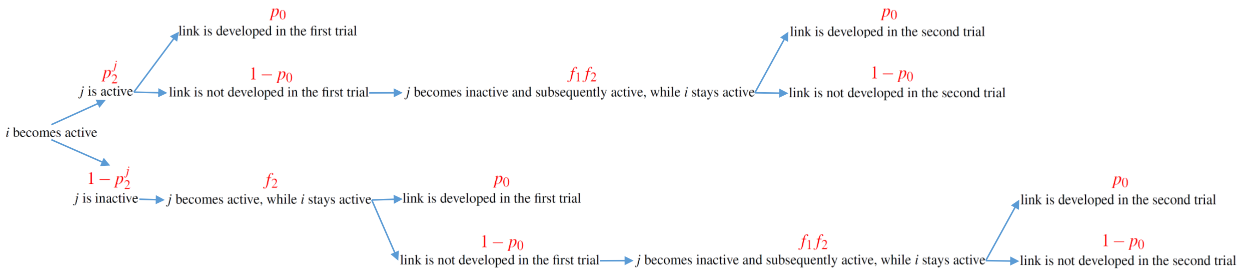
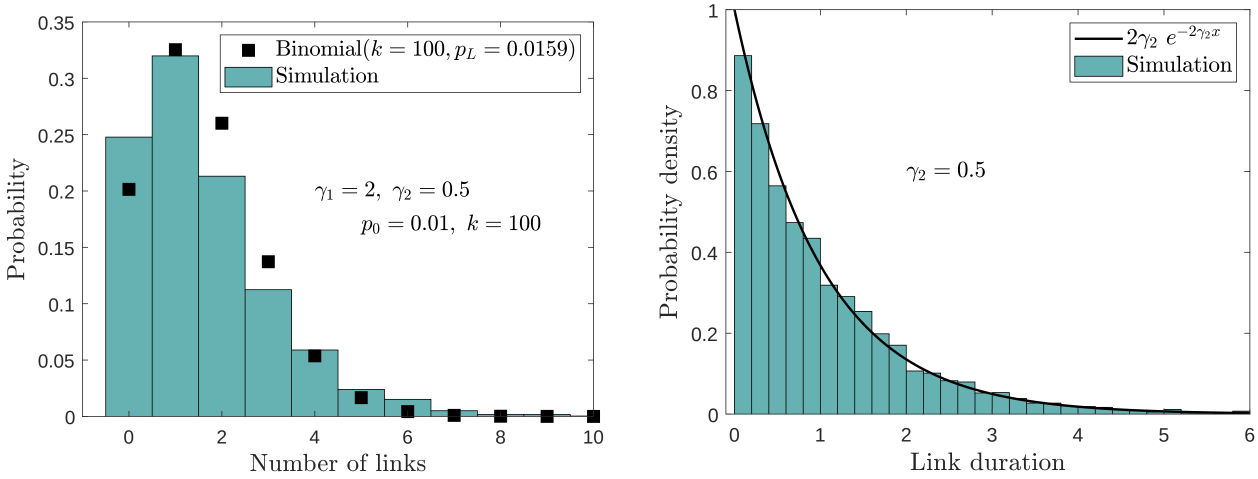

p_0
link is developed in the first trial
p_2^j
j is active
$1-p_0$
link is not developed in the first trial
$f_1 f_2$
j becomes inactive and subsequently active, while i stays active
p_0
link is developed in the second trial
$1-p_0$
link is not developed in the second trial
i becomes active
$1-p_2^j$
j is inactive
f_2
j becomes active, while i stays active
p_0
link is developed in the first trial
$1-p_0$
link is not developed in the first trial
$f_1 f_2$
j becomes inactive and subsequently active, while i stays active
p_0
link is developed in the second trial
$1-p_0$
link is not developed in the second trial

Figure 4.2: *Different processes that result in link establishment between nodes i, j while node i is active. The values in red are the probabilities for each step in the process.*

where $p_1^j = \gamma_2^j/(\gamma_2^j + \gamma_1^j)$ and $p_1^j + p_2^j = 1$. To confirm this result we performed a simulation to find the probability of existence of link between the nodes i, j during the periods that i is active and the probability we obtained from the simulation perfectly matches this theoretical result.

If we assume that node i has k_2 neighbors in the layer $\mathbb{L}_2$ and all these nodes have the same activity rates, γ_1, γ_2, then the distribution of number of developed links, during the period that i is active, follows binomial distribution $P(N_L = r) = \binom{k_2}{r} p_L^r (1-p_L)^{k_2-r}$ where $p_L = P_{j|i}$ is obtained from equation 4.1. Moreover, duration of the links is exponentially distributed with the rate $2\gamma_2$. Figure 4.3a shows the probability mass function of number of developed links obtained from a simulation in which we counted number of established links in each period that node i was active. For this simulation, we assumed that node i has $k_2 = 100$ potential neighbors in the layer $\mathbb{L}_2$. In addition, figure 4.3b shows the distribution of link durations. These figures show the result of simulation follows the expected theoretical distributions.

From the expression for $P_{j|i}$ in equation 4.1 , we can easily determine the average number of links l^i that node i develops every time it becomes active, namely, $l^i = \sum_j P_{j|i}$. As an example, consider a population where each node has k_2 potential links in $\mathbb{L}_2$ and the values of γ_2^i and γ_1^i are the same for all the nodes ($\gamma_2^i = \gamma_2$ and $\gamma_1^i = \gamma_1$ $\forall i$). In this case, the duration of temporal links are distributed exponentially with the average value of $(2\gamma_2)^{-1}$.



(a)

(b)

Figure 4.3: *Distribution of number of developed links (**a**) and their duration (**b**), during the period a node is active. The relevant model parameters are shown in panel (**a**).*

Moreover, the average number of active links for any node i during its active period becomes

$$l^i = \frac{k_2 p_0 p_2 (1 + p_1)}{1 - 0.5 p_2 (1 - p_0)}, \tag{4.2}$$

where we have used that $f_2 = p_2$ when the nodes have the same γ_1 and γ_2.

4.2.2 SIS Epidemics on Two-Layer Temporal Networks

In this section we develop a mean-field type approximation to describe the spreading of infection on the temporal network introduced in section 4.2.1. Next, we discuss the relevance of such an approximation to the exact spreading process.

The susceptible-infected-susceptible (SIS) model has been adopted for studying several STDs because it assumes no immunity after recovering from infection and, hence, it allows for multiple re-infections. This is the case, for instance, for Chlamydia and gonorrhoea where little or no immunity is acquired after infection[88;89]. In this model, each node is either susceptible (S) or infectious (I). We assume the infection and recovery processes are Poisson processes, where an infectious node recovers at a rate δ and transmits the infection to a susceptible neighbor at a rate β. When a susceptible node is in contact with several infectious nodes, it is assumed each infected neighbor acts independently. Thus, the susceptible node contracts the infection with a rate that is the sum of the rates of all the independent infection processes.

Combining the network model and the SIS spreading process, we deduce that each node can assume one of four different states: $\mathcal{S}_1$ susceptible and inactive, $\mathcal{S}_2$ susceptible and active, $\mathcal{I}_1$ infectious and inactive, $\mathcal{I}_2$ infectious and active. If S_1^i, S_2^i, I_1^i and I_2^i represent the probabilities that the node i is in one of the four states in the mean-field approximation, the equations for the time evolution of S_1^i, S_2^i, I_1^i and I_2^i can be written as

$$\dot{S}_1^i = -\gamma_1^i S_1^i + \gamma_2^i S_2^i + \delta I_1^i - \beta \sum_j a_1^{ij} S_1^i (I_1^j + I_2^j), \tag{4.3a}$$

$$\dot{I}_1^i = -\gamma_1^i I_1^i + \gamma_2^i I_2^i - \delta I_1^i + \beta \sum_j a_1^{ij} S_1^i (I_1^j + I_2^j), \tag{4.3b}$$

$$\begin{aligned}\dot{S}_2^i = {} & -\gamma_2^i S_2^i + \gamma_1^i S_1^i + \delta I_2^i - \beta \sum_j a_1^{ij} S_2^i (I_1^j + I_2^j) \\ & - \beta' \sum_j a_2^{ij} S_2^i I_2^j,\end{aligned} \tag{4.3c}$$

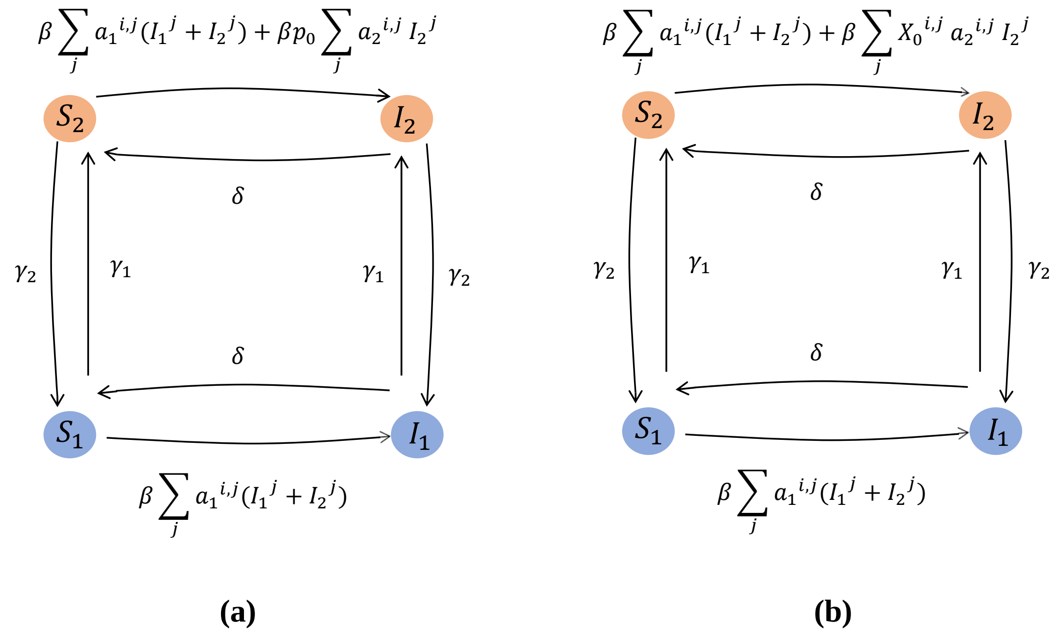
$$\begin{aligned}\dot{I}_2^i = {} & -\gamma_2^i I_2^i + \gamma_1^i I_1^i - \delta I_2^i + \beta \sum_j a_1^{ij} S_2^i (I_1^j + I_2^j) \\ & + \beta' \sum_j a_2^{ij} S_2^i I_2^j,\end{aligned} \tag{4.3d}$$

where $\beta' = p_0 \beta$. In the equations above, a_1^{ij} is an element of the adjacency matrix A_1 for layer $\mathbb{L}_1$ with $a_1^{ij} = 1$, if the nodes i and j form a steady partnership, and $a_1^{ij} = 0$ otherwise. Similarly, a_2^{ij} is the (i, j) element of the adjacency matrix A_2 corresponding to layer $\mathbb{L}_2$. It is important to note that, when p_0 has different values for each pair, we can absorb p_0 in the adjacency matrix A_2 and the elements of the A_2 become the pair-specific probabilities of developing casual partnerships.

The first term on the r.h.s. of equation (4.3a) reflects the fact that the inactive susceptible node i becomes active with a rate γ_1^i and the second term indicates if the node i is in the state $\mathcal{S}_2$ it goes to the inactive state with a rate γ_2^i. The third term originates from the recovering process of inactive infected nodes. In the fourth term, each addend is the multiplication of the probability that the node i is inactive susceptible and the probability that a permanent neighbor of node i is infected.

In equation (4.3c), we take into account the two different sets of neighbors that propagate infection to the active susceptible node i. The fourth term on the r.h.s. arises from the contagion propagation by infectious steady partners of node i. In the fifth term, every summand is the multiplication of the probability that the node i is in the state $\mathcal{S}_2$ and the probability that a potential neighbor of node i in the activity layer $\mathbb{L}_2$ is infectious and also active. When the nodes i and j are active and they are neighbors in $\mathbb{L}_2$, they develop a link with probability p_0. Hence, the summation in the fifth term of this equation is multiplied by p_0.

