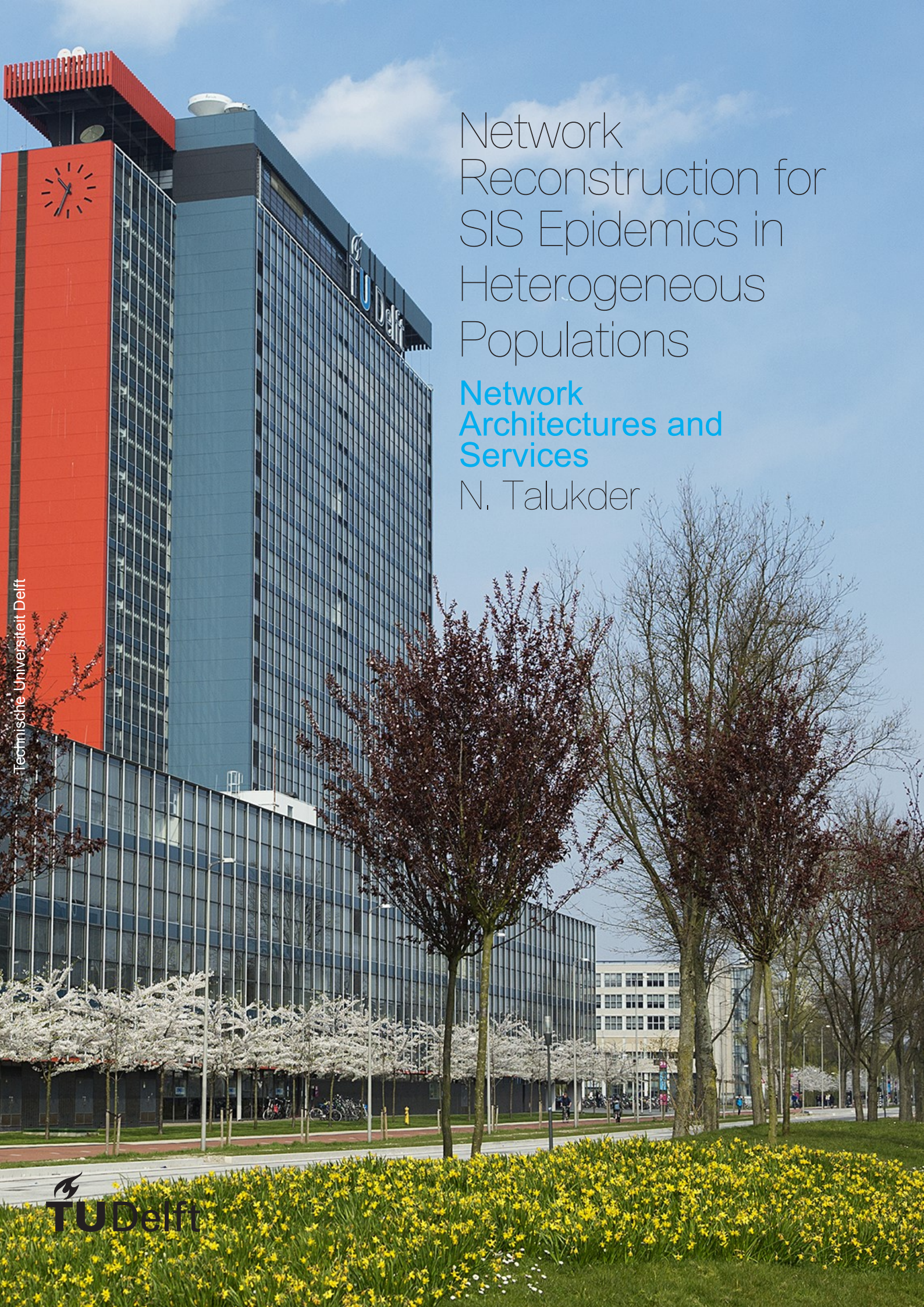


Network Reconstruction for SIS Epidemics in Heterogeneous Populations

Network Architectures and Services

N. Talukder

Technische Universiteit Delft

TU Delft

Network Reconstruction for SIS Epidemics in Heterogeneous Populations

Network Architectures and Services

by

N. Talukder

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday June 17, 2021 at 11:00 AM.

Student number:	4946081	
Project duration:	October 20, 2020 – June 17, 2021	
Thesis committee:	Prof. dr. ir. P. F. A. Van Mieghem,	TU Delft
	Dr. ir. R. Kraaij,	TU Delft
	Dr. B. Prasse,	TU Delft, daily supervisor
	Ir. M. Achterberg,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Abstract

Exact network reconstruction from observations of the SIS process in discrete time would be very useful if possible, with implications for tracking the spread of infectious diseases, trends and news on social media. It could provide estimates for the strength of links in a network and the contribution of individual nodes to the spread of an epidemic within a network as well as the underlying structure.

This Thesis provides a method for evaluating heterogeneous parameters where each node has a randomly distributed curing probability and each link between two nodes has a randomly distributed infection probability. The parameters are computed via maximum likelihood estimation using between 10^2 and 10^4 observations of the SIS process on networks ranging in size from 15 to 55 nodes, for both directed Erdős-Rényi and Barabási-Albert graphs. We vary the network size to demonstrate that for a fixed level of accuracy, the number of required observations increases exponentially with the number of nodes for both the whole network and a subset of links and nodes. We further demonstrate that it may require fewer observations to reconstruct certain nodes based on the degree of the node or reconstruct links based on the degree of the node to which the link is incident. Additionally, if we interpret 10^6 or more observations as the number of required observations where reconstruction becomes infeasible, a network size of 500 would be infeasible for reconstructing the full network and the approximate limit for partial network reconstruction.

The Thesis is extended to look at the SI and SIR models, achieving a similar exponential increase in the number of observations required as the network size increases, for a fixed error.

Preface

This Thesis is the final part of my degree in Embedded Systems, software and networking (MSc), it has been quite a journey and I have really appreciated learning about the different areas within epidemic and network research.

This Thesis has been completed in the Network, Architecture and Services (NAS) group at TU Delft. I have completed my Thesis under the supervision of Piet Van Miegham whom I would like to thank for the opportunity to study and complete research within the NAS group. A huge amount of thanks go to my daily supervisor Bastian Prasse who provided invaluable guidance and feedback throughout my Thesis.

Many thanks go to my family and friends for their support during this period and helping me get through a variety of challenges.

N. Talukder
Delft, June 2021

Contents

List of Symbols

p_{ER}	Link probability for the Erdős-Rényi model
$p_{\mathrm{BA},i}(k_i)$	Probability of attachment for an additional node, for the existing node i with k_i links
p_{WS}	Link rewiring probability for a node in the WS model
d_i	Degree of node i
A	$N \times N$ adjacency matrix with elements a_{ij}
$\tilde{B}$	$N \times N$ infection rate matrix with elements $\tilde{\beta}_{ij}$, the infection rate to node i from node j
B	$N \times N$ infection probability matrix with elements $\beta_{ij} = \Delta t \tilde{\beta}_{ij}$, infection probability to node i from node j
β_i	$N \times 1$ vector representing the sum of the infection probabilities incident on a node, $\beta_i = \sum_{j=0}^{N} \beta_{ij}$
z_k	k-th eigenvalue of the effective infection probability matrix W, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$
$\beta_{\max}$	Preset maximum infection rate between two nodes, $\beta_{ij} \in [0, \beta_{\max}]$
$\chi_i[k]$	Weighted degree of node i at time k, $\chi_i[k] = \sum_{j \neq i} \beta_{ij} x_j[k]$
$\chi[k]$	$N \times 1$ weighted degree vector at time k, $\chi[k] = (\chi_1[k], \ldots, \chi_N[k])^T$
$x_i[k]$	SIS viral state of node i at time k
$x[k]$	$N \times 1$ SIS viral state vector at time k, $x[k] = (x_1[k], \ldots, x_N[k])^T$
$X[n]$	$n \times N$ matrix of observations up to time n, $(x[1], \ldots, x[n])^T$
$\tilde{\delta}_i$	Curing rate of node i
$\tilde{\delta}$	$N \times 1$ curing rate vector, $\tilde{\delta} = (\tilde{\delta}_1, \ldots, \tilde{\delta}_N)^T$
δ_i	Curing probability of node i, $\delta_i = \tilde{\delta}_i \Delta t$
δ	$N \times 1$ curing probability vector, $\delta = (\delta_1, \ldots, \delta_N)^T$
$\delta_{\max}$	Maximum curing probability possible for a node, $\delta_i \in [0, \delta_{\max}]$
η_δ	Error in curing probability, averaged over all nodes, $\eta_\delta = \lvert\delta_i - \hat{\delta}_i\rvert / N$
η_β	Average error of infection probability all links $\eta_\beta = \sum_{(i,j) \in S_q} \frac{\lvert\beta_{ij} - \hat{\beta_{ij}}\rvert}{\lvert S_q \rvert}$, where S_q is the set of all links
η_d	Average error of the weighted degree of a node $\eta_d = \lvert\beta_i - \hat{\beta}_i\rvert / N$
θ	Parameter vector, $\theta = (\delta_1, \ldots, \delta_N, \beta_{11}, \ldots, \beta_{NN})$
$f_{\theta_i}(\theta_i)$	Likelihood function for the parameters, θ_i of node i
$f_\theta(\Theta)$	The sum of the likelihood functions of the parameters for all nodes, $f_\theta(\Theta) = \sum_{i=1}^{N} f_{\theta_i}(\theta_i)$
$f_{\theta_i \mid X[n]}(\theta_i)$	The negative log-likelihood of the parameters of node i, θ_i conditioned on the observations $X_i[n]$ of node i.
$f_{\theta \mid X[n]}(\theta)$	The sum of the negative log-likelihoods over all nodes, $f_{\theta \mid X[n]}(\theta) = \sum_{i=1}^{N} f_{\theta_i \mid X[n]}(\theta_i)$
N	Number of nodes
n	Number of observations
K	Number of runs for each graph
R_0	Basic reproduction number
W	$N \times N$ effective infection probability matrix with elements $w_{ij} = \tilde{\beta}_{ij} / \tilde{\delta}_i$, effective infection probability to node i from node j
$G(\mathcal{N}, \mathcal{L})$	Graph with nodes $\mathcal{N}$ and links $\mathcal{L}$
$\hat{\theta}_{\mathrm{ML}}$	Maximum likelihood estimate for the parameter set θ

1 Introduction

The increased presence of epidemic research in the media has resulted in a renewed interest in the field. The pandemic outbreak of the Coronavirus disease, COVID-19, and its current impact across the globe vindicate the effort to both produce more accurate forecast models and improve epidemic tracing.

In order to study epidemics effectively, the fundamental dynamics of the spreading process have to be defined. This is not always a trivial question for an infectious disease as there are several methods of transmission such as via a vector or contact with contaminated items. In addition to the transmission routes of a virus, the curing process is important for describing epidemics. There could be different methods for curing, individuals could cure by themselves or depend on a connection to another, 'curing' individual, e.g., a doctor which provides the cure. Most epidemic models assume that a healthy individual can get infected via the contact to at least one infectious individual, and that infected individuals cure independently of the contact with other individuals. When we determine the collection of links between individuals in a population, we obtain a *contact network,* which can become very complex even under idealised conditions. There are complications that occur when mapping any contact network which appears as noisy data when analysing the network. We consider the simplest case with self-curing and transmission from contact with one or more individuals to avoid this as much as possible.

In the case of COVID-19, the disease enters a new population via a handful on individuals or even a particular individual and then propagates through the rest of the population by passing from one person to another. Determining the origin of the disease is very useful as it can allow for an epidemic to be contained, by isolating the source or for all possible chains of transmission to be identified. By checking the medical records and death certificates for symptoms of the disease the path can, in theory, be retraced to the origin. There are many examples from history of particular people being identified as the originator of a disease, such as the attribution of the introduction of smallpox to Central Americas to a Spanish soldier [33]. In this example, the source was traced by hand to a single event, which gave some certainty to where the initial transmission occurred. The problem of locating the source becomes much more difficult if there are many possible events where a transmission can occur such as locally, within neighbourhoods, or from travel between remote regions. Here analysing all the possibilities for the spread of the infection person by person is very cumbersome and slow due to imperfect memory.

A modern approach to tracing the origin of an epidemic outbreak is contact tracing apps that monitor the proximity of an app user to a person later confirmed to be positive for the disease from a positive *loop-mediated isothermal amplification* (LAMP) or *polymerase chain reaction* (PCR) test, an example of which is the *CoronaMelder* app [37]. There are challenges such as the fact only a fraction of the population is currently using the app. Consequently, using all the data gathered from app users would not constitute a complete set of observations. Other recent efforts in monitoring the contact network between individuals for disease control include *Test-and-Trace* schemes, which map a network over which a disease spreads by taking an infected individual and determining anyone they were in contact with over a short period [17]. Test-and-Trace schemes rely on gathering as much data as possible and backtrace to find who could have come into contact with the infected individual. Both contact tracing apps and Test-and-Trace schemes provide some detail about the contact network and may help contain an epidemic outbreak. Data from the COVID-19 outbreak in South Korea suggests that keeping the number of infections low can be achieved by gathering as much information as possible

in the begin stages and isolating the individuals [29].

In this thesis we reconstruct the contact network from observations of the epidemic. More precisely, we consider the best-case scenario where the health status (*healthy* or *infected*) of every individual is accessible at every time. We therefore explore the full potential of monitoring the health status of individuals for uncovering the contact network and, ultimately, for predicting and controlling an epidemic outbreak.

In [40] it is shown that there is a fraction of the population that is responsible for a disproportionality high number of infections. Specifically 10% of the infected population can sometimes be responsible for up to 45% of further infections. Identifying potential super-spreaders would be very useful for predicting and mitigating future outbreaks. This relates to individual probabilities of transmission suggesting that a focus on high risk connections is perhaps better than identifying all contacts to a person [23]. In this thesis, we aim to determine how accurately connections in the contact network can be estimated, to identify which individuals pose the greatest risk and would be responsible for the greatest number of reinfections.

Lastly, we stress that, beyond the spread of infectious diseases, modern epidemiology encompasses a broad spectrum of spreading phenomena. The spread of news, posts and opinions via modern communication can be approached in a way similar to an infectious disease where instead of a disease, news is spread. An example of this is the spread of fake news on Twitter [1], currently one of the largest and most effective media for the dissemination of news. Twitter allows any news to be relayed at the press of a button to anyone following the person re-tweeting the information. By keeping track of who is re-tweeting what, one can obtain an indication of who is particularly susceptible to fake news and also obtain insights on how fake news spreads. Twitter can give every user a particular score or probability to believe news stories, which is analogous to them being infected by a disease.

From a network point of view, Twitter resembles a heterogeneously weighted network where users correspond to nodes and links correspond to followings. The chance of retweeting varies from user to user, leading to a greater prevalence of certain news amongst certain groups [9]. This thesis considers a heterogeneously weighted network to account for the fact that certain interactions are much more likely to occur than others.

1.1. Related Work

The main previous work on which this thesis is based is [34], where the *susceptible-infected-susceptible* (SIS) model is considered. Here it is shown that for a network with homogeneous spreading parameters, an exact network reconstruction is only possible for an infeasibly large number of observations. In this thesis, we extend the network reconstruction approach in [34] to heterogeneous spreading parameters. A discussion of some of the problems encountered in reconstructing networks from observed epidemics is provided in [3].

Network reconstruction can also be used in conjunction with community detection. Peixoto [32] developed a network reconstruction method and community detection method, which was applied to uncover higher-order voting patterns in the lower chamber in the Brazilian parliament and demonstrated the unity of various parties. In [32], a high number of observations of the SIS process are considered with a maximum of 10^3 infections for each node.

Network reconstruction techniques that rely on Bayesian estimation, similar to those employed in this thesis, are presented in [16] and [27]. In [16], the diffusion of web-links or phrases through a network with homogenous spreading parameters is investigated via the *susceptible-infected* (SI) model. In [16], the author proposes the `NETINF` algorithm to find the network which maximised the probability of the observed diffusion. In [27], Bayesian network reconstruction is extended to networks with heterogeneous spreading parameters for an SI, SIS and *susceptible-infected-removed* (SIR) model. In [16], clustering based on types of source was observed when analysing phrases. Some phrases spread more quickly in certain circles than others such as the spread of political phrases in political websites and accounts compared to in entertainment websites and accounts.

There is a limit to the network size that can be estimated accurately with a given number of observations of dynamics on the respective network. This thesis aims to find this limit computationally by evaluating the number of observations as a function of the network size, for a given level of accuracy.

1.2. Document Structure

Chapter 2 introduces the models we used to generate networks and real-world applications including the Erdős-Rényi, Barbási-Albert and Watts-Strogatz models. The background of epidemic modelling is given, with a focus on compartmental models. We introduce the epidemic model in Chapter 3, which is based on a discrete-time Markov chain. Chapter 4 formulates the network reconstruction problem as an optimisation problem. Chapter 4 applies the reconstruction method to networks of different sizes and different random graph models to obtain estimates of the network and the spreading parameters. Chapter 5 applies the estimation to the SI and SIR processes. Chapter 6 states the conclusions of this thesis and gives an outlook to future work.

2

Background

2.1. Network Science

Network science studies the patterns that emerge in complex, interconnected systems. Applications of network science include monitoring the performance of telecommunication networks and electrical grids, studying the interactions between proteins in living organisms via biological networks [10] and analysing clustering in financial markets [24]. Network science can also be used to predict the behaviour of complex systems using knowledge of the underlying graph such as the degree distribution of nodes.

Social science provides several good examples of the relation between the structure and the behaviours of a network. For example, the small-world experiment run by Milgram [26] which aimed to find out how quickly an unaddressed and named letter could find a person via a chain of social connections across the United States. The result showed that the average number of intermediate steps was around 6. This behaviour could be attributed to the structure of the network of connections which later was modelled by Watts and Strogatz [43]. In this chapter we introduce how graph theory provides the basic models used within epidemic modelling.

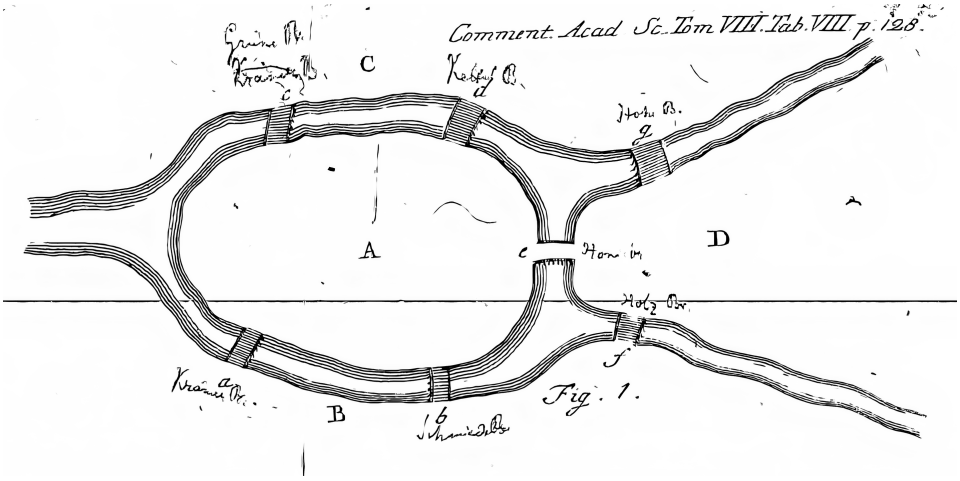
2.1.1. Graph Theory

Nodes are discrete entities, they can, for example, be individuals in a population or computers in a network. Links connect pairs of nodes, in the case of a populations this could be interactions between individuals. We denote a graph by $G(\mathcal{N}, \mathcal{L})$, where $\mathcal{N} = \{1, ..., N\}$ is the set of N nodes and $\mathcal{L}$ is the set of links. Any graph corresponds to an adjacency matrix A, whose entries a_{ij} are defined as:

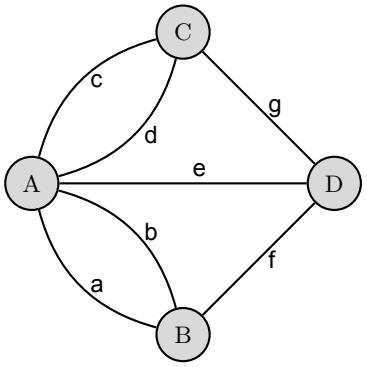
$$a_{ij} = \begin{cases} 1 \text{ for } (i,j) \in \mathcal{L} \\ 0 \text{ otherwise} \end{cases} \tag{2.1}$$

Graph Theory originates from the Königsberg bridge problem. The city of Königsberg consisted of four landmasses connected by seven bridges which presented the residents with a puzzle: *is it possible to traverse all the bridges without crossing the same bridge multiple times?* In 1736, Euler [14] demonstrated that the problem could be studied by representing the city as a graph as shown in Figure 2.1a. This particular topology cannot be traversed with a single path as the graph created based on the city is neither *Eulerian* nor *semi-Eulerian,* meaning that the total number of nodes with an odd degree is zero or at most two, respectively. As shown in Figure 2.1c, there are more than two nodes with an odd degree, which is sufficient to show a trail cannot exist for the given topology in Figure 2.1b.

Graph theory has been used extensively to map and study networks. Social networks are an example of the scale of human interaction networks people currently use day-to-day, with over 1.69 billion [11] active users on Facebook alone. Modern research into real-world networks often characterises a network via graph metrics. An important metric is *connectedness* which defines whether a graph is connected or not, meaning a path exists between each pair of nodes. Examples of real world connected graphs includes the telephone network, where any two phones may be connected in order for a call to be established. Further important metrics are *closeness* and *connectivity.* Closeness, $Cl_i(G)$ is defined as the reciprocal of the total hopcount of all the shortest paths that start at node i and end at any other



(a) The centre of the city of Königsberg, with regions A to D and bridges a to g.



(b) Graph representation of the seven bridges of Königsberg.

Area (Node)	Degree
A	5
B	3
C	3
D	3

(c) Table representation of Königsberg.

Figure 2.1: Euler's analysis of the Königsberg bridge problem. 2.1a, including the figure from the 1736 paper [14] and a simplification of the problem to nodes and links in 2.1b. Table 2.1c shows there are more than two nodes with an odd degree, demonstrating that the graph is neither semi-Eulerian nor Eulerian.

nodes in the graph. The larger the average shortest path, the larger the closeness Cl_i. Therefore the closeness Cl_i gives a measure of the participation of a node in a network. Connectivity, $\lambda(G)$ measures how resilient a connected graph is, that is the smallest number of nodes or edges that leads to the graph becoming disconnected. There are many more useful metrics that are used in network analysis [41].

2.1.2. Random Graph Models

Random graphs are useful for simulating graphs with an irregular or unknown structure. Many real world networks grow or are created randomly. This includes even small networks such as a group of people who have been just introduced to each other, where it is very difficult to predict who will end up speaking with whom, during a given period of time. The best way to represent this group and analyse the network structure is via a random graph.

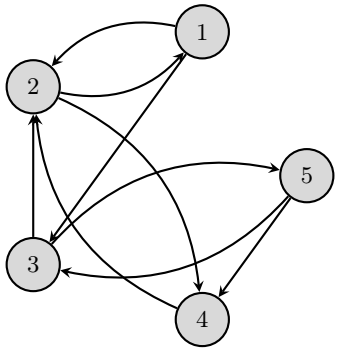
Random graphs are a type of graph where the links between nodes are determined probabilistically. The origin of random graph theory starts with the Hungarian mathematicians Erdős and Rényi who proposed the Erdős-Rényi (ER) model [13] and they proceeded to study the properties of various random graphs. One version of the ER model consists of a constant number of links which are randomly permuted between different nodes resulting in uniform random graphs, denoted by $G_m(N)$, where the number of links, $L = m$.

The second type of ER model results in binomial random graphs and was first proposed by Edgar Gilbert [15]. Here, the absence or presence of any link is modelled by a Bernoulli random variable with success probability p_{ER} and the distribution (2.2). Figure 2.2 shows an example of a $G_{p_{\mathrm{ER}}}(N)$ graph.

$$a_{ij} \sim \text{Ber}(p_{\text{ER}}),\ i \neq j,\ a_{ii} = 0 \tag{2.2}$$

$$A_{ER} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

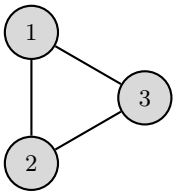
(a) Adjacency matrix A_{ER} for the graph $G_{0.7}(N, L)$



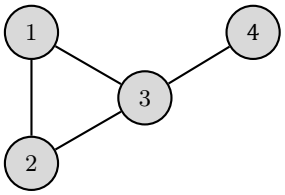
(b) 5 node ER-directed graph, $G_{0.7}(N, L)$

Figure 2.2: An example of a 5 node random binomial graph with directed links.

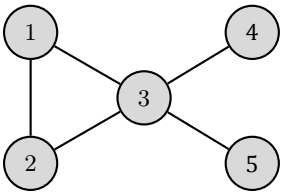
An alternative to the ER-model was found by Barabási and Albert when analysing the structure of the World Wide Web [4]. The scale-free nature of the network was a revelation as previous random graph models, including the ER model, did not even closely reproduce a scale-free graph. Barabási and Albert introduced the Barabási-Albert model which includes a power law degree distribution allowing networks produced to be scale-free. The key of component of the BA model was *preferential attachment* which models the growth of a network: new nodes attach, i.e. form a link, to the highest degree nodes with the greatest probability. More specifically, if a new node is added to the network, the probability $p_{\text{BA},i}(k_i)$ of attaching node i is proportional to the degree d_i of node i. This allows the probability to be defined as $p_{\text{BA},i}(k_i) = \frac{d_i}{d_{total}}$. The BA model has two input parameters: m and m_0. The initial nodes m_0 form a complete graph, as given in Figure 2.3a. Each further node attaches to m nodes already in the graph as depicted in Figure 2.3b and 2.3c.



(a) Step 1 with $m_0 = 3$ nodes.



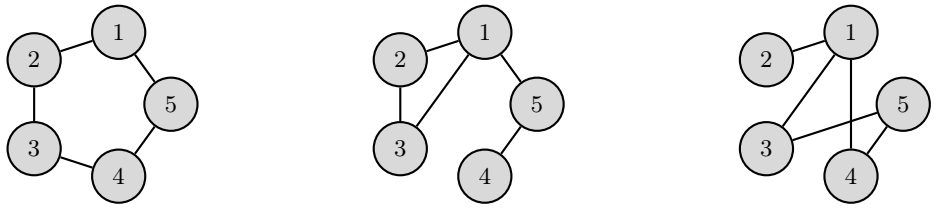
(b) Step 2. Adding node 4.



(c) Step 3. Adding node 5.

Figure 2.3: A BA graph with $N = 5$ nodes, where one node and $m = 1$ links are added sequentially using preferential attachment. The graph grows from the initial complete graph given in Subfigure 2.3a by attaching two nodes sequentially. First in Subfigure 2.3b, where the probability of the new node attaching to node 3 is $p_{\text{BA},3} = d_3/d_{total} = 2/6$. Then, in Subfigure 2.3c the probability increases to $p_{\text{BA},3} = 3/8$, since the number of degrees on node 3 increases.

Another prominent random graph model was proposed by Watts and Strogatz [43], who observed that quite often there was high clustering and short paths that were inconsistent with ER graphs in a seemingly random real-world network. They suggested introducing irregularity gradually into a network by rewiring a regular graph, moving it from a not-random, regular graph to becoming a random graph, such as those based on the ER-model. This allows for a scale of randomness in graph between a regular and ER-graph. Rewiring reduces the length of the average shortest path and increase clustering. They proposed a graph model now known as the Watts-Strogatz (WS) or 'Small World' model, the name deriving from the Small World experiment by Milgram. These graphs are created by starting from a regular ring structure with an average degree of d_{WS} as depicted in Figure 2.4a. Rewiring is then used to change the current links attached to individual nodes with the probability p_{WS}, the link to replace the current one is also chosen randomly. An example of WS random graphs are shown in 2.4b and 2.4c.



(a) Without rewiring, $p_{\mathrm{WS}} = 0$. (b) Rewiring with $p_{\mathrm{WS}} = 0.2$. (c) Rewiring with $p_{\mathrm{WS}} = 0.7$

Figure 2.4: An example of a WS graph with $N = 5$ nodes and a mean degree of $k_{WS} = 2$, demonstrating the rewiring process.

2.2. Epidemics on Networks

Epidemiology is a field that has gained a tremendous amount of prominence in recent years. Epidemiology studies the spread and effect of epidemics and can for example be used to predict the total number of expected infections and possible deaths due to a pandemic outbreak. Various epidemic models have been proposed and studied, providing differing levels of accuracy and complexity.

Bernoulli introduced the first epidemic model [5] in 1760, to analyse the efficacy of inoculation via an early form of immunisation. The mathematics he presented in his model used differential equations to show that inoculation would increase the average life expectancy for everyone, regardless of age, despite the fact inoculation had a significant chance of causing death.

A fundamental concept in modern epidemiology is to separate the population into two or more compartments. The four most commonly used compartments are *susceptible* (S), *exposed* (E), *infected* (I) and *removed* (R). The motivation for the compartmental model is that a set of differential equations can be formulated detailing the transition rates between each compartment. Solving the set of differential equations produces epidemic curves that predict the total number of individuals in each compartment. Early epidemiology did not consider a contact network, and each compartment was considered as a *homogeneous* population with indistinguishable and exchangeable individuals. The first study to use compartments to produce a deterministic model of an epidemic was published in 1927 by Kermack and McKendrick [21]. The solution to the model did indeed reproduce the curve of the infections in the epidemic.

An early example of taking into account the contact network of a population was the survey conducted on Cholera outbreak in London [39]. It was demonstrated that using data gathered from infected individuals could be enough to isolate the cause of a the disease, in this case an infected water source was found. Modern epidemiology also includes a focus on treating individuals or groups as discrete entities connected to each other as a network. Thus, modern epidemiology strongly relies on graph theory and network science.

Super-spreaders are a real world effect observed in certain individuals or events in pandemics [40], computer systems and even in social networks with the concept of influencers [6], where certain nodes have a significantly larger impact on the total number of infections than others. This can be due to a considerably higher degree on these nodes or a particularly high values for the infection rate to adjacent nodes [30].

There is a very broad range of problems that can be described by epidemic models, beyond infectious diseases. This includes the propagation of malware across an established computer network such as the internet. A cybersecurity department could know probabilistically what malware could be causing problems by having an estimate of the current prevalence of the malware. Observations such as reports by computers that have detected the malware could provide enough information to create a model and generate estimates of the origin and the danger posed by malware to computers in the same network.