Equations (4.3) describe approximately the (stochastic) spreading model whose exact mathematical description requires tracking the probability of the system being in any of 4^N possible states which is intractable. Our numerical simulations show these approximate equations lead to nodal infection probabilities that are upper bounds for the infection probabilities in the exact spreading model. In the following, we give an intuitive picture to justify the result of our simulations. Readers familiar with continuous-time Markov chain and the mean-field approximation of SIS process over static one-layer network[90] may recognize that the equations (4.3) are the N-intertwined approximation of a continuous Markov process similar to our model but with a difference. In contrast to the exact description of our model, for this Markov process a link in layer $\mathbb{L}_2$ is activated whenever the nodes at both ends of the link are active with the infection transmission through the link being $\beta' = p_0\beta$ instead of β. Figure 4.4a shows the nodal transitions in the Markov process. Instead, in our model, when both ends of a link are active, the link becomes activated with probability p_0, and transmits infection with rate β if one of the nodes is infected. Figure 4.4b shows the nodal transitions in our model. Our simulations show that the equations (4.3) give an upper bound for the nodal infection probabilities in the Markov process described above which, in turn, are higher than those of our stochastic model. To explain these results we invoke the argument in[28], where the authors show the Markovian SIS process over a static one-layer network is upper bounded



$\beta \sum_j a_1{}^{i,j}(I_1{}^j + I_2{}^j) + \beta p_0 \sum_j a_2{}^{i,j}\, I_2{}^j$
S_2
I_2
δ
γ_2
γ_1
γ_1
γ_2
δ
S_1
I_1
$\beta \sum_j a_1{}^{i,j}(I_1{}^j + I_2{}^j)$
(a)
$\beta \sum_j a_1{}^{i,j}(I_1{}^j + I_2{}^j) + \beta \sum_j X_0{}^{i,j}\, a_2{}^{i,j}\, I_2{}^j$
S_2
I_2
δ
γ_2
γ_1
γ_1
γ_2
δ
S_1
I_1
$\beta \sum_j a_1{}^{i,j}(I_1{}^j + I_2{}^j)$
(b)

Figure 4.4: *The figures show the diagrams of node transitions among different node states. The rate of each transition is specified on the arrow that indicates the transition. (a) shows a diagram of the Markov process which is discussed in section 4.2.2, and (b) shows diagram of the exact process. In these figures $I_1^j = 1$ ($I_2^j = 1$) if node j is infected and inactive (active), otherwise it is zero. In diagram (b) $X_0^{i,j}$ is a Bernoulli random variable that has value one with probability p_0. This random variable is drawn each time a pair of active nodes (i, j) with a potential link between them occurs, regardless of their disease status.*

by the N-intertwined approximation. In fact, equation (4.3b) would be an exact equation for the Markov process if we replace in this equation $S_1^i(I_1^j + I_2^j)$ with $\Pr(x_i = \mathcal{S}_1, x_j = \mathcal{I}_1 \text{ or } \mathcal{I}_2)$, which is the joint probability that node i is inactive and susceptible, and node j is infected. Moreover, since two neighboring nodes can only enhance the infection probabilities of each other and their activity states are independent, the infection states would be non-negatively correlated. In other words, when we know node j is infected, the expectation to observe node i in the susceptible state is less than the case when we do not know the state of node j,

$$\Pr(x_i = \mathcal{S}_1 | x_j = \mathcal{I}_1 \text{ or } \mathcal{I}_2) \leq \Pr(x_i = \mathcal{S}_1).$$

If we rewrite the inequality above as

$$\Pr(x_i = \mathcal{S}_1, x_j = \mathcal{I}_1 \text{ or } \mathcal{I}_2) \leq S_1^i(I_1^j + I_2^j),$$

we can see the addends in equation (4.3b) are upper bounds for the corresponding terms, $\Pr(x_i = \mathcal{S}_1, x_j = \mathcal{I}_1 \text{ or } \mathcal{I}_2)$, in the exact equation for the Markov process. Since these terms appear with positive sign, they only increase the infection probability. Using the same argument about the correlation of nodal infection in equation (4.3d), we expect the N-intertwined approximation in equation (4.3) gives an upper bound for the nodal infection probabilities in the Markov process and our simulations show that it is in fact an upper bound. In order to compare the nodal infection probabilities in the Markov model and the exact description of our stochastic model, consider an instance where at time t_1 one end of an $\mathbb{L}_2$ link is active susceptible while the other end is active infected. If t_2 is the later instant when either the infectious node recovers or one of the nodes becomes inactive, in the Markov process, the probability for transmission of infection through the link is $1 - e^{-p_0\beta(t_2-t_1)}$. But in our model this probability of transmission is $p_0(1 - e^{-\beta(t_2-t_1)})$ which is always smaller. Thus, we expect the infection probabilities in our model will be upper bounded by the probabilities from the Markov process which are in turn smaller than the values obtained from the N-intertwined approximation in equation (4.3). This property of equations (4.3) is particularly useful in controlling the infection spreading. In fact, if any initial infection that is governed by equation (4.3) dies out we know that the infection can not survive in our model.

4.3 Epidemic Threshold

In this section, we analyze the disease-free equilibrium of the SIS spreading equations (4.3), and we find a condition that guarantees the exponential die out of any small initial infection that is introduced in the population. A bifurcation analysis similar to the one in[90;91] shows that, when this condition is not satisfied, there exists another equilibrium state that it is not disease-free. We first present the analysis when the network layers are random regular network, and later we provide the epidemic threshold for a generic network with an arbitrary structure.

Consider the case where $\mathbb{L}_1$ and $\mathbb{L}_2$ are regular random networks of degree k_1 and k_2, respectively. Moreover, let us assume that all the nodes have the same transition rates, i.e. $\gamma_j^i = \gamma_j > 0 \, \forall \, i$ $(j = 1, 2)$. This means that, for any node, the probability of being active is $p_2 = \gamma_1/(\gamma_1 + \gamma_2)$, and similarly for being inactive $(p_1 = \gamma_2/(\gamma_1 + \gamma_2))$. Hence, $S_j^i = p_j - I_j^i$ $(j = 1, 2)$. Introducing this relation in the previous system and summing the equations for the infected nodes in each state, we have

$$\begin{aligned}
\dot{I}_1 &= (\beta k_1 p_1 - (\gamma_1 + \delta))I_1 + (\beta k_1 p_1 + \gamma_2) I_2 \\
&\quad -\beta k_1 \sum_j \left(\sum_i a_1^{ij} I_1^i \right) (I_1^j + I_2^j) \\
\dot{I}_2 &= (\beta k_1 p_2 + \gamma_1) I_1 + (\beta p_2 (k_1 + p_0 k_2) - (\gamma_2 + \delta)) \, I_2 \\
&- \beta \sum_j \left(\sum_i a_1^{ij} I_2^i \right) (I_1^j + I_2^j) - \beta p_0 \sum_j \left(\sum_i a_2^{ij} I_2^i \right) I_2^j,
\end{aligned}$$

where $I_1 = \sum_i I_1^i$ and $I_2 = \sum_i I_2^i$ are the expected number of inactive and active infected nodes, respectively. Let us now approximate the sums $\sum\limits_i a_l^{ij} I_l^i$ by $k_l I_l / N$, which is a good approximation as long as the degree distribution has low variance (as in regular random or Ërdos-Rény networks) and the mean degree is high. Then, after dividing both sides of the equations by N, we have the following system of equations for the disease prevalence

$\rho_j = I_j/N$ in each layer:

$$\begin{aligned}\dot{\rho}_1 &= (\beta k_1 p_1 - (\gamma_1 + \delta))\rho_1 + (\beta k_1 p_1 + \gamma_2)\rho_2 \\ &\quad -\beta k_1 \rho_1(\rho_1 + \rho_2) \end{aligned} \tag{4.4}$$

$$\begin{aligned}\dot{\rho}_2 &= (\beta k_1 p_2 + \gamma_1)\rho_1 + (\beta p_2(k_1 + p_0 k_2) - (\gamma_2 + \delta))\rho_2 \\ &\quad -\beta \rho_2(k_1(\rho_1 + \rho_2) + p_0 k_2 \rho_2). \end{aligned} \tag{4.5}$$

To study the linear stability of the disease-free equilibrium (DFE), we consider the Jacobian matrix of the previous system around the DFE

$$J_0 = \begin{pmatrix} \beta k_1 p_1 - (\gamma_1 + \delta) & \beta k_1 p_1 + \gamma_2 \\ \beta k_1 p_2 + \gamma_1 & \beta p_2(k_1 + p_0 k_2) - (\gamma_2 + \delta) \end{pmatrix}.$$

One can see that the discriminant Δ of the characteristic equation $\det(J_0 - \lambda I) = 0$ is always positive. Precisely, after some algebra and using that $p_1 + p_2 = 1$, we end up with

$$\Delta = (\beta(k_1 - k_2 p_0 p_2) + \gamma_1 + \gamma_2)^2 + 4\beta k_2 p_0 p_2(\gamma_1 + \beta k_1 p_2) > 0,$$

which implies that J_0 has two distinct real eigenvalues $\lambda_1 > \lambda_2$. Therefore, to guarantee that λ_1 traverses 0 when using a tuning parameter of interest, we need that $\text{trace}(J_0) < 0$. Then, the condition for $\lambda_1 = 0$ follows from $\det(J_0) = 0$ which is equivalent to

$$\beta k_2 p_0 p_2(\beta k_1 p_1 - (\gamma_1 + \delta)) = (\beta k_1 - \delta)(\gamma_1 + \gamma_2 + \delta). \tag{4.6}$$

The previous condition defines a polynomial of degree 2 for the critical value of β, β^*. It is easy to see that this equation has two real roots $0 < \beta_1 < \beta_2$. Since we want the value of β for which λ_1 goes from negative to positive, $\beta^* = \beta_1$. Fig. 4.5 shows the dependence of β^* with the transition rate γ_1 obtained by solving the previous equation for $\gamma_1 = \gamma_2$. So, in this figure, the probability p_2 for a node of being active is always $1/2$. However, although a node always spends half of its time with partners in $\mathbb{L}_2$, the figure reveals that how it visits this layer (short and frequent visits or longer but less frequent ones) affects the spread of the disease.

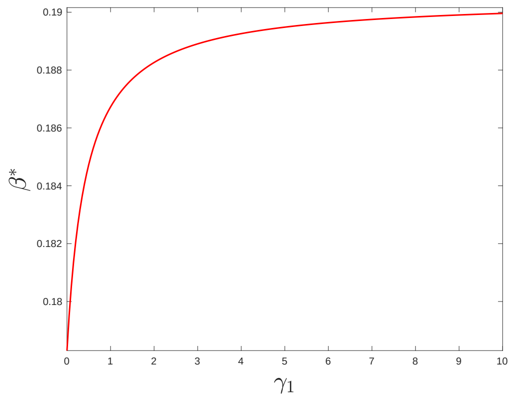


Figure 4.5: *Critical value of β as a function of γ_1 in regular random networks. Parameters: $k_1 = 4$, $k_2 = 50$, $p_0 = 0.1$, $\delta = 1$, $\gamma_2 = \gamma_1$ ($p_2 = 0.5$).*

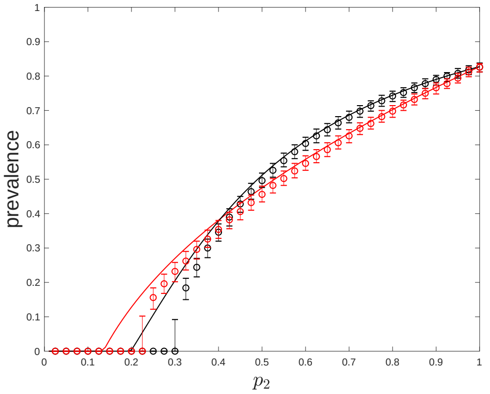


Figure 4.6: *Disease prevalence as a function of p_2 in regular random networks. Circles show, for each set of parameters values, the median of the prevalence* in networks of size 500 after 1000 runs of the Markov process approximated by the mean-field model, and error bars show the corresponding interquartile range. Parameters: $k_1 = 4$, $k_2 = 50$, $p_0 = 0.5$, $\beta = 0.2$, $\delta = 1$, $\gamma_1 = 0.01$ (red), $\gamma_1 = 10$ (black).

A second feature of the mean-field (MF) model is the possibility of having a lower prevalence at the endemic equilibrium for values of γ_1 leading to lower epidemic thresholds. We illustrate this fact in Fig. 4.6 where a bifurcation curve from the DFE is shown using the probability of being active, p_2, as a tuning parameter. Note that, in order to have p_2 as a bifurcation parameter, the epidemic has to die out if inactive individuals alone ($\gamma_1 = 0$, $p_1 = 1$) are not enough to sustain the epidemic, i.e., if $\beta k_1/\delta < 1$. As expected, the higher p_2 is, the higher the prevalence because infection transmission routes in $\mathbb{L}_2$ are used longer. The figure also shows a quite surprising fact: although $\gamma_1 = 0.01$ leads to a lower epidemic threshold in terms of p_2 when compared to that of $\gamma_1 = 10$, it also leads to a lower equilibrium prevalence for $p_2 > 0.37$. We can also observe this feature of the solutions in the output of the simulations over regular random networks of the Markov process corresponding to the mean-field model (see section 4.2.2). These simulations have been performed using the Gillespie algorithm until a final time T=600. Finally, Fig. 4.6 also reveals that the MF model underestimates the epidemic threshold observed from the stochastic simulations. As discussed in section 4.2.2, this is due to the higher infection probabilities assumed under the mean-field approach.