We consider exact networks, where the individual links and nodes are estimated. The links and nodes for a network in an epidemic are often weighted to reflect the strength of infection and curing. The weights can be identical, such as for an epidemic on a homogenous network [34]. Alternatively it can vary depending on the link or node such as in a heterogeneous network. The advantage of heterogeneous parameters is that it can more accurately reflect a real world network such as twitter. We can vary the infection probability between users as the susceptibility or the chance of a user believing a certain news article will vary based on the source. Therefore, a node i will vary the chance of infection based on the source node j. For the case of curing, we can also assume a user will verify news differently leading to

a heterogeneous curing probability, where each node can have a unique curing probability.

3

Discrete-Time SIS Epidemic Model

Various models are used in epidemiology where nodes are classed into different compartments [31, 28, 22], with a node representing an individual in a population. These compartments include: infected (I), when a node carries the disease; and susceptible (S), when the node is not infected however is able to gain the disease. The viral states considered in this thesis are S and I. In the SIS model, nodes can be reinfected once they are cured, hence they can transition from S to I and from I to S.

The SIS model describes several real-world situations well where a node changes its viral state several times such as a reoccurring infectious disease [2]. Further applications of the SIS model include modelling networks where misinformation is being spread. Here each individual shares, or chooses not to share, several different blog posts or tweets.

A *Markovian* SIS process simplifies probability calculations within the model as there is no memory of all previous states when calculating the probabilities of each step, and all transitions are independent of each other[1]. The Markovian SIS process thus allows us to work with independent probabilities for transitions for the viral state of individual nodes. A sequence of transitions for the viral state vector $x[k]$, $X[n]$ is a sequence of independent events with independent probabilities and a total probability that is the product of the individual probabilities. A node can be in either one of two viral states: susceptible or infected. There are two transition rates which determines how quickly a node gets infected or cured. A continuous-time process is often approximated using discrete time. Real-world observations of epidemics are recorded in discrete time such as the results of tests being reported periodically. This means that whilst a process may be continuous for the spread of an infectious disease, the analysis is done in discrete time. For social media, iterations are limited by physical factors that make the process itself a discrete-time process with a minimum size time step. This is due to set refresh or tick-rate for a server which means any interactions done will take at least the refresh rate for the system to change state.

3.1. The SIS process

For the discrete-time SIS process, each node i has a defined state at any time k where $x_i[k] = 1$ and $x_i[k] = 0$ indicate an observation of node as infected or healthy respectively:

$$x_i[k] \in \{0, 1\}, \ \forall i \in \mathcal{N}, \forall k \ \in \mathbb{N}_n \tag{3.1}$$

Where $\mathbb{N}_n$ is the set of integers from 0 to n, the total number of observations. For a network with N nodes we define the $N \times 1$ viral state vector $x[k] = (x_1[k], ..., x_N[k])^T$ and hence $x[k] \in \{0, 1\}^N$. A node i can be infected by any node j that is adjacent and infected, where a link exists, $(i, j) \in \mathcal{L}$ and and $x_j[k] = 1$.

We can calculate the probability of a particular viral state sequence given some parameters, $\Pr[X[k]|\theta]$ in terms of the transitions between states, the values of δ and B form the parameter vector for all nodes, $\theta = (\delta_1, ..., \delta_N, \beta_{11}, ..., \beta_{NN})$. To define the transition probabilities for each node, we make two assumptions.

[1]Non-Markovian processes can also be used to generate observations for an SIS process [42].

1. The transitions between states, $\{x[k-1] \to x[k], k = 2, \ldots, n\}$ are Markovian. This implies the probability of a transition to the current state, $\Pr[x[k-1] \to x[k]]$, is only a function of the previous observation $x[k-1]$ and the parameter set θ. Concretely, this gives $\Pr[x[k-1] \to x[k]] = \Pr[x[k]|x[k-1], \theta]$.

2. The process $\{x_i[k-1] \to x_i[k]\}$ is independent of $\{x_j[k-1] \to x_j[k]\}$ for any two nodes, where $i \neq j$, at any fixed time k.

Using assumptions 1 and 2 we can describe the SIS process as a Markov chain, where the next state $x[k+1]$ of the system only depends on the current state $x[k]$. This means transitions between states in the Markov chain are independent and the total probability of series of transition can be calculated as the product of the the transitions at different times k.

We condition the probability of the next state based on the parameters and the observed current state $x_i[k]$. This allows us to define four transition probabilities:

1. I → S, curing of an infected node, $\Pr[x_i[k+1] = 0 | x_i[k] = 1, \theta] = \delta_i$, where δ_i is the curing rate of node i.

2. I → I, nodes remains infected, $\Pr[x_i[k+1] = 1 | x_i[k] = 1, \theta] = 1 - \delta_i$. As the node will either be cured or remain infected, the events of curing and not curing are mutually exclusive so we can define this probability based on the transition in 1.

3. S → I, infection of a healthy node, $\Pr[x_i[k+1] = 1 | x_i[k] = 0, x[k], \theta] = \sum_{j=1}^{N} x_j[k]\beta_{ij}$. We need to take into account the sum of all adjacent and infected nodes, $\sum_{j=1}^{N} x_j[k]\beta_{ij}$, as each link to node i has a probability of causing an infection.

4. S → S, node remain healthy, $\Pr[x_i[k+1] = 0 | x_i[k] = 0, x[k], \theta] = 1 - \sum_{j=1}^{N} x_j[k]\beta_{ij}$. The events of infection and remaining healthy are mutually exclusive for a node.

The model allows for multiple nodes to change state from $x_i[k]$ to $x_i[k+1]$ simultaneously. This is different to the model used in [34] where at most one node could change its viral state at time k.

The basic reproduction rate R_0 can be used to indicate the growth of the epidemic, generally the number of infected nodes should increase for $R_0 > 1$ and decrease for $R_0 < 1$ as this indicates the ratio of infection to curing. It remains an open question as to whether there is an exact epidemic threshold at $R_0 = 1$ when multiple node transitions occur simultaneously in the SIS process. To obtain R_0, we first define the effective infection probability matrix W with elements w_{ij} as:

$$w_{ij} = \frac{\beta_{ij}}{\delta_i} \quad \forall i \neq j \tag{3.2}$$

In line with [12, 18, 19], we define R_0 as the largest eigenvalue of W, the eigenvector z_k corresponds to the largest eigenvalue.

$$Wz_k = R_0 z_k \tag{3.3}$$

3.2. Transitions in the SIS process

We provide a detailed example of the Markov chain and transition matrices for a three-node network to demonstrate the complexity for the number of possible transitions. Further, we demonstrate the difference between single and multiple node transitions for complete and non-complete networks.

The state of each node can be represented by a segment in the circle, labelled with a single value in the binary sequence. In this case each of the three nodes has 2 states so the viral state vector is $x[k] \in \{0, 1\}^3$, thus the total number of viral states is eight. We compare the SIS model in Section 3.1 to the SIS model of [34] by illustrating the two methods of transitions between these eight states in Figures 3.2 and 3.3, which correspond to a change in state for a single node or multiple nodes, respectively.

We further show by example that a complete graph will allow for all the transitions between Markov states of any graph with the same number of nodes. This shows that the complete graph is always a possible solution, given a finite number of observations as we cannot guarantee that every possible transition or link has been used.

Figure 3.1 provides a network on which the SIS process can occur, with the parameters given in (3.4) and (3.5) for the infection and curing probabilities respectively. We visually represent the state of the system via the sectors of the circle where the colour of a section indicates the state of the node. This is used later to show the various Markov states of the network.

$$\mathrm{B} = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} \\ \beta_{21} & 0 & \beta_{23} \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix} \tag{3.4}$$

$$\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \tag{3.5}$$

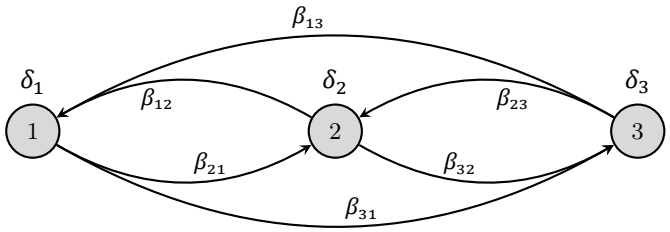
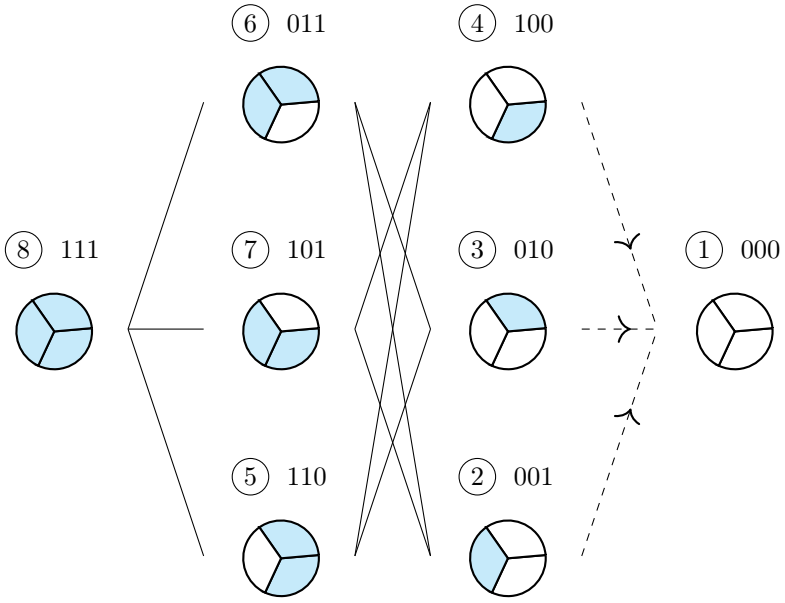


Figure 3.1: A 3 node complete graph where each the variables for infection and curing between nodes is described by the matrix B and vector δ from (3.4) and (3.5) respectively.

The SIS model in [34] considered a maximum of one node state change per time step, this approach is demonstrated in Figure 3.2. The approach here, as shown in Figure 3.3 allows for multiple transitions simultaneously.

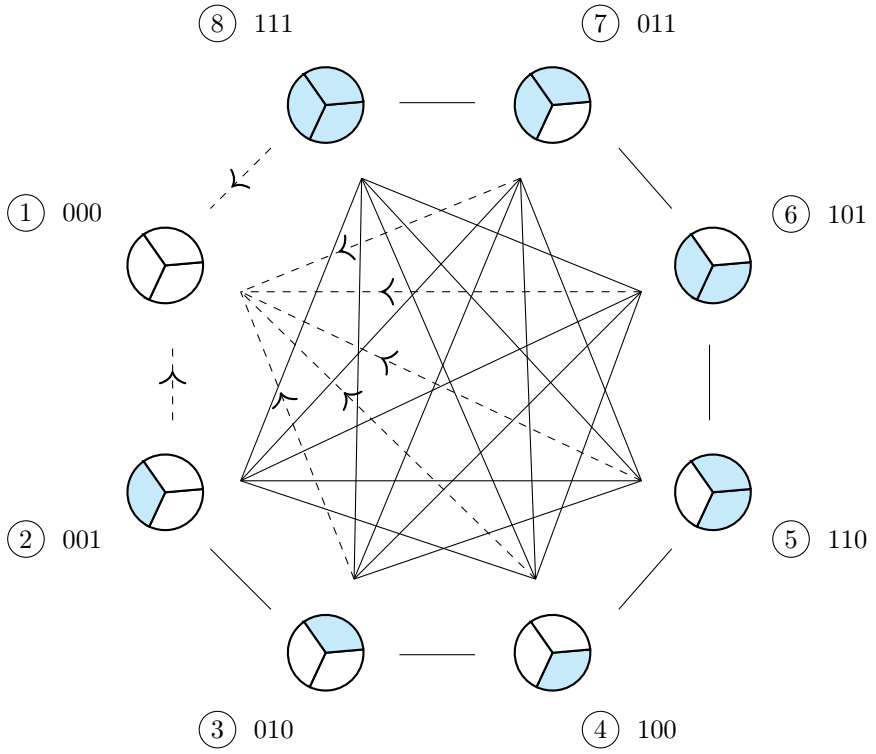


(a) SIS Markov chain of [34], where at most one node i changes its viral state $x_i[k]$ from time k to $k+1$. The states are labelled from the completely infection state (8) to the all-healthy state (1).

$$
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\delta_1 & 1-[\delta_1+\beta_{12}+\beta_{13}] & 0 & 0 & \beta_{12} & \beta_{13} & 0 & 0 \\
\delta_2 & 0 & 1-[\delta_2+\beta_{23}+\beta_{31}] & 0 & \beta_{23} & 0 & \beta_{21} & 0 \\
\delta_3 & 0 & 0 & 1-[\delta_2+\beta_{13}+\beta_{32}] & 0 & \beta_{31} & \beta_{32} & 0 \\
0 & \delta_2 & \delta_1 & 0 & 1-[\delta_2+\delta_3+\beta_{21}+\beta_{31}] & 0 & 0 & \beta_{21}+\beta_{31} \\
0 & \delta_3 & 0 & \delta_1 & 0 & 1-[\delta_3+\delta_1+\beta_{32}+\beta_{12}] & 0 & \beta_{32}+\beta_{12} \\
0 & 0 & \delta_3 & \delta_2 & 0 & 0 & 1-[\delta_3+\delta_2+\beta_{23}+\beta_{12}] & \beta_{23}+\beta_{12} \\
0 & 0 & 0 & 0 & \delta_1 & \delta_2 & \delta_3 & 1-[\delta_1+\delta_2+\delta_3]
\end{bmatrix}
$$

(b) Transition probability matrix P, whose elements P_{ij} represent the transition from state j to state i.

Figure 3.2: The Markov chain of the various states of the SIS process for the network in Figure 3.1. At most one node can be infected or cured in each time step. The solid lines indicate the two-way transitions between states. The dashed lines indicate a one-way transition to the all-healthy state, labelled 000, where the disease has died out.



(a) SIS Markov chain of section 3.1 for the complete graph, where multiple nodes can change their viral state $x_i[k]$ from time k to $k+1$ simultaneously.

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\delta_1 & 2-[(1+\delta_1)(1+\beta_{12})(1+\beta_{13})] & \delta_1\cdot\beta_{12} & \delta_1\cdot\beta_{13} & \delta_1\cdot\beta_{12}\cdot\beta_{13} & \beta_{13} & \beta_{12} & \beta_{12}\cdot\beta_{13} \\
\delta_2 & \delta_2\cdot\beta_{21} & 2-[(1+\delta_2)(1+\beta_{21})(1+\beta_{23})] & \delta_2\cdot\beta_{23} & \beta_{21} & \delta_2\cdot\beta_{21}\cdot\beta_{23} & \beta_{23} & \beta_{21}\cdot\beta_{23} \\
\delta_3 & \delta_3\cdot\beta_{31} & \delta_3\cdot\beta_{32} & 2-[(1+\delta_3)(1+\beta_{31})(1+\beta_{32})] & \beta_{32} & \beta_{31} & \delta_3\cdot\beta_{31}\beta_{32} & \beta_{31}\cdot\beta_{32} \\
\delta_2\cdot\delta_3 & \delta_2\cdot\delta_3\cdot(\beta_{21}+\beta_{31}) & \delta_3 & \delta_2 & 2-[(1+\delta_2)(1+\delta_3)(1+\beta_{21}+\beta_{31})] & \delta_2\cdot(\beta_{21}+\beta_{31}) & \delta_2\cdot(\beta_{31}+\beta_{21}) & \beta_{21}+\beta_{31} \\
\delta_1\cdot\delta_3 & \delta_3 & \delta_1\cdot\delta_3\cdot(\beta_{12}+\beta_{32}) & \delta_1 & \delta_1\cdot(\beta_{12}+\beta_{32}) & 2-[(1+\delta_1)(1+\delta_3)(1+\beta_{12}+\beta_{32})] & \delta_2\cdot(\beta_{12}+\beta_{32}) & \beta_{12}+\beta_{32} \\
\delta_1\cdot\delta_2 & \delta_2 & \delta_1 & \delta_1\cdot\delta_2\cdot(\beta_{13}+\beta_{23}) & \delta_1\cdot(\beta_{13}+\beta_{23}) & \delta_2\cdot(\beta_{13}+\beta_{23}) & 2-[(1+\delta_1)(1+\delta_2)(1+\beta_{13}+\beta_{23})] & \beta_{13}+\beta_{23} \\
\delta_1\cdot\delta_2\cdot\delta_3 & \delta_2\cdot\delta_3 & \delta_1\cdot\delta_3 & \delta_1\cdot\delta_2 & \delta_3 & \delta_2 & \delta_1 & 2-[(1+\delta_1)(1+\delta_2)(1+\delta_3)]
\end{bmatrix}$$

(b) Transition probability matrix P, whose elements P_{ij} represent the transition from state j to state i.

Figure 3.3: Infection and curing for the discrete-time SIS process where multiple nodes can change their viral state in each time step, for the network in Figure 3.1.

We give another example of the SIS process for an alternative graph to the complete graph. More specifically, we consider the path graph in Figure 3.4. Any transitions that required the links β_{13} and β_{31} are no longer possible, these changes are highlighted in the matrix entries in (3.6). Comparing the Markov chains of Figure 3.2b and Figure 3.3b demonstrates that many of the same walks are possible for both the complete and non-complete graph. Thus, it is difficult to differentiate the complete graph and the graph in Figure 3.4, unless one of the transitions such as from state (2) to (8) occurs.

$$\mathrm{B} = \begin{bmatrix} 0 & \beta_{12} & 0 \\ \beta_{21} & 0 & \beta_{23} \\ 0 & \beta_{32} & 0 \end{bmatrix} \tag{3.6}$$

$$\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \tag{3.7}$$

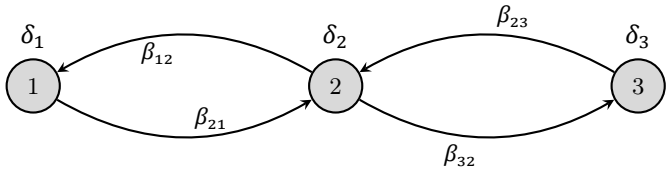
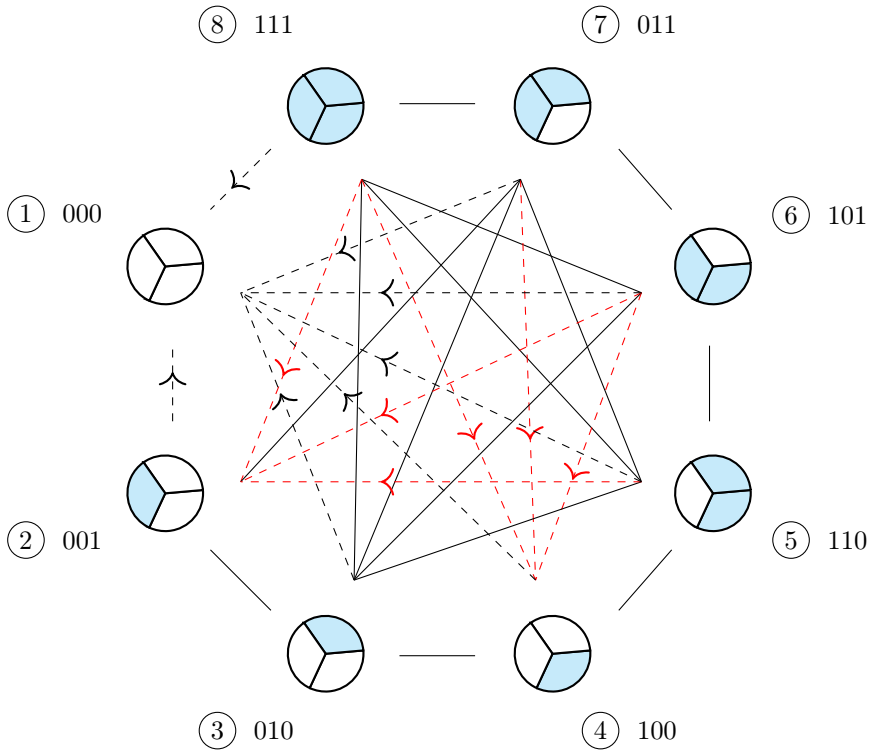


Figure 3.4: A 3 node line graph example. The diagonal for the B matrix is always 0, for other elements it depends on the existence of a link as given in Figure 3.6.



(a) SIS Markov chain of section 3.1 for the graph in Figure 3.4, where multiple nodes can change their viral state $x_i[k]$ from time k to $k+1$ simultaneously. The states are labelled with the circled numbers.

$$
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\delta_1 & 1-[\delta_1+\delta_1\cdot\beta_{12}+\beta_{12}+\beta_{12}\cdot\beta_{13}] & \delta_1\cdot\beta_{12} & 0 & 0 & 0 & \beta_{12} & \beta_{12}\cdot\beta_{13} \\
\delta_2 & \delta_2\cdot\beta_{21} & 2-[(1+\delta_2)(1+\beta_{21})(1+\beta_{23})] & \delta_2\cdot\beta_{23} & \beta_{21} & \delta_2\cdot\beta_{21}\cdot\beta_{23} & \beta_{23} & \beta_{21}\cdot\beta_{23} \\
\delta_3 & 0 & \delta_3\cdot\beta_{32} & 1-[\delta_2+\delta_2\cdot\beta_{32}+\beta_{32}+\beta_{31}\cdot\beta_{32}] & \beta_{32} & 0 & 0 & \beta_{31}\cdot\beta_{32} \\
\delta_2\cdot\delta_3 & \delta_2\cdot\delta_3\cdot(\beta_{21}+\beta_{31}) & \delta_3 & \delta_2 & 2-[(1+\delta_2)(1+\delta_3)(1+\beta_{21}+\beta_{31})] & \delta_2\cdot(\beta_{21}+\beta_{31}) & \delta_2\cdot(\beta_{31}+\beta_{21}) & \beta_{21}+\beta_{31} \\
\delta_1\cdot\delta_3 & \delta_3 & \delta_1\cdot\delta_3\cdot(\beta_{12}+\beta_{32}) & \delta_1 & \delta_1\cdot(\beta_{12}+\beta_{32}) & 2-[(1+\delta_1)(1+\delta_3)(1+\beta_{12}+\beta_{32})] & \delta_2\cdot(\beta_{12}+\beta_{32}) & \beta_{12}+\beta_{32} \\
\delta_1\cdot\delta_2 & \delta_2 & \delta_1 & \delta_1\cdot\delta_2\cdot(\beta_{13}+\beta_{23}) & \delta_1\cdot(\beta_{13}+\beta_{23}) & \delta_2\cdot(\beta_{13}+\beta_{23}) & 2-[(1+\delta_1)(1+\delta_2)(1+\beta_{13}+\beta_{23})] & \beta_{13}+\beta_{23} \\
\delta_1\cdot\delta_2\cdot\delta_3 & \delta_2\cdot\delta_3 & \delta_1\cdot\delta_3 & \delta_1\cdot\delta_2 & \delta_3 & \delta_2 & \delta_1 & 2-[(1+\delta_1)(1+\delta_2)(1+\delta_3)]
\end{bmatrix}
$$

(b) Probability Matrix, P with element P_{ij} representing the transition from state j to state i.

Figure 3.5: Infection and curing for the discrete time SIS process where multiple nodes can change state in each time step, for the network in Figure 3.4. Differences of the complete network Markov chain from Figure 3.3 are given in red.

4

Network Reconstruction for the SIS Epidemic Model

Network reconstruction involves obtaining the underlying adjacency matrix, A and potentially the node and link characteristics such as curing and infection rates or probabilities. This is achieved by taking measurements of the sequence of observations of the changes in viral state, $X[n]$. Current state-of-the-art methods used of for network reconstruction involves using either a Maximum A-Posteriori (MAP) or the related Maximum Likelihood (ML) estimators to obtain an estimate for the parameters of the network from the observations $X[n]$.

MAP estimation requires prior knowledge of the model used to produce the graph for example whether it is directed or undirected and whether the links were generated using the BA, ER or WS model from Section 2.1.2. It has been shown that while MAP estimation is at least as accurate as ML estimation, the tendency for both is the same for the growth in the number of observations required to estimate a network of size N [34]. Additionally, ML estimation was sufficient to establish variation in the number of observations required to gain a level of accuracy for exact reconstruction.

4.1. Maximum Likelihood Estimation

For Maximum Likelihood estimation, we aim to find the most probable parameter set from the value of θ that maximises the likelihood function (4.1), where $X[n] = (x[1], \ldots, x[n])$ is the $N \times n$ matrix of observations.

$$\underset{\theta}{\operatorname{argmax}} \ \Pr[X[n]|\theta] \tag{4.1}$$

The likelihood function is the probability for a viral state sequence $X[n]$ conditioned on the parameter set of the network θ. We take the value of $\hat{\theta}$ that maximises $\Pr[X[n]|\theta]$ as our estimate of θ. The specific parameters we include within θ are given in (4.2).

$$\theta = (\beta_{11}, \ldots, \beta_{NN}, \delta_1, \ldots \delta_N) \tag{4.2}$$

Assumption 1 states that subsequent changes to viral state are independent. The probability of the viral state sequence up to time n is the product of probability of each individual transitions

$$\Pr\left[X[n]\middle|\theta\right] = \Pr\left[x[1]\middle|x[0], \theta\right] \cdot \Pr\left[x[2]\middle|x[1], \theta\right] \cdot \ldots \cdot \Pr\left[x[n]\middle|x[n-1], \theta\right] \tag{4.3}$$

For the whole network, we can use Assumption 2 where all transitions on different nodes that occur simultaneously are independent so the total probability for all nodes is the product of transitions of individual nodes. This allows the total probability for all observations of all nodes $i \in \mathcal{N}$ to be calculated as

$$\begin{aligned}\Pr\left[X[n]\middle|\theta\right] &= \prod_{k=0}^{n-1} \Pr\left[x[k+1]\middle|x[k],\theta\right] \\ &= \prod_{k=0}^{n-1}\prod_{i=1}^{N} \Pr\left[x_i[k+1]\middle|x[k],\theta\right]\end{aligned} \tag{4.4}$$

We aim to find the maximum

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}} \prod_{k=0}^{n-1}\prod_{i=1}^{N} \Pr\left[x_i[k+1]\middle|x[k],\theta\right] \tag{4.5}$$

To make this more easily computable, it can be expresses as the sum of negative logarithms

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmin}} \sum_{k=0}^{n-1}\sum_{i=1}^{N} -\log\left(\Pr\left[x_i[k+1]\middle|x[k],\theta\right]\right) \tag{4.6}$$

with the constraints

$$0 \leq \beta_{ij} \leq \beta_{\max} \;\forall\; i \neq j \tag{4.7}$$

$$0 \leq \delta_i \leq \delta_{\max} \;\forall\; i \tag{4.8}$$

The objective function for each node i is given in (4.9), which we derive in Appendix A.

$$\begin{aligned}\Pr\left[x_i[k+1]\middle|x[k],\theta\right] =& x_i[k]\left(x_i[k+1] + (1-2x_i[k+1])\delta_i\right) \\ &+ (1-x_i[k])\left((1-x_i[k+1]) + (2x_i[k+1]-1)\left(\sum_{j=1}^{N} x_j[k]\beta_{ij}\right)\right)\end{aligned} \tag{4.9}$$

To solve the optimisation problem (4.6), we use the `fmincon` function [25] in MATLAB, which uses an existing interior point algorithm designed for large-scale non-linear programming [7]. Since (4.6) is a convex optimisation problem, the interior point algorithm could guaranteeably find the global minimum [20].

Concretely, the problem is to determine the values of δ and B, given that the observation set $X[n]$ was generated. In an unweighted network where $\beta_{ij} \in \{0, \beta_{\max}\}$, finding the minimum of (4.9) is NP-hard [35]. In this case, a brute force method quickly runs into complexity issues as this method would consider all possible networks that can be generated from N nodes, that is $2^{L_{\max}}$ where, $L_{\max} = \binom{N}{2}$ is the maximum number of links. By considering a weighted network we reduce the complexity so the problem is convex and solvable in polynomial time as the values of β_{ij} are continuous in the range $\beta_{ij} \in [0, \beta_{\max}]$.

The estimation process requires several runs for each network size to increase the reliability of an estimate of the error in the calculated values of δ_i or β_{ij}. The observations are generated probabilistically, it can not be guaranteed that δ, and B are the most likely values to generate the observations and so we aim to calculate a reliable average error in the estimate. There is a chance of being unlucky and that the observations generated are not very representative of the parameters. For example if the probability of a node becoming infected is 99% there us no guarantee that the observation generated will be an infection, 99% of the time. Similarly, if 99 out of 100 times, a node got infected by the same adjacent node, the link weight was not necessarily 0.99, this was simply the best estimate.

4.1.1. Maximum a posteriori estimation for uniform prior distributions

For the case of a uniform prior, we demonstrate MAP estimation is equivalent to ML estimation. Starting from Bayes' theorem with the prior probability $f_\theta(\Theta)$, defined as the probability distribution function of the parameter set θ; the likelihood function $\Pr[X[n]|\theta]$ and a posterior probability $f_{\Theta|X[n]}(\theta)$, the probability distribution function of the parameter set θ conditioned on the viral state sequence $X[n]$. This gives the overall formula for the posterior distribution

$$f_{\Theta|X[n]}(\theta) = \frac{\Pr[X[n]|\theta] \cdot f_\theta(\Theta)}{\Pr[X[n]]} \tag{4.10}$$

Given that the probability of the observation set $\Pr[X[n]]$ is not a function of the parameters and the prior distributions in 4.11 are uniform and independent therefore $f_\Theta(\theta)$ is constant over the ranges in (4.7) and (4.8).

$$f_\Theta(\theta) = f_\Theta(\beta_{11}) \cdot \ldots f_\Theta(\beta_{NN}) \cdot f_\Theta(\beta_{11}) \cdot f_\Theta(\delta_1) \cdot \ldots \cdot f_\Theta(\delta_N) = \text{constant} \tag{4.11}$$

This shows that MAP estimation and ML estimation provide the same estimate $\hat{\theta}$

$$\underset{\theta}{\text{argmax}}\ \Pr[X[n]|\theta] = \underset{\theta}{\text{argmax}}\ f_{\Theta|X[n]}(\theta) = \hat{\theta} \tag{4.12}$$

4.2. Simulation Settings

Two different graph models are used to generate adjacency matrices for Barabási-Albert undirected and Erdős-Rényi directed graphs. We consider networks sizes from $N = 15$ to $N = 55$. For the Erdős-Rényi directed graphs, we use the link probability $p_{\text{ER}} = 0.3$ for all graphs. For the Barabási-Albert graphs, the initial complete graph has $m_0 = 3$ nodes with each additional node connecting to $m = 3$ nodes.