The Disease-Free Equilibrium on General Networks

The analysis of disease-free equilibrium we presented for random regular networks can be generalized for any generic network structure with heterogeneous activity parameters. In this section we focus on the stability analysis of the disease-free equilibrium of the SIS spreading equations 4.3, and we find a condition that guarantees the exponential die out of any small initial infection that is introduced in the population.

For the SIS spreading equations, it is a straightforward observation that the disease-free state given by

$$S_1^i = p_1^i = \frac{\gamma_2^i}{\gamma_2^i + \gamma_1^i}, \quad S_2^i = p_2^i = \frac{\gamma_1^i}{\gamma_2^i + \gamma_1^i}, \quad I_1^i = 0, \quad I_2^i = 0, \tag{4.7}$$

is an equilibrium state. In equation (4.7), p_1^i and p_2^i are the probabilities that node i is active and inactive, respectively, at the steady-state of the continuous-time Markov chain that governs the activity of node i. Here, we study the evolution of the initial infection around the disease-free equilibrium using the corresponding linearized version of SIS spreading equations. In the analysis that comes later, we use state variables $I^i = I_1^i + I_2^i$ and I_2^i instead of

I_1^i, I_2^i. Particularly, this choice of variables directly leads to a relation between the network structure and the model parameters such that, if it is satisfied, the disease-free equilibrium is exponentially stable. If we choose I_1^i, I_2^i, we would need extra algebraic manipulation to get the same relation.

If $\mathbf{I}^i$ and $\mathbf{I}_2^i$ represent small perturbations from the disease-free equilibrium, using the linearized version SIS spreading equations we obtain the following linear dynamical system

$$\dot{\mathbf{I}}^i = -\delta \mathbf{I}^i + \beta \sum_j a_1^{ij} \mathbf{I}^j + \beta' \sum_j a_2^{ij} p_2^i \mathbf{I}_2^j, \tag{4.8a}$$

$$\dot{\mathbf{I}}_2^i = -(\gamma_2^i + \gamma_1^i)\mathbf{I}_2^i + \gamma_1^i \mathbf{I}^i - \delta \mathbf{I}_2^i + \beta \sum_j a_1^{ij} p_2^i \mathbf{I}^j \tag{4.8b}$$

$$+ \beta' \sum_j a_2^{ij} p_2^i \mathbf{I}_2^j,$$

that determines the evolution of the state variables

$$X = (\mathbf{I}^1, \cdots, \mathbf{I}^N, \mathbf{I}_2^1, \cdots, \mathbf{I}_2^N).$$

We can write equations (4.8) as $\dot{X} = JX$ where $J = B - D$ with

$$B = \begin{pmatrix} \beta A_1 & \beta' p_2 A_2 \\ \beta p_2 A_1 + \gamma_1 & \beta' p_2 A_2 \end{pmatrix}, \quad D = \begin{pmatrix} \overline{\delta} & 0 \\ 0 & \overline{\delta} + \gamma_1 + \gamma_2 \end{pmatrix}.$$

In the definition of matrices B and D above, $p_2, \gamma_1, \gamma_2, \overline{\delta}$, are diagonal matrices whose entries are the corresponding parameters for different nodes. It is well known that the linear system is stable if the stability modulus

$$\alpha(J) := \max\{\Re(\lambda) | \lambda \in \text{spectrum of } J\} < 0.$$

In the following, we show there exists a threshold β^* such that for any value of the transmission rate $\beta < \beta^*$ the disease-free equilibrium is exponentially stable, i.e. $\alpha(J) < 0$.

Lemma 1. *If γ_1¿0, $p_0 > 0$ and $\beta > 0$, then the following statements hold:*

a) There is a real eigenvalue of J, denoted by $\lambda_{\max}(J)$, such that any other eigenvalue λ satisfies $\Re(\lambda) \leq \lambda_{\max}(J)$, and the eigenvector Z corresponding to $\lambda_{\max}(J)$ is unique and positive, $Z > 0$.

b) $\min_i \sum_k J_{ik} \leq \lambda_{\max}(J) \leq \max_i \sum_k J_{ik}$

c) If there exists a vector $X \geq 0$ such that $JX \leq \mu X$, then $X > 0$ and $\lambda_{\max}(J) \leq \mu$ with $\lambda_{\max}(J) = \mu$ if and only if X is a multiple of Z.

Proof. Let us prove that the matrix B is irreducible. Consider the associated graph G_B with $2N$ nodes $\{v_1, \ldots, v_N, w_1, \ldots, w_N\}$ such that there is a directed link from node i to node j if and only if $B_{ij} > 0$. As it is well known, B will be irreducible if and only if G_B is strongly connected, that is, for any pair of nodes x, y there is a path of links in G_B from x to y. Recall that, by hypothesis, $\mathbb{L}_2$ is connected and, in consequence (since A_2 is symmetric), strongly connected. Observe that $p_2 > 0$, because $\gamma_1 > 0$. The following facts about the structure of the graph G_B follow from the four blocks defining the matrix B:

A) The block βA_1 implies that the subgraph of G_B induced by the nodes $\{v_1, \ldots, v_N\}$ is isomorphic to $\mathbb{L}_1$

B) The lower right block $\beta' p_2 A_2$ implies that the subgraph of G_B induced by the nodes $\{w_1, \ldots, w_N\}$ is isomorphic to $\mathbb{L}_2$

C) The diagonal entry γ_1 in the block $\beta p_2 A_1 + \gamma_1$ implies that there is a link from w_i to v_i for every $1 \leq i \leq N$

D) The upper right block $\beta' p_2 A_2$ implies that for every $1 \leq k \leq N$ there are links from v_i to some nodes w_k.

Now, to prove that there is a path between any pair of nodes in G_B, we must consider four cases. If the pair has the form $\{v_i, v_j\}$, by (D) there is a link from v_i to some w_k, by (B) there is a path from w_k to w_j (since $\mathbb{L}_2$ is strongly connected) and by (C) there is a link from w_j to v_j. The existence of a path for the three remaining forms of the pair, $\{v_i, w_j\}$, $\{w_i, v_j\}$ and $\{w_i, w_j\}$, follows analogously using (A–D).

From the definition of J, we have $J = B - D$ where B is a nonnegative matrix and D is a nonnegative diagonal matrix. If we assume $\tau = \max_k D_{kk}$ then matrix $C = B - D + \tau I$, with I denoting the identity matrix, is also nonnegative. Since B is irreducible, then C becomes irreducible. Now we can use Perron-Frobenius theorem for non-negative irreducible matrices[92] to show the statements of Lemma 1 hold for the matrix $C = J + \tau I$. Since the eigenvectors of J are similar to the eigenvectors of C and the eigenvalues of J can be obtained by subtracting τ from the eigenvalues of C, we deduce the statements of Lemma 1 also hold for J. □

If we assume β^* is the transmission rate for which $\lambda_{\max}(J_{\beta^*}) = 0$ and $Z_{\beta^*} > 0$ is the corresponding eigenvector, using Lemma 1 it is straightforward to show that for any $\beta < \beta^*$ we have $J_\beta Z_{\beta^*} \leq 0$. Next, we can use the last part of Lemma 1 and conclude $\lambda_{\max}(J_\beta) < 0$. This shows that, if $\beta < \beta^*$, the disease-free equilibrium is exponentially stable. Moreover, to prove the existence of β^*, we can use statement (b) of Lemma 1 and consider the limiting cases $\beta \to 0$ and $\beta \to \infty$ to show that there are β_1 and β_2 such that $\lambda_{\max}(J_{\beta_1}) < 0$ and $\lambda_{\max}(J_{\beta_2}) > 0$. Since $\lambda_{\max}(J_\beta)$ is a continuous function of β there should be a β^* such that $\lambda_{\max}(J_{\beta^*}) = 0$.

In the proof of Lemma 1, the irreducibility condition on B was derived from the positivity of the rates γ_1, p_0 and β and our standing assumption that both layers were connected. If for some reason one wants to relax these assumptions, then one cannot assure that the graph G_B of the proof is strongly connected. In this case, we can separate it into strongly connected components and the threshold analysis which was presented in this section can be done on different components separately. Particularly, for an individual i that never gets active we have $\gamma_1^i = 0$ or equivalently $p_2^i = 0$. In such a case we can see the node that corresponds to I_2^i in the associated graph G_B is disconnected from the rest of nodes and the threshold analysis can be carried out by eliminating the row and column for I_2^i in the J matrix. In fact, if in the matrix J we exclude all those rows and columns that correspond to I_2 for the individuals that never get active we can see the resulting matrix B is irreducible if and only if union of the two layers, $\mathbb{L}_1$ and $\mathbb{L}_2$, is strongly connected.

As we have shown, the threshold value β^* is the smallest transmission rate β for which the eigenvalue problem $J_\beta Z = 0$ has a nontrivial solution. Writing $Z = (Z_1, Z_2)^T$ with $Z_1 = \mathbf{I}$ and $Z_2 = \mathbf{I}_2$, we have $Z_1 = \beta/\delta\,(A_1 Z_1 + p_0 p_2 A_2 Z_2)$ from equaling the r.h.s. of (4.8a) to 0. Then, replacing Z_1 by this expression in the term $\gamma_1^i Z_1^i$ of the r.h.s. of (4.8b) and rearranging terms, we can rewrite the eigenvalue problem as $\tau B^\star Z = Z$, where

$$B^\star = \begin{pmatrix} A_1 & p_0 p_2 A_2 \\ p_2 A_1 & p_0 p_2^\star A_2 \end{pmatrix}, \tag{4.9}$$

$\tau = \beta/\delta$ is the so-called effective spreading rate[10], and $p_2^\star$ is a diagonal matrix such that

$$(p_2^\star)_{i,i} = p_2^i \frac{1 - p_2^i + \overline{\gamma}_2^i p_2^i}{1 - p_2^i + \overline{\gamma}_2^i}$$

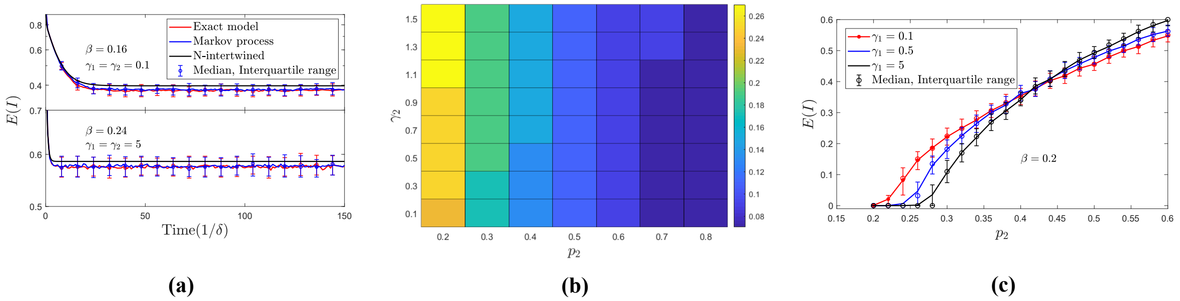
with $\overline{\gamma}_2^i = \gamma_2^i/\delta$. From the expression $\tau B^\star Z = Z$, we can find the threshold value β^* from

$$\tau^* = \frac{\beta^*}{\delta} = \lambda_{max}^{-1}(B^*). \tag{4.10}$$

We can see this threshold depends not only on p_2 and p_0 but also on γ_2. Hence, it captures the effect of the probability for the existence of a casual link and its duration. Indeed, in section 4.2.1 we explained that the probability for the existence of a temporal link between any nodes i, j (casual partnership) in the steady state is $p_2^i p_2^j p_0$ and the expected duration of the link is $(\gamma_2^i + \gamma_2^j)^{-1}$. Thus, this threshold value is different from that of a static network with a link of weight $p_2^i p_2^j p_0$ between nodes i and j.