We initially obtain the curing and infection probabilities for simulating the SIS process. We generate 900 adjacency matrices A and infection probability matrices B' for each network size, where the elements for B' are generated uniformly at random in the range $[0, \beta_{\max}]$. We then take the element-wise product of A and B' to obtain B. We generate 900 vectors $\delta = (\delta_1, \ldots, \delta_N)^T$ of the curing probabilities uniformly at random in the range $[0, \delta_{\max}]$. We can therefore obtain 900 parameter sets θ with which to generate observations $X[n]$.

The curing probability for each node is generated from the distribution

$$\delta_i \sim U(0, \delta_{\max}), \quad i \in \mathcal{N} \tag{4.13}$$

The infection probabilities for single links are random variables that have a uniform distribution, where the maximum value is $\beta_{\max} = 1/(N-1)$. This means the maximum value of the infection probability incident on a node, the weighted degree β_i is 1, where $\beta_i = \sum_{j=1}^{N} \beta_{ij}$.

$$\beta_{ij} \sim U(0, \beta_{\max}), \quad (i,j) \in \mathcal{L} \tag{4.14}$$

The initial state is an all-one vector, $x[0] = u = (x_1[0] = 1, x_2[0] = 1, \ldots, x_N[0] = 1)^T$. The viral state where all the nodes are cured, $x[k] = ([x_1[k] = 0, x_2[k] = 0, \ldots, x_N[k] = 0)^T$ is an absorbing Markov state. In order for the process network to continue the SIS process, the health status of all the nodes is reset so that $x[k+1] = u$ if $x[k] = 0$. The next viral state is generated using Algorithm 3 from Appendix B. This is repeated for 10^4 times to simulate $n = 10^4$ observations.

The basic reproduction number R_0 is specified by the infection probability matrix B and the curing probability δ, see (3.3). We want to set the basic reproduction rate ourselves, $R_0 = R_{0,\text{required}}$ so we can choose a value that maximises the accuracy of the estimated network. The effective infection probability of each link has to be scaled accordingly from the current value, $R_{0,\text{current}}$.

$$w_{ij,\text{required}} = w_{ij,\text{current}} \cdot \frac{R_{0,\text{required}}}{R_{0,\text{current}}} \tag{4.15}$$

We can adjust w_{ij} by keeping β_{ij} constant and varying δ_i, due to the relation in (3.2):

$$\delta_{i,\text{required}} = \delta_{i,\text{current}} \cdot \frac{R_{0,\text{current}}}{R_{0,\text{required}}} \tag{4.16}$$

Testing several sets for values of $R_0 = 0.6$ to $R_0 = 2.4$ can be done to establish if there is a value that increases the accuracy. We set the value $R_0 = 1.4$ as this value meant fewer observations could be used. Hence we consider the best case for network reconstruction and saves time for computing an estimate for a fixed level of accuracy.

4.3. Numerical Evaluation: Exact Network Reconstruction

The accuracy of an estimate of the adjacency matrix is measured using the AUC (Area Under Curve) metric. An AUC of 1 indicates a perfectly reconstructed network where every link and absence of a link is correctly identified. An AUC of 0.5 indicates that the average chance of an estimated link being correct is completely random and an AUC of 0 is where the graph estimated is the complement, or exact opposite of the actual graph. The AUC uses the area under an ROC (Receiver Operator Characteristic) curve which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The greater the area under the curve, the better the curve is at classifying the data correctly.

The estimation for each link was either a True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values given in Table 4.1. The fractions of links identified as TP or FN are used to calculate the True Positive Rate (TPR) via (4.17) and the fraction of links defined as FP or TN are used to calculate the False Positive Rate (FPR) via (4.18).

	Actual=1	Actual=0
Estimate=1	TP	FP
Estimate=0	FN	TN

Table 4.1: The table used to determine whether the estimate of link is a true positive, true negative, false positive or false negative.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.17}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4.18}$$

Figure 4.1 shows the greater the area under the curve, the better the curve is at classifying the data correctly. Each curve is generated by varying the threshold value that separates links into each of the four categories in Table 4.1. This is used to demonstrate the increase in accuracy of an estimate for a greater number of observations.

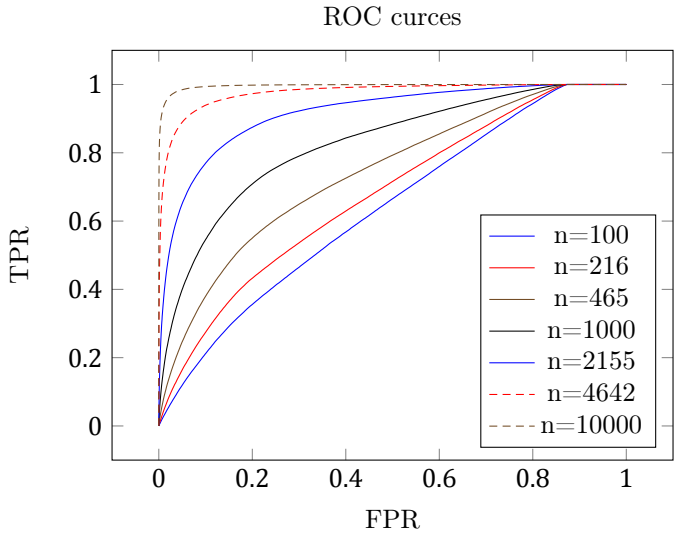
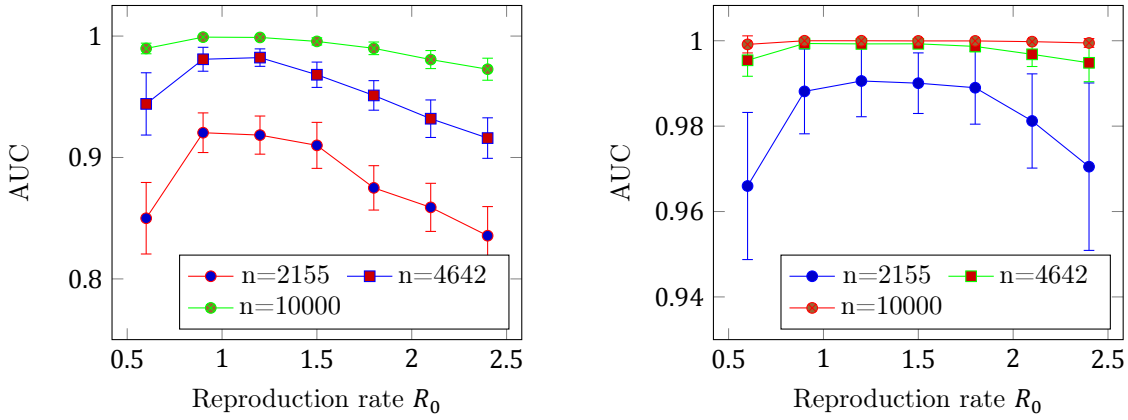


Figure 4.1: ROC curves for N = 55, where the area under each curve (AUC) compares the accuracy of the estimate of the adjacency matrix.

4.3.1. Reproduction rate

Figure 4.2 demonstrates the variation in AUC for different R_0 values. This indicated that a network with a value between 1.4 to 1.6 would have the best estimates of the adjacency matrix A. This may be due to the number of infection and curing observations being maximised for a R_0 value in that range.



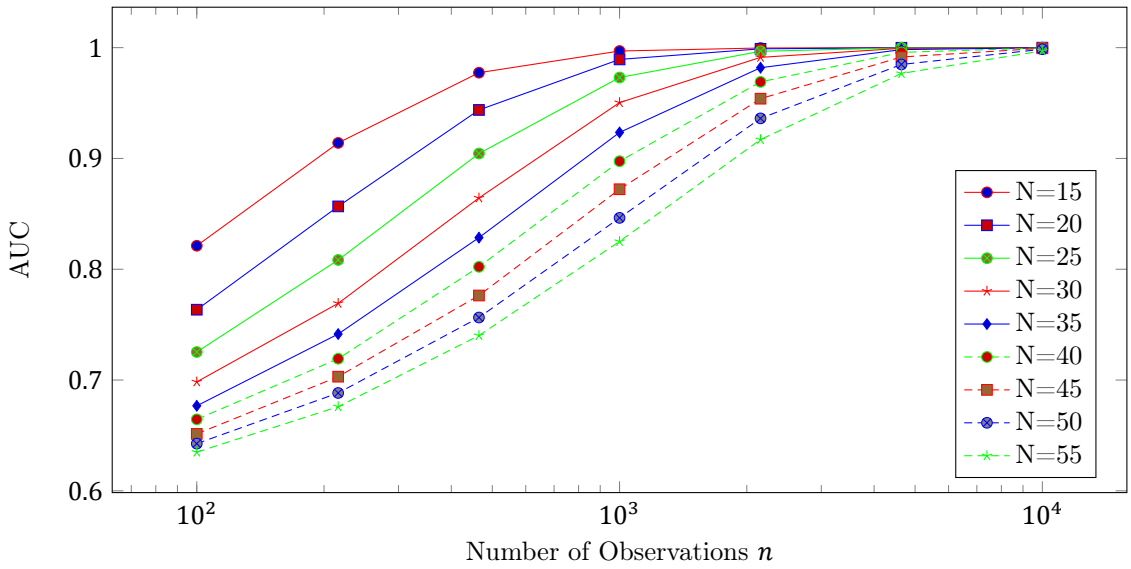
(a) Varying the reproduction rate for the Erdős-Rényi directed model.

(b) Varying the reproduction rate for the Barabási-Albert model.

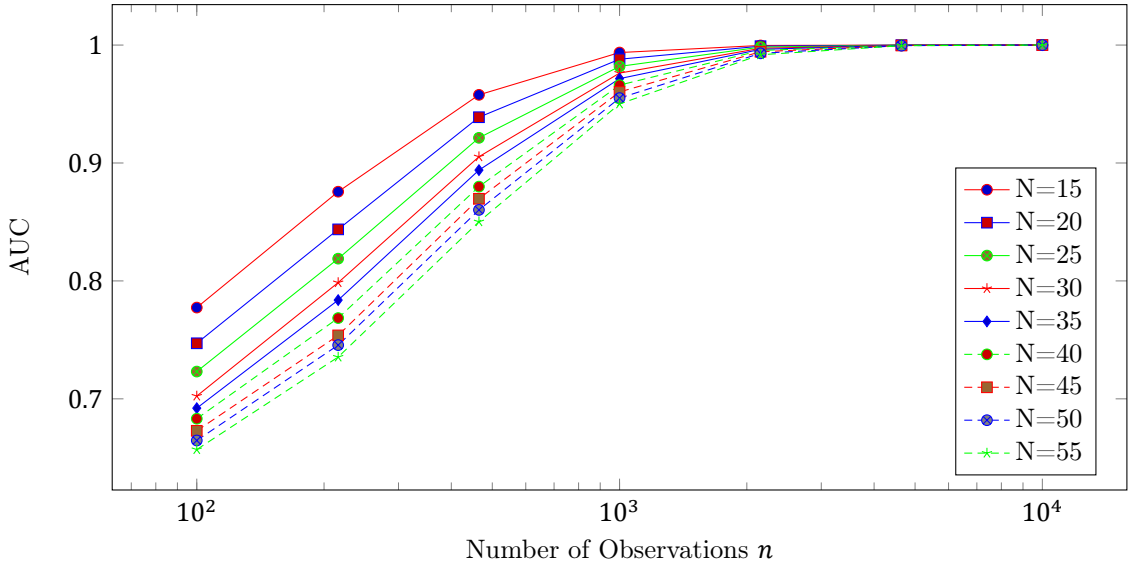
Figure 4.2: Graphs showing the effect of changing the basic reproduction number R_0 on the AUC of a reconstructed network of size $N = 55$.

4.3.2. Reconstruction accuracy versus the observation length

We evaluate the accuracy of the estimate of the adjacency matrix, giving the results in Figure 4.3. As expected, the AUC value converged quicker to AUC = 1 for a graph with a lower number of nodes, this is likely due to more link involvement for a set number of observations for a smaller network size, providing more data for an estimation for that link. The Erdős-Rényi directed graph converged more slowly which may be due to the degree distribution with fewer low degree nodes compared to a Barabási-Albert graph.



(a) AUC versus the number of observations n for a directed Erdős-Rényi graph.



(b) AUC versus the number of observations n for a Barabási-Albert graph.

Figure 4.3: Network reconstruction accuracy as a function of the number of observations n, for various network sizes, on a semi-logarithmic scale.

We present the absolute error of the node curing probability δ_i and the relative error in the link infection probability β_{ij} in Figures 4.4 and 4.5 for Erdős-Rényi directed and Barabási-Albert graphs, respectively. A relative error was used for as the larger the number of nodes N the smaller the value of $\beta_{\max} = 1/(N-1)$. A relative error meant the error in the estimate of a link for different network sizes could be compared.

We calculate the average error in curing probability for a node as

$$\eta_\delta = \frac{1}{N}\sum_{i=1}^{N} |\delta_i - \hat{\delta}_i| \tag{4.19}$$

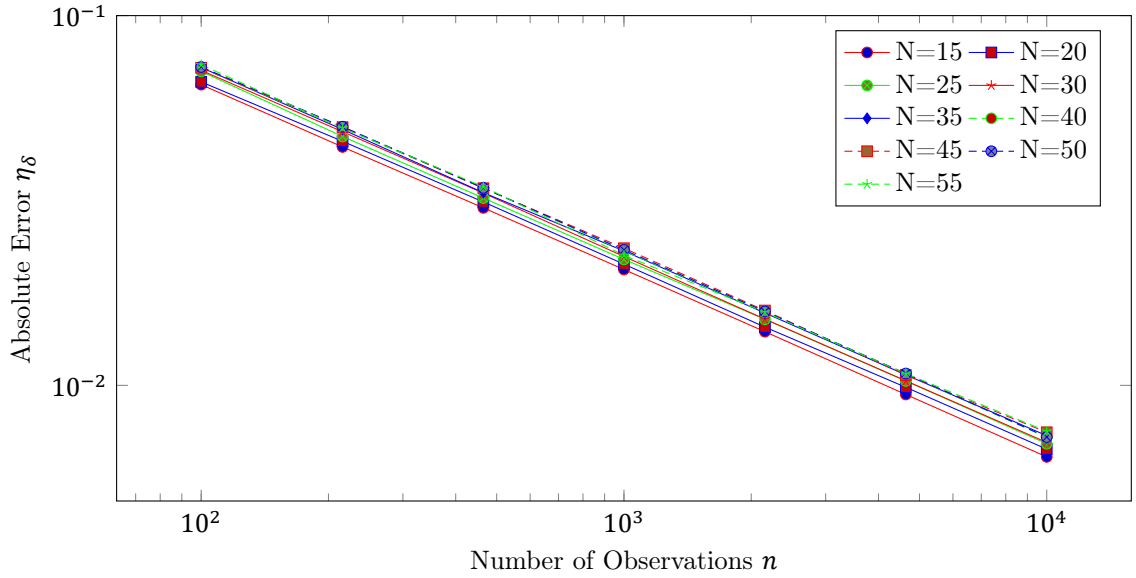
and the infection probability for a link

$$\eta_\beta = \frac{1}{(N-1)N}\sum_{i\neq j} \frac{|\beta_{ij} - \hat{\beta}_{ij}|}{\beta_{ij}} \tag{4.20}$$

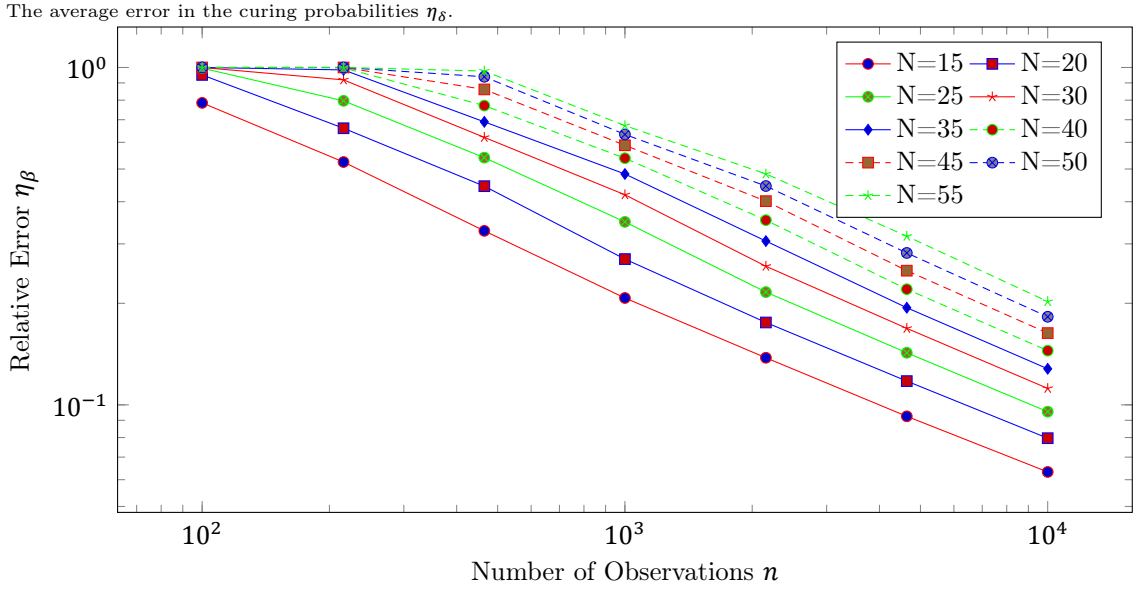
We can see in Figures 4.4a and 4.5a that over the range $n = 10^2$ to $n = 10^4$, the number of observations required to decrease the error in the curing probability estimate, increases at a greater than linear rate.

For the link infection error, η_β, Figures 4.4b and 4.5b show that for a smaller number of observations $n < 500$ for Erdős-Rényi graphs and $n < 300$ for Barabási-Albert graphs, larger networks $N > 35$ have a relative error close to 1. This may be due to the majority of the links remaining unused and estimated to be 0, leading to a relative error $\eta_\beta = 1$. In both cases the same exponential decrease is also observed in the link infection probability as for the curing probability, with $\eta_\beta \to 0$ as $n \to \infty$.

We can see that for η_δ there is not much of a reliance on network size on the accuracy on a estimate unlike η_β, where there is a clear increase in accuracy for smaller networks. The effect of network size on η_β is more pronounced for Erdős-Rényi directed graphs than Barabási-Albert graphs due to the scale-free nature of Barabási-Albert graphs where there is always a high proportion of low degree nodes, regardless of network size. We later analyse the effect of accuracy of an estimate of the link infection probability for a node i with degree d_i.

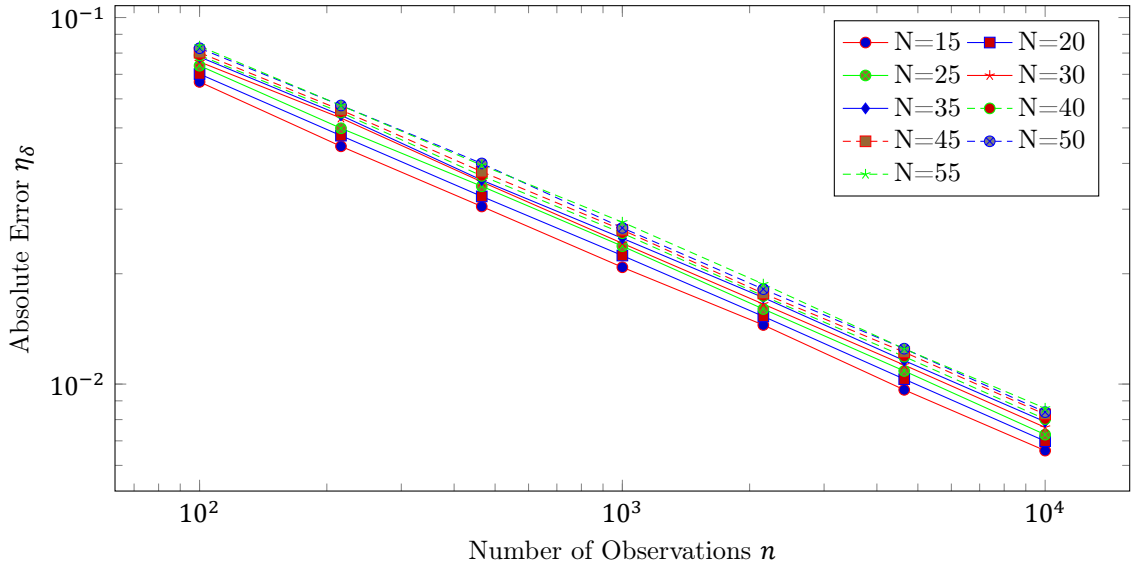


(a) The average error in the curing probabilities η_δ.

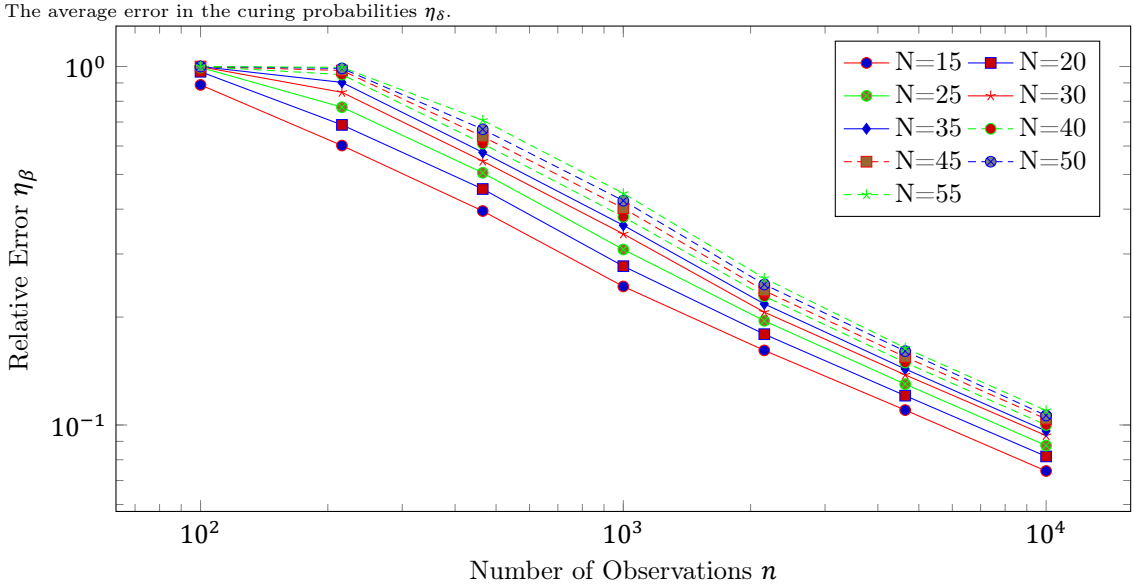


(b) The average error in the infection probabilities η_β.

Figure 4.4: Average error in infection and curing for individual nodes and links, η_δ and η_β, for an Erdős-Rényi directed graph, on a log-log scale.



(a) The average error in the curing probabilities η_δ.



(b) The average error in the infection probabilities η_β.

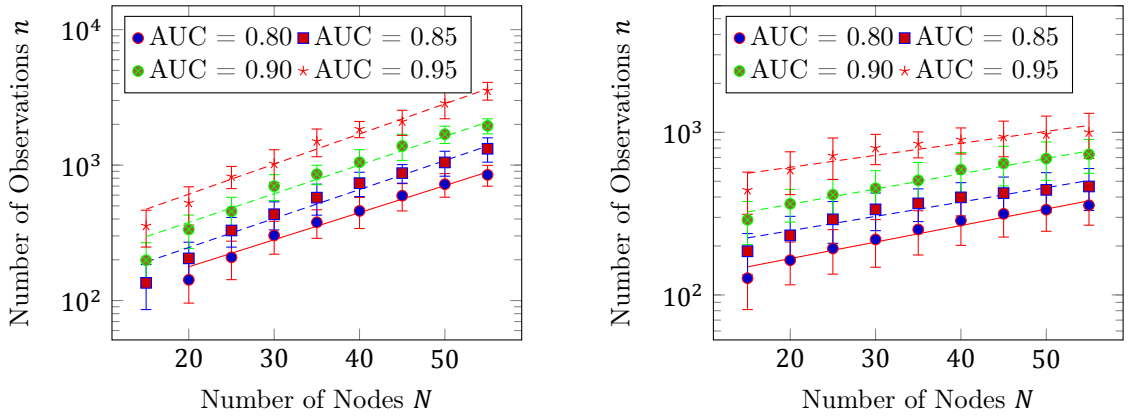
Figure 4.5: Average error in infection and curing for individual nodes and links η_δ and η_β for a Barabási-Albert graph on a log-log scale.

4.3.3. Required observation length for large networks

We apply an exponential fitting to the data to each sub-figure in 4.6, 4.7 and 4.8 in the form of (4.21), where the number of observations n is a function of the network size N for a given error. We obtain estimates for the values of C and g in (4.21) which can be used to compare the fitted function and extrapolate the number of observations from the network size for graphs of $N = 500$.

$$n(N) = C10^{gN} \tag{4.21}$$

Figure 4.6 show the exponential type increase in the number observations required for an AUC of 0.80, 0.85, 0.90 and 0.95. This shows that constructing the adjacency matrix to 80% accuracy and above may require an exponentially increasing number of observations. This would suggest a completely infeasible number of observations would be required for a pandemic even within a small, isolated population of $N = 5000$ would required somewhere in the region of $n > 10^{110}$ observations which would be completely infeasible to collect. Table 4.2 suggests that the number of observations required to for a network size of $N = 500$ would require very high number of observations, $n > 10^{12}$ for a level of accuracy AUC > 0.8 for Erdős-Rényi directed graphs and $n > 10^{7}$ for Barabási-Albert graphs. We see that the exponent g is approximately the same for different errors with $g \approx 0.5$ for Erdős-Rényi directed graphs and $g \approx 0.2$ for Barabási-Albert graphs, for differing levels of accuracy. There is some inaccuracy in the function approximated, which may be noted from Table 4.2b as the number of required observations for $N = 500$ is slightly lower that should be expected for AUC $= 0.85$ as it is lower than the value for AUC $= 0.8$.



(a) Interpolation of AUC for the Erdős-Rényi directed graphs. (b) Interpolation of AUC for Barabási-Albert graphs.

Figure 4.6: The number of observations n as a function of network size N. This indicates the function may be exponential suggesting that for larger network sizes, it would quickly become infeasible to get enough observations to reconstruct the adjacency matrix.

Error A	C	g	n(N=500)
0.95	222	0.022	$1.3 \cdot 10^{13}$
0.9	145	0.020	$4.6 \cdot 10^{12}$
0.85	92	0.021	$4.3 \cdot 10^{12}$
0.8	71	0.020	$6.6 \cdot 10^{11}$

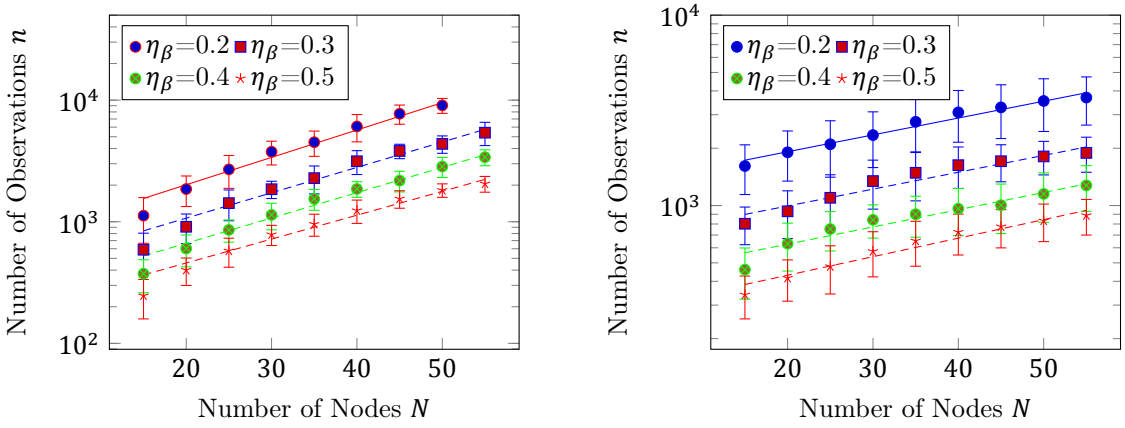
(a) Reconstruction of Erdős-Rényi directed graphs.

Error A	C	g	n(N=500)
0.95	434	0.007	$2.1 \cdot 10^{6}$
0.9	234	0.009	$1.2 \cdot 10^{7}$
0.85	165	0.009	$4.3 \cdot 10^{6}$
0.8	105	0.010	$1.3 \cdot 10^{7}$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.2: Table for fitting (4.21) to the calculated AUC to get C and g values for different errors. The number of required observations is extrapolated to $N = 500$.

Reconstructing the infection probabilities of links β_{ij} appears particularly challenging as the level of accuracy is very low. A relative error of $\eta_\beta = 0.1$ was not achieved for $n = 10^4$ observations. Figure 4.7 shows the increase in the number of observations. Table 4.3 shows that for a network with $N = 500$ nodes, we would require more than 10^7 observations. For an error of 0.5 in 4.3b however we can note that the extrapolated number of observations $n = 1.90 \cdot 10^7$ for a network of $N = 500$ nodes is greater than for an error of 0.3 and 0.4, which may be due to some error in the values of C and g for the function approximation. The exponent of the exponential fitting in Figure 4.7a for η_β in Erdős-Rényi graphs is similar to the one for the error in A as $g \approx= 0.5$ in Table 4.3a. The value of the exponent for the exponential fitting in Barabási-Albert graphs is also similar for the weighted and unweighted links, where we obtain $g \approx 0.2$ in Table 4.3b. This suggests the number of observations required for network reconstruction grows at the same rate for a given unweighted link error and weighted link error.



(a) Interpolation of η_β for th Erdős-Rényi graphs.

(b) Interpolation of η_β for Barabási-Albert graphs.