4.4 Simulations

In the following, we present the results from the simulation of the exact process (Fig. 4.4b), which we described in section 4.2.2, over random regular networks. These simulations clarify



(a) **(b)** **(c)**

Figure 4.7: *Results of numerical and stochastic simulations of the spreading processes on random regular graphs, discussed in section 4.4. Panel (a) shows the comparison of different approximate processes with the exact process; panel (b) shows the epidemic threshold of the exact process, as a function of p_2 (probability of being active in $\mathbb{L}_2$) and the parameter γ_2, which is proportional to the inverse of expected duration of active potential links; panel (c) shows how the infection prevalence in the metastable state is affected by different parameters in the exact process. Error bars show the median and the interquartile range.*

the relation between the exact process and its approximating counterpart, which are the N-intertwined equations and the stochastic Markov process (Fig. 4.4a) we described in section 4.2.2. In addition, using the simulations we explore effect of the model parameters on the infection spreading.

As usual in the setting of continuous-time stochastic simulations, we use the well-known Gillespie algorithm[93] in all experiments. All random regular networks are generated using the configuration model algorithm[94]. We run this algorithm twice and independently to get layers $\mathbb{L}_1$ and $\mathbb{L}_2'$. To get the empty intersection condition, we extract $\mathbb{L}_2$ from $\mathbb{L}_2'$ by deleting the links that are common to $\mathbb{L}_1$. When the respective prescribed degrees k_1 and k_2 are small with respect to the number N of nodes, this happens with very small probability, and the obtained graph $\mathbb{L}_2$ has a mean degree very close to k_2. The reported experimental prevalence values are always computed by averaging over several hundreds of independent realizations of the stochastic process, each corresponding to a particular random initial condition with a fixed number of infected individuals (see the captions of figures for more details on each experiment). Since the average prevalence does not show the distribution of prevalence across independent simulations, we calculated the median and the interquartile range and show them as error bars in the figures.

Here we assumed a population of 500 nodes, and for the layer $\mathbb{L}_1$, we generated a random regular network where each node has four neighbors, while for the layer $\mathbb{L}_2$ we used a random regular network of degree 50. Moreover, for the layer $\mathbb{L}_2$ we have assumed $p_0 = 0.5$. In these simulations, to estimate the expected number of infected nodes in the Markov process and the exact process, we generated 200 independent realizations and calculated the average, median and interquartile range for the total number of infected nodes over these realizations at different times. Moreover, we assumed that all the nodes were infected and active at $t = 0$. Figure 4.7a shows the infection prevalence curves obtained from the N-intertwined approximation, the Markov process, and the exact spreading process. As we expect, the N-intertwined equations provide an upper bound for the prevalence values obtained from the Markov process and the exact process.

4.4.1 Impact of Partnership Duration on the Epidemic

In Fig. 4.7b, we have shown the epidemic threshold, obtained from the simulation of the exact process, as a function of the probability of being active, p_2, and γ_2, which is proportional to the inverse of the expected duration of active links. From this figure, we can see that, when p_2 increases, the epidemic threshold decreases. However, when the number of active nodes is small (lower values of p_2), the epidemic threshold increases as the duration of links decreases (larger values of γ_2). This effect of the link duration in $\mathbb{L}_2$ on the epidemic threshold is also clear from Fig. 4.7c. In this figure, we see that reaching the metastable state requires less active nodes (smaller p_2) when the link duration is longer (note that, for a given p_2, γ_2 is proportional to γ_1).

Figs. 4.7c and 4.7b show that the epidemic threshold decreases with increasing link duration and this result can appear counter-intuitive, especially when we consider sexual networks. For instance, one can conclude that, due to the higher epidemic threshold, a population in which the sexual behavior is dominated by short-duration casual partnerships is less vulnerable to epidemics, compared to a population with longer-duration casual partnerships. We can interpret this result thinking about two counterbalancing processes; the probability of transmission during the partnership duration, and the frequency of changing partners. With short-duration casual partnerships, the number of partners in a given interval is greater but the probability of transmission is smaller; with long-duration casual partnerships, the number of partners is smaller but the probability of disease transmission is higher. Our numerical simulations show that these two processes do not obtain a complete balance; rather, the duration of the partnership plays a more important role than the number of partners. For this reason, the threshold is increased by short-duration partnerships in spite of the increased number of partners. If this result seems counterintuitive we need to keep in mind that they are obtained by keeping a constant value for the infection transmission rate.

To understand the role of infection transmission rate in explaining our results, we can compare the two scenarios (long-duration casual partnership and short-duration casual partnership) keeping the same probability of infection transmission per sexual intercourse, instead of the same infection probability per unit of time. In this case, we impose a similar average number of sexual intercourse in these two scenarios, which in turn corresponds to different infection rates of the disease. To understand the difference, assume a population where the average number of sexual intercourse for an individual in a long-term casual partnership is once per week, and the probability of infection transmission in an intercourse with an infected individual is 0.5. Consequently, the infection transmission rate in partnerships with duration significantly larger than a week can be estimated as $\beta_p = -\ln(0.5)/7 = 0.1 \text{ day}^{-1}$. Indeed, we have set the value of β_p such that, if a partner is infected, the probability that the susceptible partner stays susceptible is multiplied by one half each week during an infection period of length T days; in other words, $e^{-\beta_p T} = (0.5)^{T/7}$. In contrast, assuming the same probability of infection during an intercourse with an infected individual, for short-duration partnerships with one intercourse, we can assign the transmission rate by solving

$$\int_0^\infty \alpha^{-1} e^{-t/\alpha}(1 - e^{-\beta t})\, dt = 0.5, \tag{4.11}$$

for β, while for consistency in the mathematical modeling, we can assume a duration of $\alpha = 0.5$ day for short-duration partnerships. In the equation above, $\alpha^{-1}e^{-t/\alpha}$ is the probability

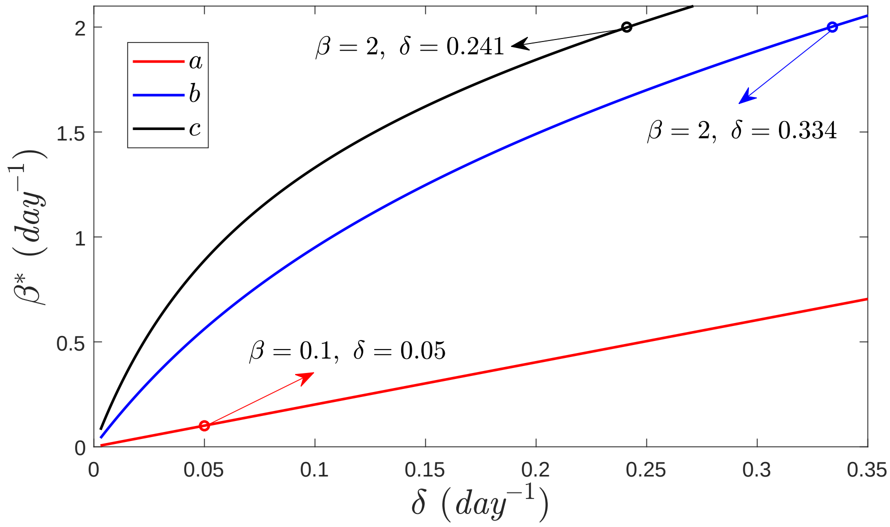


Figure 4.8: *Infection transmission rate threshold as a function of the recovery rate for three different temporal networks discussed in section. Case **a** corresponds to partnerships of 60 days duration and cases **b**, **c** correspond to casual sexual encounters.*

density for a link with duration t, and $1-e^{-\beta t}$ is the probability for transmission of infection from an infected node within the period that the link exists. Thus, the integral in equation 4.11 gives the expected value of transmission probability, which is easily computed. A simple expression for β follows from equation 4.11, namely, $\beta = \alpha^{-1}$. Therefore, the equivalent transmission rate for a short-duration partnership becomes $\beta_e = 2$ day^{-1}.

In order to compare the vulnerability of populations with different duration of partnership under this viewpoint, in Fig. 4.8 we have plotted the epidemic threshold, obtained from equation 4.6, corresponding to three different sets of model parameters,

a. $\gamma_2^{-1} = 120$ days, $\gamma_1^{-1} = 1$ day, $l = 1$, $k_1 = 0$

b. $\gamma_2^{-1} = 1$ day, $\gamma_1^{-1} = 6$ days, $l = 1$, $k_1 = 0$

c. $\gamma_2^{-1} = 1$ day, $\gamma_1^{-1} = 13$ days, $l = 1$, $k_1 = 0$

Among these selections of parameters, case **a** corresponds to partnerships with an average duration of $(2\gamma_2)^{-1} = 60$ days. Moreover, using equation 4.2, we set $k_2 p_0$ such that the average number of links in each activity period is $l = 1$. For simplicity, we assumed that there is no static links, $k_1 = 0$. Contrarily, cases **b** and **c** correspond to sexual encounters with frequencies that are once per week or two weeks, respectively. Indeed, case **c** is more comparable to case **a** because they provide a similar average number of sexual intercourses per year.

From the curve for case **a**, in Fig. 4.8 we conclude that the epidemic threshold β^* is greater than the estimated value of transmission rate in partnership, $\beta_p = 0.1$, when the average recovery rate, δ, is greater than 0.05 day^{-1} or, equivalently, for the expected average recovery time of $\delta^{-1} < 20$ days. Hence, the infection dies out when the recovery time is smaller than 20 days. On the other hand, for case **c**, which corresponds to sexual encounters with the frequency of once per two weeks, the transmission rate $\beta_e = 2$ is smaller than the epidemic threshold only for the expected recovery time $\delta^{-1} < (0.241)^{-1} \approxeq 4$ days. This suggests that the population with sexual encounter behavior is more vulnerable than the population with partnership behavior.

4.5 Summary

In this chapter, we have developed a network model that incorporates the process of switching between two network layers –steady and casual partners– driven by individual activities, which define the propensity of individuals to be engaged in casual partnerships. Hence, the temporal characteristic of the model appears as a consequence of the existence of such partnerships. This scenario is suitable for studying the dynamics of sexually transmitted diseases in real communities where casual partners are not always disclosed in partner-notification programs. These partnerships are modeled by considering the activation of links drawn from a set of potential links. Each of these links is activated with probability p_0 when the nodes at both ends are willing to develop a casual partnership, i.e., when both nodes are active.

The model incorporates two ingredients, namely, change of partners and partnership duration, that have also been discussed in pair formation models. The contribution of casual partnerships to the disease spread by increasing the basic reproduction number has already been highlighted in these models (see, for instance, Eq. (5) in[95] and Eq. (3) in[96]). Here, our model allows us to assess the role of the layer of casual partnerships by quantifying its utilization through the activity probability parameter p_2, and by considering the mean duration of partnerships by means of γ_2. In particular, we have studied how different parameters of the model affect the epidemic threshold and the disease prevalence in the metastable state –endemic equilibrium of the mean-field model. We have found that, given a fixed value of the infection transmission rate β, the prevalence of infection strongly depends on the utilization of the layer of casual partnerships and on the duration of these partnerships. Our simulations show that the epidemic threshold decreases with increasing link duration, while short partnership durations decrease the probability of disease transmission, thus increasing the threshold. Finally, and without contradiction, our analysis shows that casual sexual behavior, which implies extremely short partnerships, makes the population vulnerable to the infection spreading.

Chapter 5

Delocalized SIS Spreading [1]

5.1 Introduction

Despite the simple description of the SIS process, only a few *exact* results about the SIS process on a generic graph G have been proposed. In the SIS process, the disease–free state is an absorbing state, i.e., any initial infection will ultimately die out regardless of the infection rate[90;98]. The extinction time depends on the structure of the network, the infection rate β, curing rate δ, and the initial infection. For this model, Ganesh *et al.*[98] rigorously proved that any initial infection dies out exponentially in time if the infection strength, $\tau \triangleq \beta/\delta$, is smaller than the inverse of the spectral radius of the graph $\rho(G)$ ($\rho(G)$ is the largest eigenvalue of the adjacency Matrix for the graph). However, for the values of τ larger than $1/\rho(G)$, the process may reach *metastability* where the extinction time is exponentially long with respect to the population size and the process stays in a state that resembles equilibrium[99].

A true epidemic outbreak concerns network–wide invasion of the contact graph rather than localized infection of certain sites within the contact network. This argument leads us to the concept of infection *localization* in the SIS model over a generic graph. To illustrate this phenomenon, consider the Line–Clique graph in Fig.(5.1) consisting of two subgraphs,

[1] This chapter is a slightly modified version of our published article[97], Copyright © 2016, IEEE.

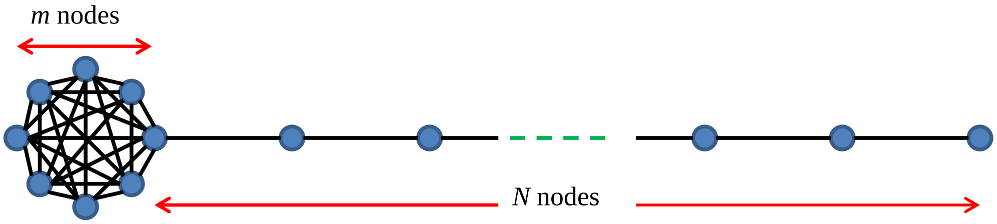


Figure 5.1: *The Line-Clique graph consisting of a complete graph of size m and a line graph of size $N >> m$. It is possible to observe a metastable state where infections mostly localize on the clique part — a tiny portion of the network.*

the clique part of size m and the line part with size $N >> m$. The spectral radii of the clique part and the line part separately are $m - 1$ and ~ 2, respectively. However, the spectral radius of the Line–Clique graph is close to that of the clique subgraph. For such a graph, any infection dies out exponentially in time as long as the infection strength τ is smaller than $1/(m - 1)$. But, what does happen if $\tau > 1/(m - 1)$? We know for $\tau \leq 1/2$, the line subgraph, considered separately, cannot sustain infections for a long time. The argument above leads to the speculation that for $1/(m - 1) \leq \tau \leq 1/2$, if the Line–Clique network with $N >> m$ reaches a metastable state, the infection should be mostly localized on the clique part of the network.

Localization of SIS process has recently been reported in the literature. Goltsev *et al.*[52] studied the steady–state solution of the mean–field approximated SIS model for τ close to $1/\rho(G)$, where the equilibrium solution is proportional to the dominant eigenvector of the contact network adjacency matrix. The major drawback of such approaches[52;100;101] is that they fully rely on approximate models in a region where they are least accurate. Mean–field models perform more accurately for large values of τ and homogeneous networks, while they can perform very poorly at steady–state and for τ close to $1/\rho(G)$.

Here, we propose a dispersion measure based on Kullback–Leibler divergence[102] that quantifies how the marginal probability of infection is far from a homogeneous spread over the nodes of the network. We show that by formulating a maximum entropy problem, we can find an upper bound for the dispersion entropy of the possible metastable state. As a result, any initial infection over the network either dies out or reaches a metastable state that has lower entropy than the upper bound. Unlike existing studies, our investigation of epidemic localization does not use mean-field approximation of the SIS process and is based on exact equations arguments. convex optimization techniques allow for efficient solution of the maximum entropy problem even for large networks. Numerous Monte Carlo simulations of the SIS model support our results.

5.2 Method

In section 1.4 we have described the SIS spreading process over networks. Equation (1.1) in that section describes the evolution of the expected value of random variable x_i, denoted by $E(x_i)$. Indeed, $E(x_i)$ is equivalent to the marginal infection probability of node i, which we denote it by p_i. Similarly, $E[x_i x_j]$ is equivalent to the joint infection probability of nodes i and j.

In the SIS process, summation of marginal infection probabilities $\sum_i^N p_i$ (N is the number of nodes in the network) provides a descriptor for the expected size of the epidemic. However, this measure does not provide any information on how the infection is distributed among the nodes. In order to study the dispersion of infection, regardless of its size, first we normalize infection probabilities and define $\overline{p}_i \triangleq p_i / \sum_i^N p_i$. Since $\sum_i^N \overline{p}_i = 1$, we propose to treat $\overline{P} = [\overline{p}_1, ..., \overline{p}_N]^T$ as a probability distribution and then utilize concept of distance between distributions to quantify the distance between $\overline{P}$ and uniform distribution $U = [\frac{1}{N}, ..., \frac{1}{N}]^T$. In particular, we use Kullback-Leibler divergence[102] which is

$$D_{KL}(\overline{P}||U) = -\sum_i^N \overline{p}_i \ln \overline{p}_i - \ln(N). \tag{5.1}$$

Therefore, to study the degree of delocalization we use entropy, which is the variable term on the r.h.s of equation above,

$$S(P) = -\sum_i^N \overline{p}_i \ln \overline{p}_i, \tag{5.2}$$

The defined entropy reaches its maximum, $\ln(N)$, when all the nodes have the same none-zero probability of infection and this leads to $D_{KL}(\overline{P}||U) = 0$.

We are particularly interested in study of infection delocalization in the metastable state. Metastable state resembles an equilibrium where the infection probability of each node stays (almost) constant, i.e., $dP(t)/dt \to 0$. To begin with, we use a simple observation from the exact SIS equations (1.1) that $dp_i/dt \leq \beta \sum a_{ij}p_j - \delta p_i$, which is due to the fact that $E[x_i x_j]$ is nonnegative. Therefore, when meta-stability is achieved, we must have

$$(\beta A - \delta I)P \geq 0, \tag{5.3}$$

where $P = [p_1, ..., p_2]^T$ and A is the adjacency matrix of the graph. Assuming condition (5.3) holds for the metastable state, in the next step we will find the most delocalized distribution that satisfies the condition.

Theorem 1. *Assuming a contact graph G, infection strength $\tau = \beta/\delta$ and initial infection probability $P(0)$, if a metastable state is achieved, the dispersion entropy of the metastable state is upper-bounded by S^* which is the solution of the following maximum entropy problem:*

$$\begin{aligned} &\mathit{maximize}: && S = -\textstyle\sum_i^N p_i \ln p_i, \\ &\mathit{subject\ to}: && (\tau A - I)P \geq 0, \\ & && \textstyle\sum_i^N p_i = 1, \\ & && P > 0. \end{aligned}$$

Proof. The solution to the above optimization problem maximizes the entropy defined in Eq. (5.2) because, instead of normalization of the probabilities, $\sum_i^N p_i = 1$ has been added to the constraints set and inequality (5.3) is linear which is not altered by scaling P. □

The maximum entropy problem can be solved efficiently for large network sizes using convex optimization tools such as CVX package[103].

Lemma 2. *If $\tau < 1/\rho(G)$, there does not exist any P that satisfies condition (5.3). Furthermore, for $\tau \geqslant 1/\rho(G)$, the constraint of Theorem 2 has a non-empty feasible set.*

Proof. Any feasible probability distribution $P > 0$ that satisfies condition (5.3) satisfies

$$P^T(\tau A - I)P \geq 0. \tag{5.4}$$

However, if $\tau < 1/\rho(G)$, matrix $(\tau A - I)$ is a negative definite matrix which cannot allow (5.4). Therefore, if $\tau < 1/\rho(G)$, there does not exist any P that satisfies condition (5.3). On the other hand, for $\tau \geqslant 1/\rho(G)$, the dominant eigenvector of A, i.e., $P = \frac{1}{||\boldsymbol{x}_1(A)||_1}\boldsymbol{x}_1(A)$, is always feasible. □

We would like to remark the existence of a distribution with a high value of dispersion entropy that satisfies condition (5.3) does not indicate the existence of a metastable state. However, *if there exists a metastable state*, our analysis assigns an upper bound to its dispersion entropy. Therefore, if the optimization problem yields a small value for entropy, the infection does not invade a large number of nodes in the metastable state, hence providing a *sufficient condition for either complete extinction of infections or their localized persistence.*

Moreover, if $\tau \downarrow 1/\rho(G)$, Lemma 2 indicates the feasible space of optimization problem is a small neighborhood including the dominant eigenvector of the adjacency matrix $\boldsymbol{x}_1(A)$, which makes $S^* \simeq S(\boldsymbol{x}_1(A))$. In this case, the results of our analysis are compatible with those of[52], except we use a different measure for localization. However, for higher values of τ, our analysis can still provide an upper bound for the delocalization of SIS process; while an analysis based on the mean-field approximation does not necessarily characterize the infection delocalization in exact SIS process.

5.3 Numerical Result

Considering the example graph depicted in Fig.(5.1), we generated a Line-Clique graph with 280 nodes in the line subgraph and 40 nodes in the clique subgraph. We used a convex

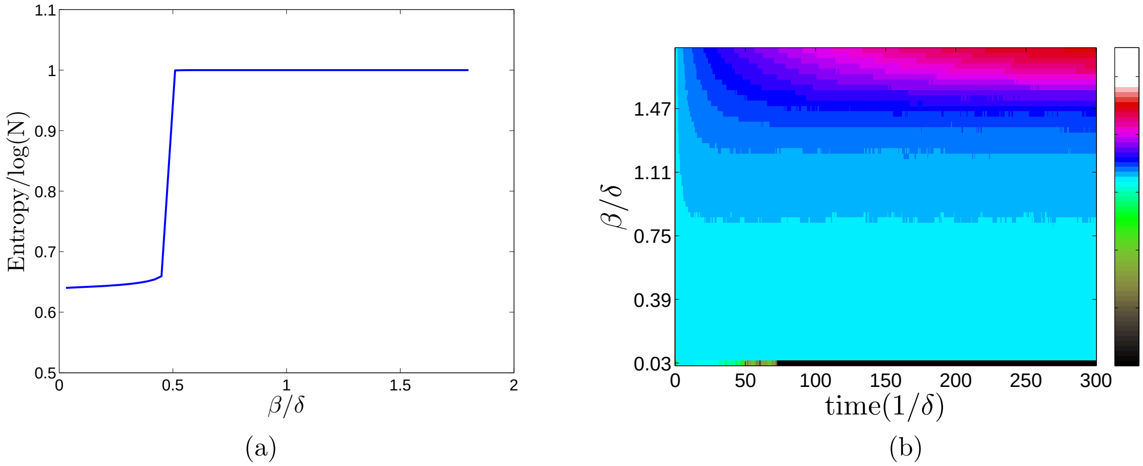


Figure 5.2: *(a) The entropy of the optimized distribution for the Line-Clique graph in Fig. 5.1. As can be seen, there is a sudden jump at $\tau = \frac{1}{2}$. (b) Monte Carlo simulation of the SIS model over the Line-Clique graph. Color represents dispersion entropy of infection probability distribution divided by* $\ln(N)$.

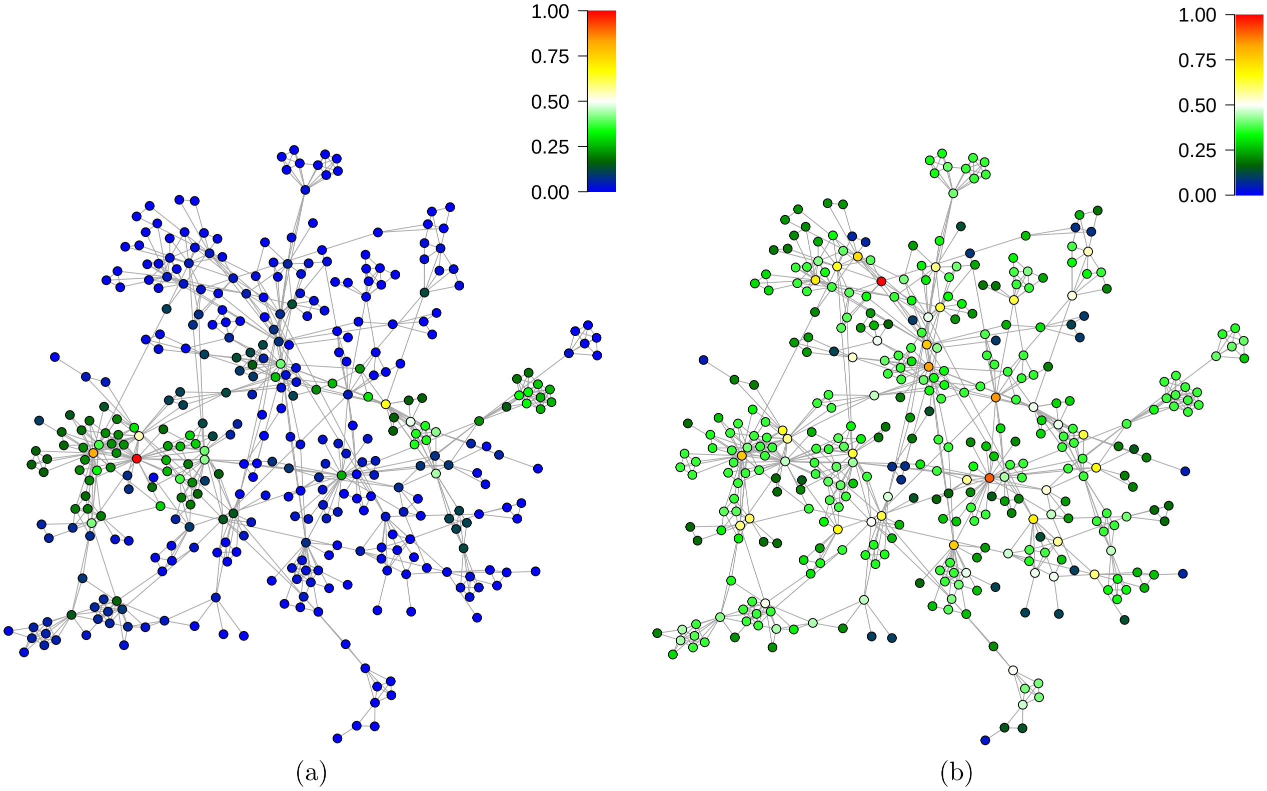
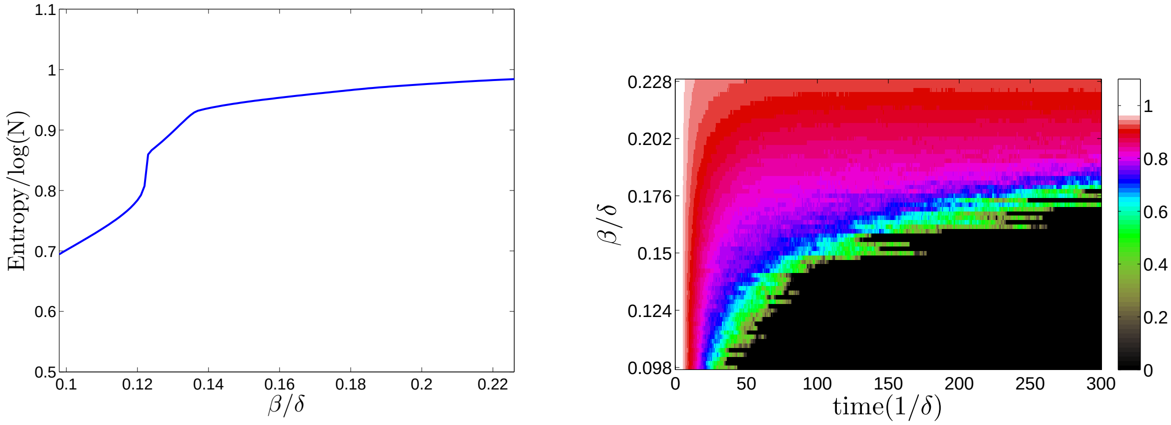