Figure 4.7: Average error in infection and curing for individual nodes and links

Error β_{ij}	C	g	n(N=500)
0.2	714	0.023	$1.3 \cdot 10^{14}$
0.3	410	0.021	$1.1 \cdot 10^{13}$
0.4	255	0.021	$6.4 \cdot 10^{12}$
0.5	185	0.020	$1.3 \cdot 10^{12}$

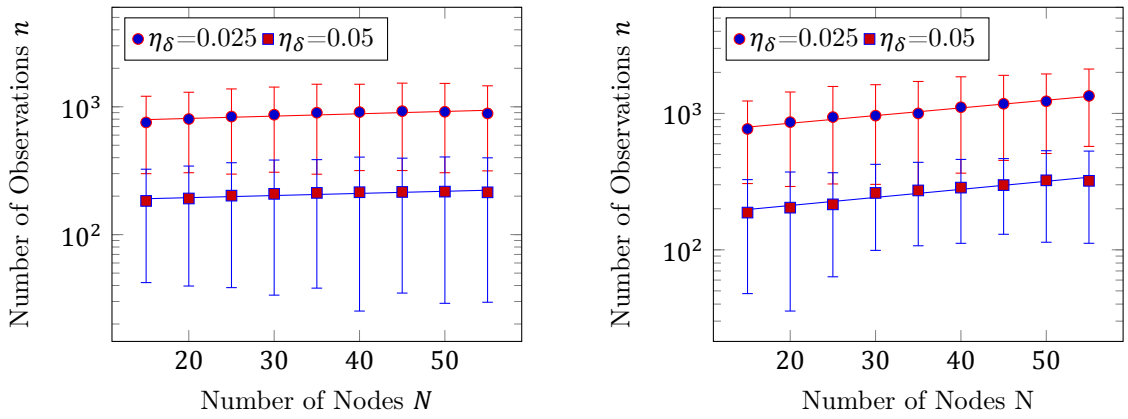
(a) Reconstruction of Erdős-Rényi directed graphs.

Error β_{ij}	C	g	n(N=500)
0.2	1289	0.009	$2.91 \cdot 10^{7}$
0.3	662	0.009	$1.71 \cdot 10^{7}$
0.4	415	0.009	$1.30 \cdot 10^{7}$
0.5	276	0.010	$1.90 \cdot 10^{7}$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.3: Table for fitting (4.21) to the estimates of β_{ij} to get C and g values for different errors. The number of required observations is extrapolated to $N = 500$.

Figure 4.8 suggests there is an exponential increase in the number of required observations to estimate the curing probability of a node δ_i however this increase is considerably weaker than for the AUC or link infection probability. The exponent obtained from the exponential fitting for Erdős-Rényi directed graphs given in Table 4.4a is $g \approx 0.004$ which is an order of magnitude less than for the exponent of the growth in the number of observations for the error in the adjacency and link infection probabilities. In Table 4.4b the exponent obtained is $g \approx 0.01$, which is half the value of the exponents in the fitting for the number of observations for a fixed error in the adjacency and link infection probabilities. The trend for the average error in the curing probability is less well defined as can be seen in Figure 4.8 where the standard deviation for estimating the number of observations n as a function of the network size N is much larger than for estimating the AUC or link infection probability. Table 4.4 demonstrates that 10^5 observations are required for Barabási-Albert graphs and 10^3 for Erdős-Rényi directed graphs. We can again note that there is some error in C and g in the function approximation as the number of observations for a network fo $N = 500$ is higher for an error of $\eta_\delta = 0.05$ compared to $\eta_\delta = 0.025$ in Table 4.4b.



(a) Interpolation of η_δ for Erdős-Rényi directed graphs.

(b) Interpolation of η_δ for the Barabási-Albert graphs.

Figure 4.8: Average error in curing probability for a node, for a constant error. This shows there is a weak dependence on the network size for evaluating the curing probability for a node.

η_δ	C	g	n(N=500)
0.025	745	0.002	$6.08 \cdot 10^3$
0.05	180	0.002	$1.22 \cdot 10^3$

(a) Trend in η_δ for Erdős-Rényi directed graphs.

η_δ	C	d	n(N=500)
0.025	654	0.006	$4.23 \cdot 10^5$
0.05	161	0.006	$1.45 \cdot 10^5$

(b) Trend in η_δ for Barabási-Albert graphs.

Table 4.4: Table for fitting (4.21) to η_δ to get C and g values for different errors. The number of required observations is extrapolated to $N = 500$.

4.4. Numerical Evaluation: Partial Network Reconstruction

It seems that reconstructing the whole network for a given level of accuracy is not possible for a feasible number of observations. An alternative could be to reconstruct parts of the network, such as a subset of links and nodes. This can be determined by separating the set of nodes $\mathcal{N}$ into nodes with a low, mid and high degree value the and set of links $\mathcal{L}$ into links incident on nodes with low, mid and high degree.

The accuracy of estimates of the link infection probability β_{ij}, the curing rate δ_i and the adjacency matrix A are calculated for nodes of different degrees. This demonstrates there is a significant improvement in the accuracy of estimates in lower degree nodes for β_{ij} and A meaning low degree nodes can be estimated for a given level of accuracy in larger networks. For curing rates however, the accuracy increases for higher degree nodes meaning that δ_i can be estimated for a given level of accuracy in larger networks.

Figure 4.9 demonstrates the increase in accuracy of an estimate of the adjacency of a node i the degree d_i of the node. The AUC in Figure 4.9 is based on the links incident to a node i that have been classified correctly as having a infection probability $\beta_{ij} > 0$ or $\beta_{ij} = 0$. We observe a decrease in accuracy for higher degree nodes in both Erdős-Rényi and Barabási-Albert graphs with $N = 55$ nodes, suggesting the trend may be independent of the type of random graph.

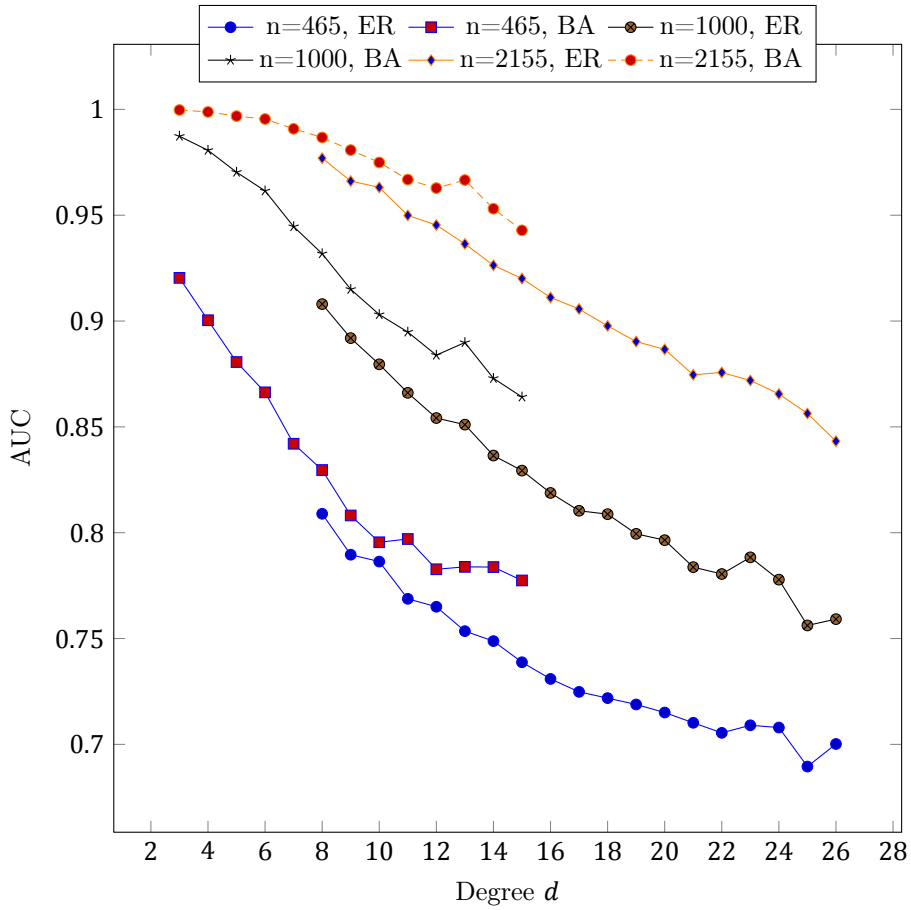
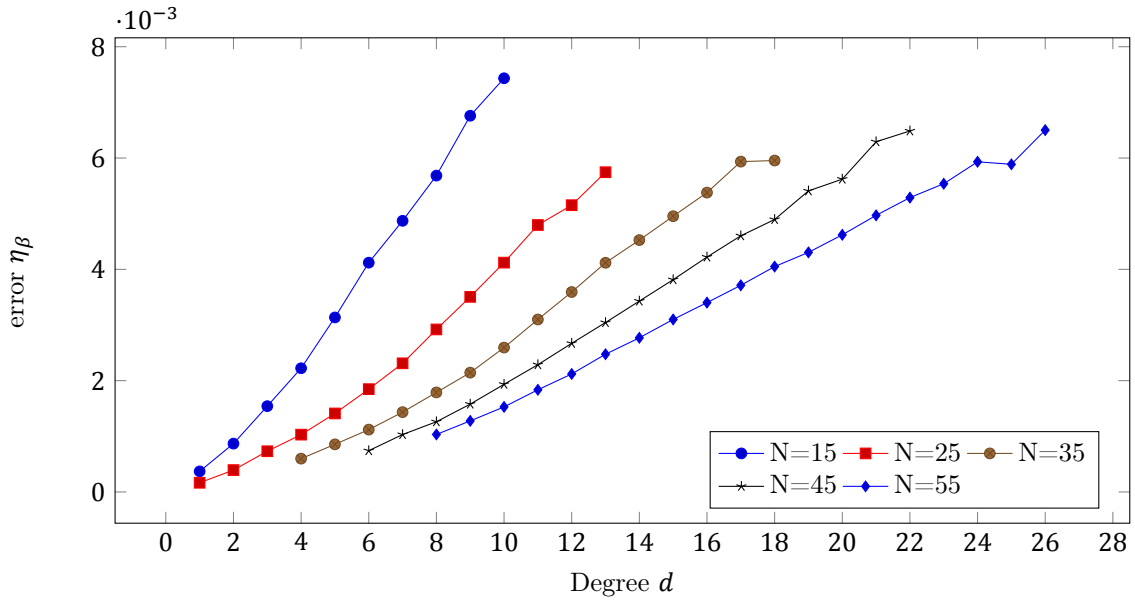
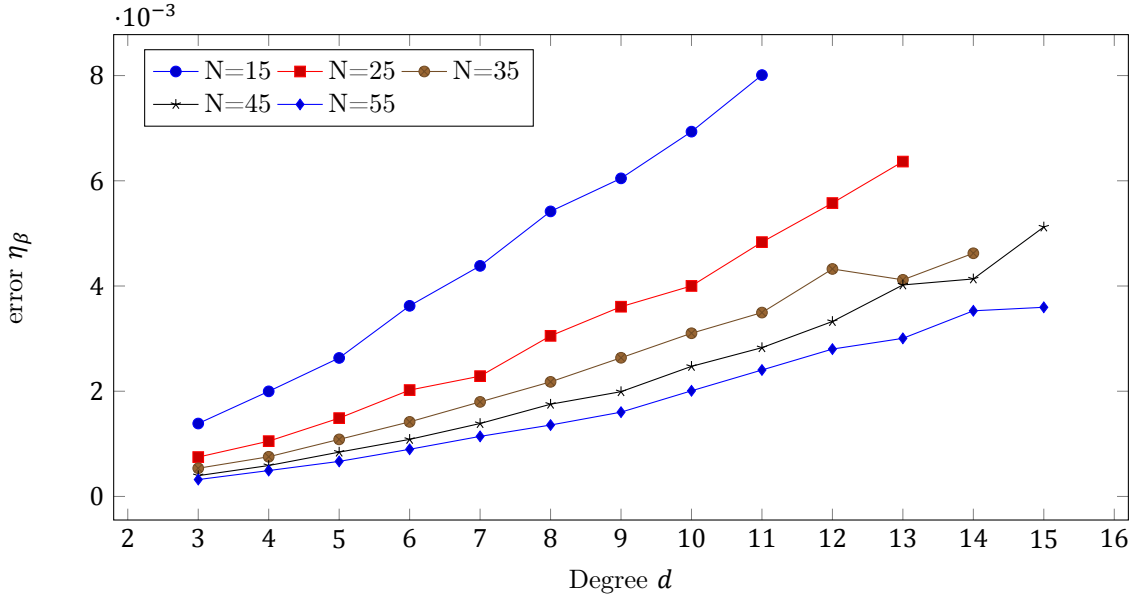


Figure 4.9: Average AUC for a node with degree d, for networks of size $N = 55$.

We can show that the accuracy in the infection probability of a link β_{ij} improves for a decreasing value of d_i. Figures 4.10a and 4.10b show the absolute error β_{ij} can decrease by more than an order of magnitude when comparing high degree and low degree nodes.



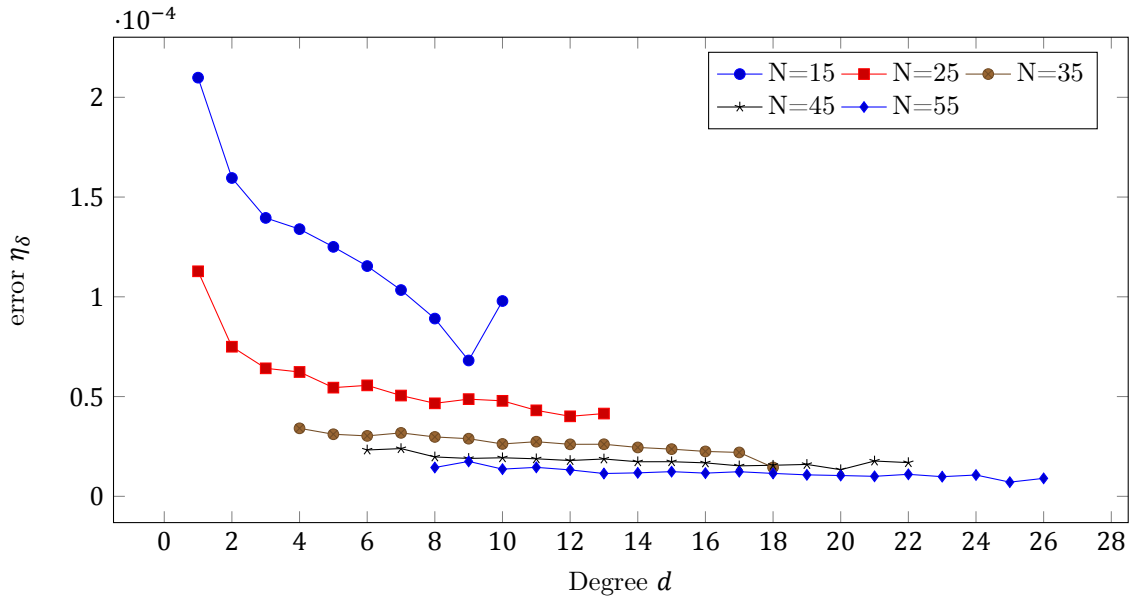
(a) Degree versus link error Erdős-Rényi directed graphs.