Figure 5.3: *(a) Optimized probability distribution for $\beta/\delta = 0.125$, showing only a few localized sites of the network have active nodes (b) Optimized probability distribution for* $\beta/\delta = 0.23$.

optimization package[103] and found a distribution with maximum entropy for different values of infection strength. The result is plotted in Fig.(5.2a). As we can see for $\beta/\delta < \frac{1}{2}$, the optimized entropy is smaller than the entropy of homogeneous distribution which is $\ln(N)$. Therefore, for $\beta/\delta < \frac{1}{2}$, if the epidemic reaches metastability, the infection will not spread to the whole nodes of the network. On the other hand, for $\beta/\delta > \frac{1}{2}$, the optimized entropy is close to $\ln(N)$. In this case, the solution of the optimization problem gives the trivial upper bound. Moreover, to show the relation between the optimized entropy and the true entropy of infection evolving over time, we performed the Monte Carlo simulation of the SIS model using GEMF-Tool package[19;38] over the line-clique graph. For this simulation, we assumed an initial condition where only one node in the clique subgraph was infected. Fig. (5.2b) shows the result of the simulation. In this figure, color represents the entropy of infection divided by $\ln(N)$. As we can see in the Fig. (5.2b), for different $\beta/\delta < \frac{1}{2}$, the dispersion entropy of infection grows very fast and stays constant with values less than the optimized entropy. In this case, the epidemic reaches meta-stability but is localized on the clique subgraph.

As another example, we chose the largest component of a coauthorship network from[39] as shown in Fig. (5.3). For this network, the spectral radius of the adjacency matrix is about 10.4. The entropy of the optimized distribution, shown in Fig. (5.4a), is an upper bound for the metastable state of SIS model over the network. Although for $0.13 < \beta/\delta < 0.15$ the optimized entropy has a considerable value, we cannot predict the existence of a metastable state. In fact, the result of Monte Carlo simulation (Fig. (5.4b)) for SIS model over the network shows the metastable state starts at much higher values of β/δ (larger than 0.2) where the optimized entropy is almost $\ln(N)$.

Moreover, as an illustration for the relation between the entropy of a distribution and delocalization of the distribution, we plotted the network and colored the nodes based on the value of its probability in the optimized distribution. In Fig. (5.3a,5.3b) , the optimized distribution for two different values of β/δ is plotted. For $\beta/\delta = 0.125$, where the optimized entropy is small, the distribution is mainly localized on a few nodes. On the other hand when β/δ increases to 0.23, the entropy of the optimized distributions increases and more nodes get involved.



(a)

(b)

Figure 5.4: *(a) The entropy of the optimized distribution normalized by* $\ln(N)$ *for coauthorships network of Fig. 5.3. (b) Monte Carlo simulation for the SIS where all the nodes were initially infected. Color represents dispersion entropy of the infection probability distribution divided by* $\ln(N)$

5.4 Summary

In summary, we investigated the infection localization of SIS process. We used dispersion entropy defined in Eq.(5.2) as a measure of delocalization. We believe, in addition to infection size, measures such as dispersion entropy are relevant in epidemic spreading processes and should be included in numerical simulations. Moreover, we find an upper bound for the infection dispersion entropy when a metastable state exist. This upper bound which depends on the infection strength suggests the maximum number of nodes that can be active in a metastable state. A small upper bound for the dispersion entropy of a metastable state provides a sufficient condition for either complete extinction of infections or their localized persistence.

Chapter 6

Interpolation of Networked Spreading Data

6.1 Introduction

In this chapter we compare two different heuristic methods to interpolate a target function, f, defined over the nodes of a network, i.e., $f : V \to \mathbb{R}$ where V is the set of network node. We assume the value of target function, f, is known for some of the nodes in the network, and by interpolation, we mean the estimation of this function over those nodes for which the value of f is unknown. Here, we apply the heuristic methods to a specific target function, which is an outcome of an SIS stochastic spreading process.

In the SIS stochastic spreading process, we define the infection prevalence at time t as the expected number of infected individuals in the network at time t. Our simulations for the SIS process show that the value of prevalence depends on the node where the initial infection starts. Here, we define the target function f such that the value of the function at any node n is the prevalence at some specific time, after the infection, started from node n, spreads through the network. Indeed, we use this function as an example to compare the heuristic methods, and in the future, we plan to perform similar studies on other functions defined on networks.

6.2 Methods

In general, an interpolation method relies on the assumption that the nodes which are close to each other have similar values. However, for any two nodes in a network, we can define different closeness measures, such as the weight of the link between the nodes (if there is no link, the weight is considered to be zero), the length of the shortest path or the effective resistance distance between them. In designing an interpolation method, one factor that we need to consider is the distance measure that we use to quantify closeness. It is reasonable to expect that the choice of the distance measure affects the accuracy of the interpolation. For instance, an interpolation method that uses the function values of adjacent nodes to estimate the value of the unknown node relies on local information, and it might not capture global information.

6.2.1 Energy Minimization

The first interpolation method which we are going to employ is from reference[104]. This algorithm assumes that the value of function $f : V \to \mathbb{R}$, defined over the graph $G = (V, E)$, is known only for a subset of nodes, $V_l \subset V$. For the rest of the nodes , $V_u = V - V_l$, value f is estimated by minimizing the quadratic energy function

$$E(f) = \sum_{i,j} w_{i,j}(f(i) - f(j))^2,$$

where $w_{i,j}$ is the weight of the link between node i, j. In this algorithm, for the nodes in V_l, f is constrained to the known values of the nodes. For the nodes in V_u we determine values of$f$ such that $E(f)$ is minimized. It can be shown that the unknown value of f is the average of f at neighboring nodes[104],

$$f(j) = \frac{1}{\sum_i w_{i,j}} \sum w_{i,j} f(i) \quad \text{for } j \in V_u.$$

This implies a notion of smoothness for f that assumes neighboring nodes have similar values of the target function f. To compute the minimization solution, we split the weight matrix $W = [w_{i,j}]$ into 4 blocks,

$$\begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix},$$

where W_{ll} is the weight matrix over the set V_l, and the elements of W_{lu} are the weight of links between nodes in V_l and V_u. A similar definition holds for W_{ul} and W_{uu}. If f_u and f_l denote the values of f over the sets V_u and V_l, respectively, from the energy minimization we obtain[104],

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l.$$

In the equation above $D_{uu} = \mathrm{diag}(d_i)$ is a diagonal matrix with entries that are the degrees, $d_i = \sum_j w_{i,j}$, of nodes in the set V_u.

6.2.2 Effective resistance

Next, we propose a novel interpolation method that relies on the notion of effective resistance distance between the nodes in a graph. We use effective resistance distance as the measure of similarity because it is a global measure over a graph, and it accounts for all paths connecting a node pair in addition to the length of the paths. Indeed, for a graph of N nodes, we can map the nodes to an $N-1$ dimensional Euclidean space through eigenvectors of the Laplacian matrix, and the effective resistance distance becomes the distance between the nodes in the embedding space. The Laplacian of a graph is defined by $L = D - W$, where W is the adjacency matrix, and D is a diagonal matrix with diagonal elements equal to the degree of nodes. If we denote the orthogonal eigenvectors of the Laplacian matrix by u_i, $(i = 1, \cdots, N)$, and the corresponding eigenvalues by λ_i, the coordinates of node n in the embedding space are[105]

$$\left(\frac{u_2(n)}{\sqrt{\lambda_2}}, \cdots, \frac{u_N(n)}{\sqrt{\lambda_n}} \right).$$

While designing the interpolation method, we assume the function defined over the nodes is smooth in the embedding space, i.e, nodes that are close to each other have similar values.

One method of interpolation that we can use to estimate the unknown values f is the radial basis function. In this method, f in the embedding space is estimated by

$$f^e(x) = \sum_k a_k \varphi(\|x - x_k\|).$$

In the equation above, the summation is over the training nodes for which we know the value of f. The vector x_k represents the set of training node coordinates in the embedding space. φ is a radial basis function, such as a Gaussian function, $\|x - x_k\|$ is the distance between two points in the embedding space, and $f^e(x)$ is the estimation of the target function for the point x. a_k is the weight associated to each function $\varphi(\|x - x_k\|)$, and can be obtained by minimizing the total error, $\sum_k (f(x_k) - f^e(x_k))^2$, calculated using the known values of function f for the training nodes.

Another method of estimation that takes effective resistance distances as input, uses a feedforward neural network method. Figure 6.1 illustrates the type of neural network we use in the estimation. In this approach, we want to estimate the value of function f at a point x in the embedding space. We first use the effective resistance distances between x and all the training points x_k in the Laplacian embedding space as the inputs for the neural network. Next, the weighted inputs are passed through several hidden layers with sigmoid activation functions to obtain the output, which is the estimated value of function $f(x)$. If we only use two hidden layers the estimated output can be written as

$$f^e(x) = \sum_s \beta^3_{sr}\ \sigma\left(\sum_q \beta^2_{rq}\ \sigma\left(\sum_k \beta^1_{qk}\ \|x - x_k\|\right)\right).$$

In this equation $f^e(x)$ is the estimation of function f at point x in the embedding space, β matrices define link weights between different layers of the neural network and σ is a sigmoid function. x_k represents a training point in the embedding space for which we know the value of function f. To find the weight matrix, we can minimize the total error, $\sum_k (f(x_k) - f^e(x_k))^2$, calculated using the known values of function f for the training nodes. However, to avoid overfitting, we can modify the total error by adding regularization terms that depends

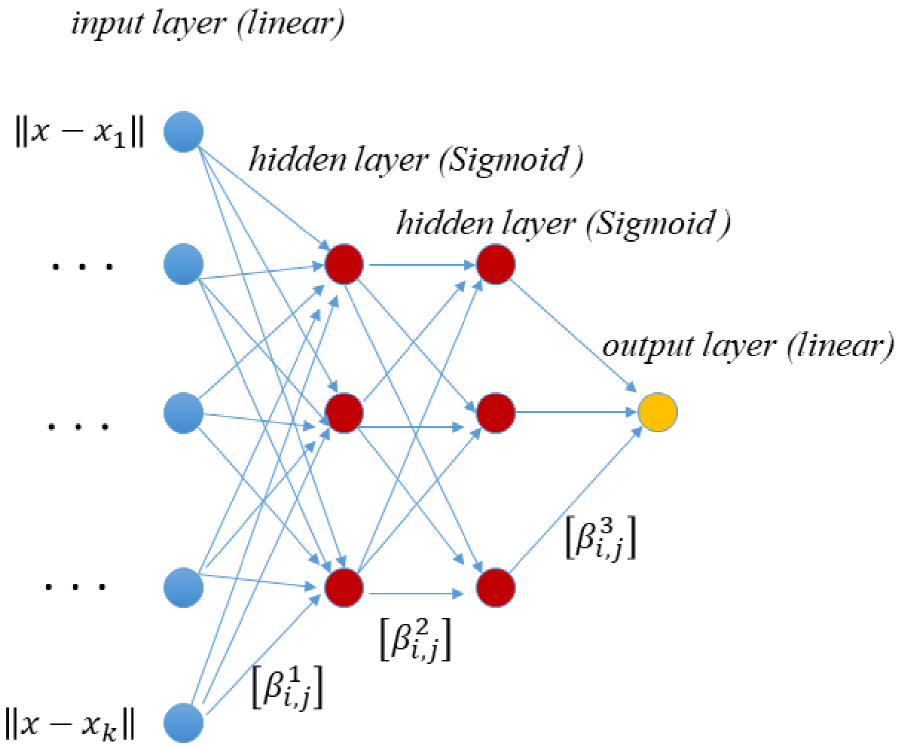


Figure 6.1: *An example of feedforward neural network*

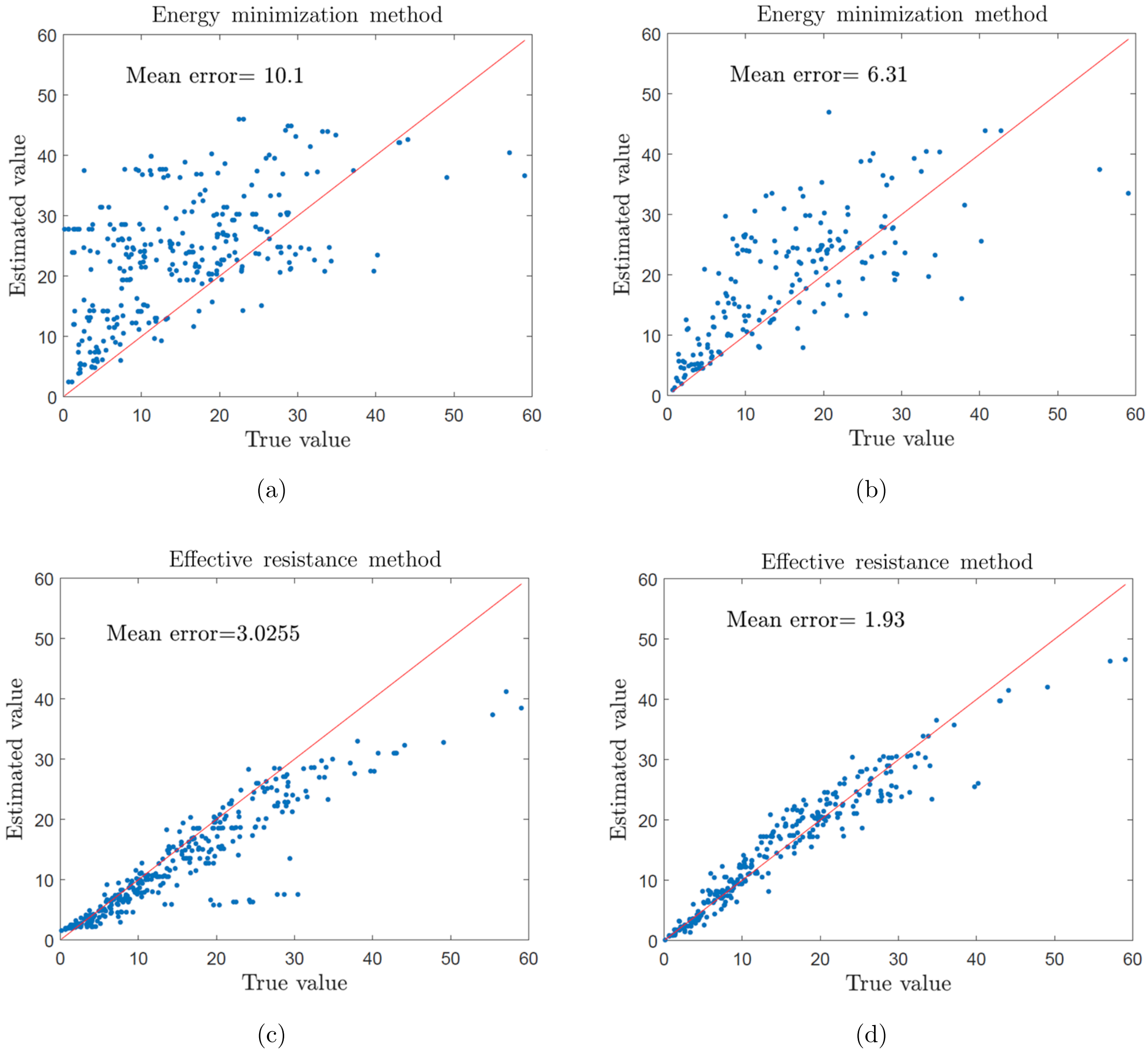
on weight matrices for the neural network.

6.3 Numerical Results

In this experiment, we used the largest component of the coauthorship network from reference[39]. The network is shown in Fig. (5.3) and it has $N = 379$ nodes. For the first step in the experiment, for each node n in the network we generated a set of 100 simulated SIS trajectories assuming the only initial infected node is n. Then, for each set of simulations, we calculated the average number of infected nodes at some time t, i.e, the infection prevalence at time t. For this experiment, we assume t is large enough that the SIS process reaches the metastable state. We denote the value of infection prevalence for each set of simulations by $I(n)$, where n refers to the initial infected node in the simulations. Indeed, we can think of $I(n)$ as a measure of impact for the node n in the SIS spreading. After calculating $I(n)$ for each node n in the graph, we used the values of I for a set of randomly chosen training nodes, denoted by V_l, to estimate I for the remaining nodes in the network, $V_u = V - V_l$. Figure 6.2 compares the estimated and true values of $I(n)$ for $n \in V_u$. Figures 6.2a, 6.2b shows the estimation results using the energy minimization method from section 6.2.1. In figure 6.2a and 6.2b, we used 20 percent and 50 percent of the nodes in the network, respectively, as the training nodes. Figures 6.2c and 6.2d show the estimation outcomes of when the radial basis function and feedforward neural network methods were employed(see section 6.2.2). In figures 6.2c and 6.2d , we used 20 percent of nodes as training nodes. We can see the feedforward network method generated more accurate estimations compared with the estimations obtained by the radial basis function. In summary, machine learning approaches coupled with effective resistance as the measure of distance among nodes is a very promising approach for estimating function values on networks.

6.4 Summary

We applied two different methods for interpolation of a function defined over a graph. The numerical results show that our approach, section 6.2.2, which uses effective resistance distance as the measure of similarity and employs the neural network as the function approximator, provides the most accurate estimation. This might be due to the fact that the function we estimated was generated via a stochastic spreading process. Such processes depend on the global characteristics of the graph, and this might be the reason that using the effective resistance provides a better estimation. In future works, we plan to use our method on different types of functions.



(a)

(b)

(c)

(d)

Figure 6.2: *The plots compare the estimated values of the infection prevalence, $I(n)$, with their true values. For plots (a) and (b), we used the energy minimization method for the estimation and used* 20 *and* 50 *percent of the nodes, respectively, for training. Plots (c) and (d) show the estimation result when we used the radial basis function and the feedforward neural network methods, respectively. In plots (c), (d) we used* 20 *percent of the nodes for training.*

Chapter 7

Conclusion

7.1 Conclusions

In this dissertation, we presented new knowledge about networked spreading processes. Particularly, even though we studied various aspects of the SIS process, the methods we developed in chapters 2,3,4 and 6 are applicable to other networked spreading models as well.

First, we studied the inverse problem of continuous–time SIS spreading over a network. We obtained the likelihood function of observing an SIS trace. Using this likelihood function and powerful Gibbs sampling, we were able to show the feasibility of reconstructing the underlying network from the observed SIS traces. However, our numerical simulations show that accurate network reconstruction requires a long observation of the network nodes.

To understand the effect of link duration in spreading processes, we developed a temporal network model where the temporariness of the links resulted from the transition of nodes between an active and an inactive state. To start our analysis, we derived the temporal characteristic of the network model. Combining the dynamics of the network and the SIS spreading, we studied the effect of link durations on the epidemic of sexually transmitted diseases. Given a fixed value of the infection transmission rate β, we found that, the prevalence of infection strongly depends on the number of casual partnerships and the duration of these partnerships. Our simulations show that the epidemic threshold decreases with increasing link duration. In addition, our analysis shows that casual sexual behavior, which implies short partnerships, makes the population vulnerable to the infection spreading.

Furthermore, we studied the infection localization of SIS processes. We proposed the dispersion entropy as a measure of infection delocalization, and found an upper bound for the infection dispersion entropy when a metastable state exist. This upper bound depends on the infection strength and quantifies the spread of infection in the metastable state. A small upper bound for the dispersion entropy of a metastable state provides a sufficient condition for either complete extinction of an epidemic or its localized persistence.

Finally, we addressed the interpolation of a function defined over the nodes of a network. The function we considered resulted from an SIS spreading unfolding over a network, and it captures the expected number of infected nodes. In particular, the value of the function on a node is the infection size obtained when the infection starts from that specific node. In general, calculating such a function for a networked spreading process requires lengthy simulations. In this dissertation we developed a method based on effective resistance distance between the nodes, and feed forward neural network to interpolate the function based on its values over a limited subset of nodes.

7.2 Future Works

In chapter 4, we proposed a temporal network model that can be combined with specific spreading dynamics to analyze epidemics in temporal networks. One assumption that we made in our analysis was that the nodes' transition time between the active and inactive states have exponential distributions. That leads to link durations that are also exponentially distributed. This distribution of the link duration is a limitation in our model, and generalization of the model to include link durations with different distributions will be relevant. One possible path that might lead to this generalization can be the change of the node transition time distributions. We expect this would, in turn, lead to non-exponential distributions for link durations. If this conjecture is correct it will be possible to write a set of mean–field equations, using phase–type distributions, to describe the spreading process over such a temporal network. A phase–type distribution can be considered as the distribution of the absorbing time in a continuous-time Markov process with several transient states and one absorbing state. It is well known that any positively valued distribution can be approximated with a phase-type distribution. The use of a phase-type distribution consists of adding some auxiliary states to the original spreading model, replacing all non-exponential distributions of inter-event times in the original model with a mixture of exponential distributions. Since all inter-event times will have exponential distributions in the modified spreading process, we can develop differential equations that describe the spreading process.

In Chapter 6 section 6.2.2, we proposed a method for interpolating networked spreading data. Although we applied the method for a specific function, we expect this method can be applied to various functions defined over a network. It is possible that only the state of a subset of nodes is available during an outbreak. We expect our method would be able to estimate the state of remaining nodes in the network using the available information about the known subset. Therefore, extending the method to other functions can have a great impact in practical cases.

Bibliography

[1] Steven Sanche, Yen Ting Lin, Chonggang Xu, Ethan Romero-Severson, Nicolas W Hengartner, and Ruian Ke. The novel coronavirus, 2019-nCoV, is highly contagious and more infectious than initially estimated. *arXiv preprint arXiv:2002.03268*, 2020.

[2] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[3] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[4] Neil M Ferguson and Geoffrey P Garnett. More realistic models of sexually transmitted disease transmission dynamics: sexual partnership networks, pair models, and moment closure. *Sexually transmitted diseases*, 27(10):600–609, 2000.

[5] Robert M May, Roy M Anderson, and ME Irwin. The transmission dynamics of human immunodeficiency virus (hiv). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 321(1207):565–607, 1988.

[6] Johannes Müller, Birgitt Schönfisch, and Markus Kirkilionis. Ring vaccination. *Journal of mathematical biology*, 41(2):143–171, 2000.

[7] Ingemar Nåsell. Stochastic models of some endemic infections. *Mathematical biosciences*, 179(1):1–19, 2002.

[8] Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.

[9] Daryl J Daley, Joe Gani, and Joseph Mark Gani. *Epidemic modelling: an introduction.* Cambridge University Press, 2001.

[10] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63(6):066117, 2001.

[11] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*, pages 25–34. IEEE, 2003.

[12] Sunetra Gupta, Roy M Anderson, and Robert M May. Networks of sexual contacts: Implication for the pattern of spread. *Aids*, 3(12), 1989.

[13] Ken TD Eames and Matt J Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the National Academy of Sciences*, 99(20):13330–13335, 2002.

[14] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[15] Sergey N Dorogovtsev, Alexander V Goltsev, and José FF Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4):1275, 2008.

[16] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65(3):036104, 2002.

[17] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87 (3):925, 2015.

[18] Piet Van Mieghem. The n-intertwined sis epidemic network model. *Computing*, 93 (2-4):147–169, 2011.

[19] Faryad Darabi Sahneh, Caterina Scoglio, and Piet Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking*, 21(5):1609–1620, 2013.

[20] István Z Kiss, Joel C Miller, and Péter L Simon. Mathematics of epidemics on networks. 2017.

[21] Jeffrey O Kephart and Steve R White. Directed-graph epidemiological models of computer viruses. In *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*, pages 343–359. IEEE, 1991.

[22] Mark EJ Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[23] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[24] Sergio Gomez, Albert Diaz-Guilera, Jesus Gomez-Gardenes, Conrad J Perez-Vicente, Yamir Moreno, and Alex Arenas. Diffusion dynamics on multiplex networks. *Physical review letters*, 110(2):028701, 2013.

[25] Igor Belykh, Martin Hasler, Menno Lauret, and Henk Nijmeijer. Synchronization and graph topology. *International Journal of Bifurcation and Chaos*, 15(11):3423–3433, 2005.

[26] Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.

[27] Osman Yağan and Virgil Gligor. Analysis of complex contagions in random multiplex networks. *Physical Review E*, 86(3):036103, 2012.

[28] E Cator and P Van Mieghem. Nodal infection in markovian susceptible-infected-susceptible and susceptible-infected-removed epidemics on networks are non-negatively correlated. *Physical Review E*, 89(5):052802, 2014.

[29] Faryad Darabi Sahneh and Caterina Scoglio. Competitive epidemic spreading over arbitrary multilayer networks. *Physical Review E*, 89(6):062817, 2014.

[30] Mina Youssef and Caterina Scoglio. An individual-based approach to sir epidemics in contact networks. *Journal of theoretical biology*, 283(1):136–144, 2011.

[31] Victor M Preciado, Michael Zargham, Chinwendu Enyioha, Ali Jadbabaie, and George Pappas. Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 7486–7491. IEEE, 2013.

[32] Azwirman Gusrialdi, Zhihua Qu, and Sandra Hirche. Distributed link removal using local estimation of network topology. *IEEE Transactions on Network Science and Engineering*, 2018.

[33] Faryad Darabi Sahneh, Aram Vajdi, Heman Shakeri, Futing Fan, and Caterina M. Scoglio. Gemfsim: A stochastic simulator for the generalized epidemic modeling framework. *J. Comput. Science*, 22:36–44, 2017. doi: 10.1016/j.jocs.2017.08.014. URL https://doi.org/10.1016/j.jocs.2017.08.014.

[34] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[35] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[36] Marian Boguná, Luis F Lafuerza, Raúl Toral, and M Ángeles Serrano. Simulating non-Markovian stochastic processes. *Physical Review E*, 90(4):042108, 2014.

[37] Piet Van Mieghem. *Performance analysis of communications networks and systems.* Cambridge University Press, 2009.

[38] GEMFsim: implementation of the generalized epidemic modeling framework in matlab, r, python, and c. https://www.ece.k-state.edu/netse/software/index.html.

[39] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[40] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009. Datasets used in this paper is available online http://toreopsahl.com/datasets.

[41] Faryad Darabi Sahneh and Caterina Scoglio. Epidemic spread in human networks. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 3008–3013. IEEE, 2011.

[42] Faryad Darabi Sahneh, Fahmida N Chowdhury, and Caterina M Scoglio. On the existence of a threshold for preventive behavioral responses to suppress epidemic spreading. *Scientific reports*, 2, 2012.

[43] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[44] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[45] Jantien A Backer, Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 january 2020. *Eurosurveillance*, 25(5), 2020.

[46] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine*, 2020.

[47] Aram Vajdi and Caterina M Scoglio. Identification of missing links using susceptible-infected-susceptible spreading traces. *IEEE Transactions on Network Science and Engineering*, 6(4):917–927, 2018.

[48] Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1): 016128, 2002.

[49] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[50] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical review letters*, 90(2):028701, 2003.

[51] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

[52] Alexander V Goltsev, Sergey N Dorogovtsev, Joao G Oliveira, and Jose FF Mendes. Localization and spreading of diseases in complex networks. *Physical review letters*, 109(12):128702, 2012.

[53] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[54] Praneeth Netrapalli and Sujay Sanghavi. Learning the graph of epidemic cascades. In *ACM SIGMETRICS Performance Evaluation Review*, volume 40, pages 211–222. ACM, 2012.

[55] Takayuki Kamei, Keiko Ono, Masahito Kumano, and Masahiro Kimura. Predicting missing links in social networks with hierarchical dirichlet processes. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

[56] Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schoelkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *International Conference on Machine Learning*, pages 793–801, 2014.

[57] Edward Choi, Nan Du, Robert Chen, Le Song, and Jimeng Sun. Constructing disease network and temporal progression model via context-sensitive hawkes process. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 721–726. IEEE, 2015.

[58] Quang Duong, Michael P Wellman, and Satinder Singh. Modeling information diffusion in networks with unobserved links. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 362–369. IEEE, 2011.

[59] Varun R Embar, Rama Kumar Pasumarthi, and Indrajit Bhattacharya. A bayesian framework for estimating properties of network diffusions. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1216–1225. ACM, 2014.

[60] Emre Sefer and Carl Kingsford. Convex risk minimization to infer networks from probabilistic diffusion data at multiple scales. In *Data engineering (ICDE), 2015 IEEE 31st international conference on*, pages 663–674. IEEE, 2015.

[61] Alfredo Braunstein and Alessandro Ingrosso. Network reconstruction from infection cascades. *arXiv preprint arXiv:1609.00432*, 2016.

[62] Zhuozhao Li, Haiying Shen, and Kang Chen. Learning network graph of sir epidemic cascades using minimal hitting set based approach. In *Computer Communication and Networks (ICCCN), 2016 25th International Conference on*, pages 1–9. IEEE, 2016.

[63] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alejandro Lage-Castellanos, and Riccardo Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112(11):118701, 2014.

[64] Mehrdad Farajtabar, Manuel Gomez-Rodriguez, Nan Du, Mohammad Zamani, Hongyuan Zha, and Le Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *Artificial Intelligence and Statistics*, 2015.

[65] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Cascade-based community detection. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 33–42. ACM, 2013.

[66] Long Tran, Mehrdad Farajtabar, Le Song, and Hongyuan Zha. Netcodec: Community detection from individual activities. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 91–99. SIAM, 2015.

[67] Andrey Y. Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E*, 90: 012801, Jul 2014.

[68] Wuhua Hu, Wee Peng Tay, Athul Harilal, and Gaoxi Xiao. Network infection source identification under the siri model. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1712–1716. IEEE, 2015.

[69] Nino Antulov-Fantulin, Alen Lančić, Tomislav Šmuc, Hrvoje Štefančić, and Mile Šikić. Identification of patient zero in static and temporal networks: Robustness and limitations. *Physical review letters*, 114(24):248701, 2015.

[70] Zhen Chen, Kai Zhu, and Lei Ying. Detecting multiple information sources in networks under the sir model. *IEEE Transactions on Network Science and Engineering*, 3(1): 17–31, 2016.

[71] Pedro C. Pinto, Patrick Thiran, and Martin Vetterli. Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.*, 109:068702, Aug 2012.

[72] Aram Vajdi, David Juher, Joan Saldaña, and Caterina Scoglio. A multilayer temporal network model for std spreading accounting for permanent and casual partners. *Scientific Reports*, 10(1):1–12, 2020.

[73] Lori Newman, Jane Rowley, Stephen Vander Hoorn, Nalinka Saman Wijesooriya, Magnus Unemo, Nicola Low, Gretchen Stevens, Sami Gottlieb, James Kiarie, and Marleen Temmerman. Global estimates of the prevalence and incidence of four curable sexually transmitted infections in 2012 based on systematic review and global reporting. *PloS one*, 10(12):e0143304, 2015.

[74] Gilla K Shapiro, Ovidiu Tatar, Arielle Sutton, William Fisher, Anila Naz, Samara Perez, and Zeev Rosberger. Correlates of tinder use and risky sexual behaviors in young adults. *Cyberpsychology, Behavior, and Social Networking*, 20(12):727–734, 2017.

[75] Xiao Zhang, Cristopher Moore, and Mark EJ Newman. Random graph models for dynamic networks. *The European Physical Journal B*, 90(10):200, 2017.

[76] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[77] Nicos Georgiou, Istvan Z Kiss, and Enrico Scalas. Solvable non-markovian dynamic network. *Physical Review E*, 92(4):042801, 2015.

[78] Philip E Paré, Carolyn L Beck, and Angelia Nedić. Epidemic processes over time-varying networks. *IEEE Transactions on Control of Network Systems*, 5(3):1322–1334, 2018.

[79] Mustapha Ait Rami, Vahid Samadi Bokharaie, Oliver Mason, and Fabian Wirth. Stability criteria for sis epidemiological models under switching policies. *Discrete and Continuous Dynamical Systems-Series B*, 19(9):2865–2887, 2014.

[80] Mohammad Reza Sanatkar, Warren N White, Balasubramaniam Natarajan, Caterina M Scoglio, and Karen A Garrett. Epidemic threshold of an sis model in dynamic

switching networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(3):345–355, 2016.

[81] David Juher, Jordi Ripoll, and Joan Saldaña. Outbreak analysis of an sis epidemic model with rewiring. *Journal of Mathematical Biology*, 67(2):411–432, Aug 2013.

[82] Andrea EF Clementi, Claudio Macci, Angelo Monti, Francesco Pasquale, and Riccardo Silvestri. Flooding time of edge-markovian evolving graphs. *SIAM journal on discrete mathematics*, 24(4):1694–1712, 2010.

[83] Michael Taylor, Timothy J Taylor, and Istvan Z Kiss. Epidemic threshold and control in a dynamic network. *Physical Review E*, 85(1):016103, 2012.

[84] Masaki Ogura and Victor M Preciado. Stability of spreading processes over time-varying large-scale networks. *IEEE Transactions on Network Science and Engineering*, 3(1):44–57, 2016.

[85] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2:469, 2012.

[86] Iacopo Pozzana, Kaiyuan Sun, and Nicola Perra. Epidemic spreading on activity-driven networks with attractiveness. *Physical Review E*, 96(4):042310, 2017.

[87] Quan-Hui Liu, Xinyue Xiong, Qian Zhang, and Nicola Perra. Epidemic spreading on time-varying multiplex networks. *Phys. Rev. E*, 98:062303, Dec 2018. doi: 10.1103/PhysRevE.98.062303.

[88] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.

[89] Trystan Leng and Matt J Keeling. Concurrency of partnerships, consistency with data, and control of sexually transmitted infections. *Epidemics*, 25:35–46, 2018.

[90] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networking (TON)*, 17(1):1–14, 2009.

[91] Faryad Darabi Sahneh. *Spreading processes over multilayer and interconnected networks*. PhD thesis, Kansas State University, 2014.

[92] Shlomo Sternberg. *Dynamical systems*. Courier Corporation, 2010.

[93] Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007. doi: 10.1146/annurev.physchem.58.032806.104637.

[94] Tom Britton, Maria Deijfen, and Anders Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377 – 1397, 2006. doi: https://doi.org/10.1007/s10955-006-9168-x.

[95] Mirjam Kretzschmar and Janneke CM Heijne. Pair formation models for sexually transmitted infections: a primer. *Infectious Disease Modelling*, 2(3):368–378, 2017.

[96] D Hansson, KY Leung, T Britton, and S Strömdahl. A dynamic network model to disentangle the roles of steady and casual partners for hiv transmission among msm. *Epidemics*, 2019.

[97] Faryad Darabi Sahneh, Aram Vajdi, and Caterina Scoglio. Delocalized epidemics on graphs: A maximum entropy approach. In *2016 American Control Conference (ACC)*, pages 7346–7351. IEEE, 2016.

[98] Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 2, pages 1455–1466. IEEE, 2005.

[99] Thomas Mountford, Daniel Valesin, and Qiang Yao. Metastable densities for the contact process on power law random graphs. *Electron. J. Probab*, 18(103):1–36, 2013.

[100] Marian Boguñá, Claudio Castellano, and Romualdo Pastor-Satorras. Nature of the epidemic threshold for the susceptible-infected-susceptible dynamics in networks. *Physical review letters*, 111(6):068701, 2013.

[101] Géza Ódor. Spectral analysis and slow spreading dynamics on complex networks. *Physical Review E*, 88(3):032109, 2013.

[102] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

[103] Inc. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0. `http://cvxr.com/cvx`, 2012.

[104] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[105] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.