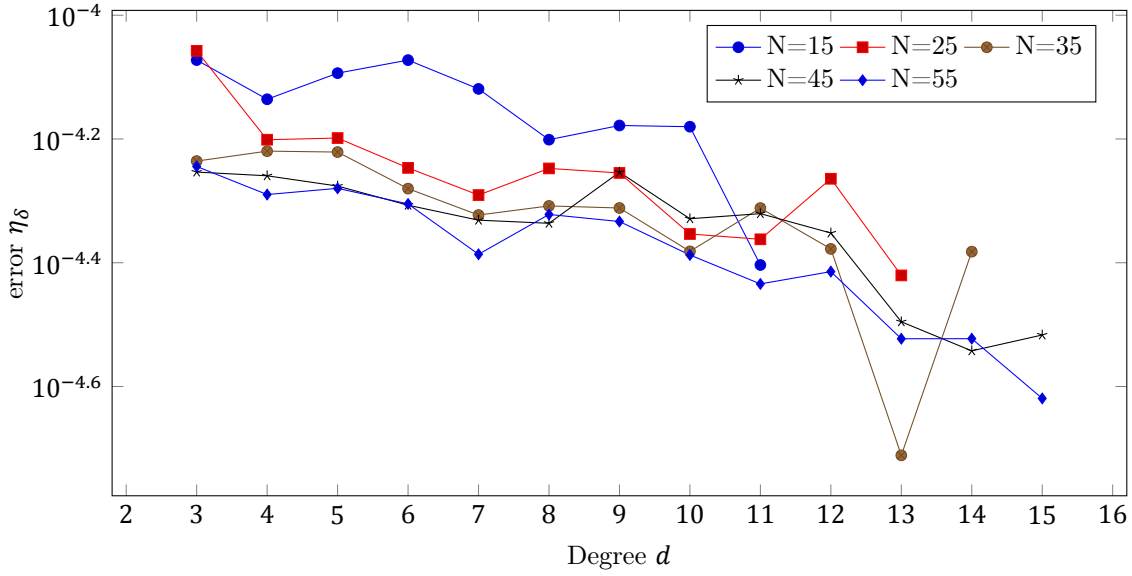
(b) Degree versus link error Barabási-Albert graphs.

Figure 4.10: Average error in infection probability for incident links with a node of degree d.

For the curing probability, Figures 4.11a and 4.11b the variation in the accuracy of an estimate of δ_i is not clearly defined other than a slight increase as the degree of a node d_i increases for $N = 15$.



(a) Degree versus node error Erdős-Rényi directed graphs.

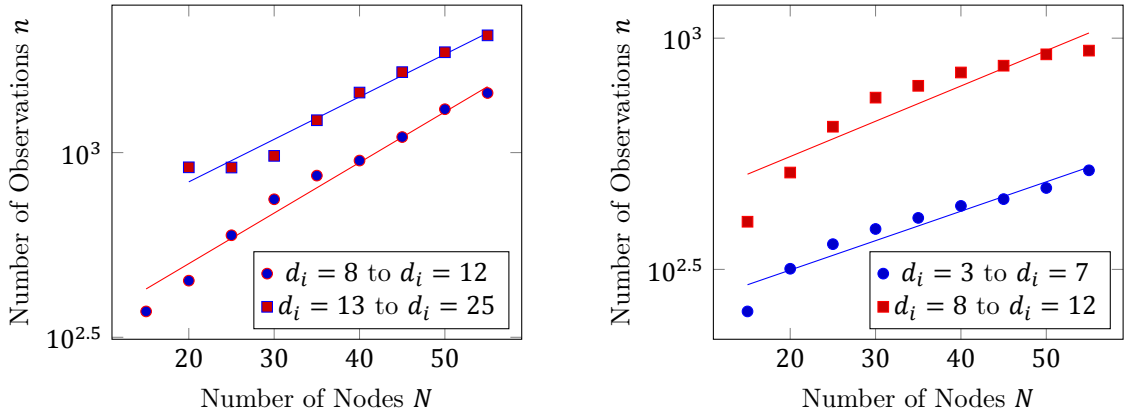


(b) Degree versus node error Barabási-Albert graphs.

Figure 4.11: Average error in curing probability for a node with degree d.

4.4.1. Error in adjacency for nodes with different degrees

The number of observations required for AUC = 0.9 is lower for nodes with a lower degree compared to nodes with a higher degree, for any network size in the range $N = 15$ to $N = 55$ as shown in Figures 4.12a and 4.12b. Table 4.5 shows that we predict that for a network size of $N = 500$ we would require more than 10^5 observations. We calculate the error in A_i, the adjacency of node i by calculating the AUC value of the estimate.



(a) Interpolation of the error in A_i for a node in Erdős-Rényi directed graphs.

(b) Interpolation of the error in A_i for a node in Barabási-Albert graphs.

Figure 4.12: Average error in the AUC for a node with adjacency A_i. This suggests that nodes with a lower degree can more easily be reconstructed than nodes with a higher degree.

Degree d_i	C	g	n(N=500)
8-12	266	0.014	$1.87 \cdot 10^9$
13-25	490	0.012	$2.82 \cdot 10^8$

(a) Reconstruction of Erdős-Rényi directed graphs.

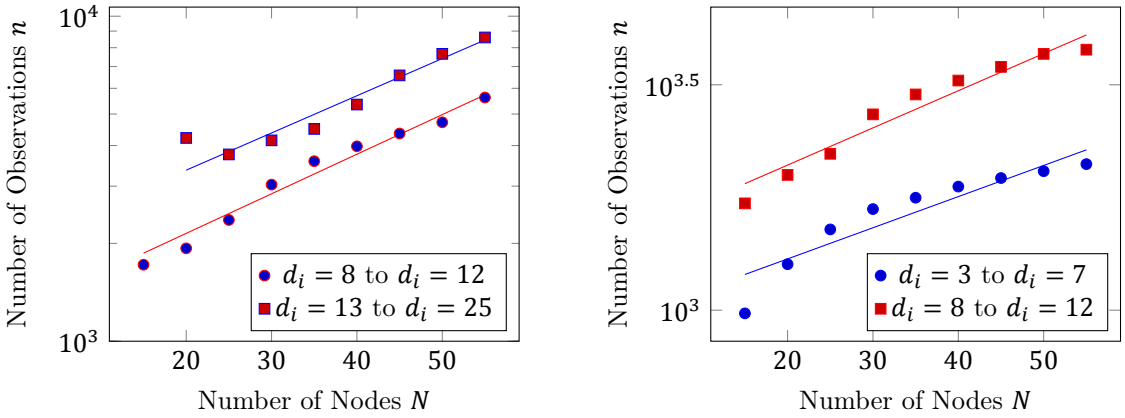
Degree d_i	C	g	n(N=500)
3-7	235	0.006	$3.55 \cdot 10^5$
8-12	390	0.008	$3.56 \cdot 10^6$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.5: Table for fitting (4.21) to error in estimates of the adjacency of nodes in different networks, to get C and g values for different errors. The number of required observations is extrapolated to $N = 500$.

4.4.2. Error in the link infection probability

Figures 4.13a and 4.13b suggest that links incident to nodes with a lower degree require fewer observations to be estimated with an error of $\eta_\beta = 0.2$. This means that it will likely be easier to reconstruct nodes in the network with a low degree. Table 4.6 suggests that over 10^6 observations are required to reconstruct links in the network attached to nodes with a degree in the range $d_i = 3$ to $d_i = 7$.



(a) Interpolation of the error in the link infection propability β_{ij} for mid and high degree nodes in Erdős-Rényi directed graphs.

(b) Interpolation of the error in the link infection propability β_{ij} for low and mid degree nodes in Barabási-Albert graphs.

Figure 4.13: Average error in the infection probability β_{ij} for a link with for different ranges of degrees. This suggests the infection probability of links with a low degree can be estimated more accurately than links with a high degree.

Degree d_i	C	g	n(N=500)
8-12	1225	0.012	$1.51 \cdot 10^9$
13-25	1983	0.011	$1.05 \cdot 10^9$

(a) Reconstruction of Erdős-Rényi directed graphs.

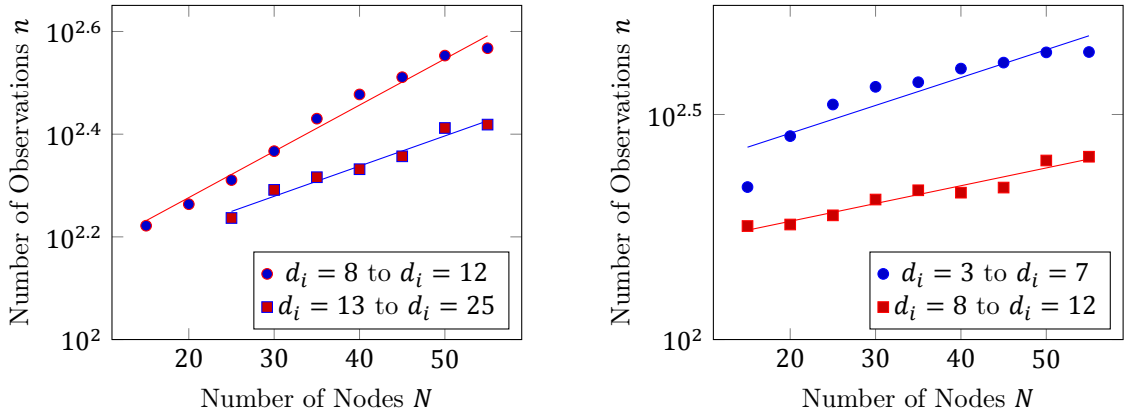
Degree d_i	C	g	n(N=500)
3-7	945	0.007	$2.70 \cdot 10^6$
8-12	1435	0.001	$1.91 \cdot 10^7$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.6: Table for fitting (4.21) to error in estimates of the link infection probability, η_β attached to a node i with degree d_i. We obtain C and g values for different ranges of degree. The number of required observations is extrapolated to $N = 500$.

4.4.3. Error in curing probability

Figure 4.14a shows that nodes degrees in the mid range required more observations for an error of $\eta_\delta = 0.05$ compared to nodes with degrees in th high range. Figure 4.14b demonstrates this is also the case for nodes in the low degree range compared to the high degree range. Table 4.7 shows that $n > 10^4$ observations are required in the best case to reconstruct particular nodes in graphs of $N = 500$.



(a) Interpolation of the error η_δ for mid and high degree nodes in Erdős-Rényi directed graphs.

(b) Interpolation of the error η_δ for mid and high degree nodes in Barabási-Albert graphs.

Figure 4.14: Average error in the curing probability η_δ for a node with degree d_i. This suggests the curing probability of nodes with a high degree can be estimated more accurately than nodes with a low degree.

Degree d_i	C	g	n(N=500)
8-12	125	0.009	$3.96 \cdot 10^6$
13-25	126	0.006	$1.10 \cdot 10^5$

(a) Reconstruction of Erdős-Rényi directed graphs.

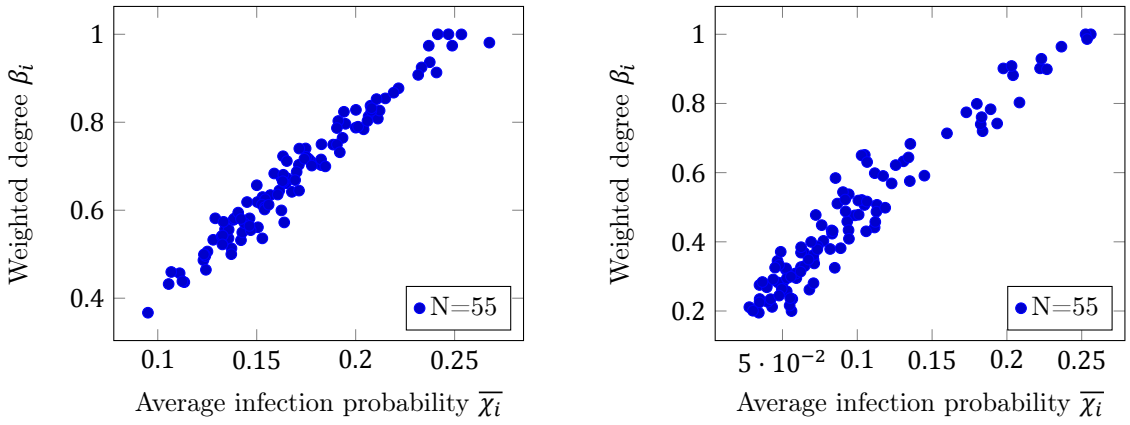
Degree d_i	C	g	n(N=500)
3-7	216	0.006	$2.69 \cdot 10^5$
8-12	153	0.004	$1.44 \cdot 10^4$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.7: Table for fitting (4.21) to error in estimates of the curing probability of nodes in different networks, to get C and g values for different ranges of degrees. The number of required observations is extrapolated to $N = 500$.

4.5. Estimating nodal infection probability

We propose an alternative method to evaluating individual link probabilities where the *weighted degree* of a node is used to estimate its infection probability. It is shown empirically in Figure 4.15 that the weighted degree β_i is highly correlated with the infection probability averaged over time, $\overline{\chi}$, where $\beta_i = \sum_{j=1}^{N} \beta_{ij}$ and $\chi_i[k] = \sum_{j \neq i} \beta_{ij} x_j[k]$ for node i. We can show the error in β_i scales less strongly with the network size N.

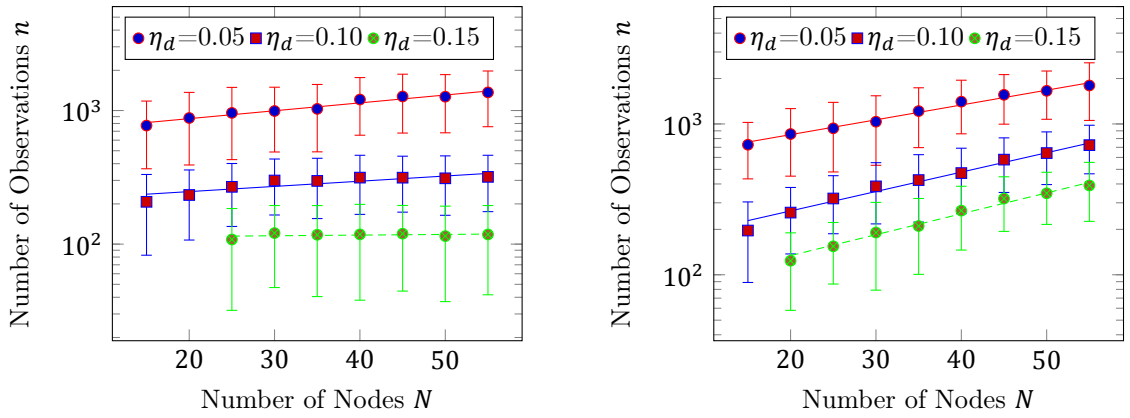


(a) Average $\chi_i[k]$ against β_i for the Erdős-Rényi directed model. (b) Average $\chi_i[k]$ against β_i for the BA model.

Figure 4.15: Correlation between average infection probability and weighted degree

In addition, the number of observations required for reconstruction is much less for β_i than β_{ij}, as demonstrated in Figure 4.16 and Table 4.8 for a fixed level of accuracy determined as

$$\eta_d = \frac{1}{N} \sum_{i=1}^{N} |\beta_i - \hat{\beta}_i| \tag{4.22}$$



(a) Interpolation of the error in β_i for the Erdős-Rényi directed model. (b) Interpolation of error in β_i for the BA model.

Figure 4.16: Average error in total infection probability for a node, for a constant error. This shows there is a weak dependence on the network size for evaluating the weighted degree of a node.

Error β_i	C	g	n(N=500)
0.05	663	0.006	$5.89 \cdot 10^5$
0.1	207	0.004	$1.82 \cdot 10^4$
0.15	112	0.0001	198

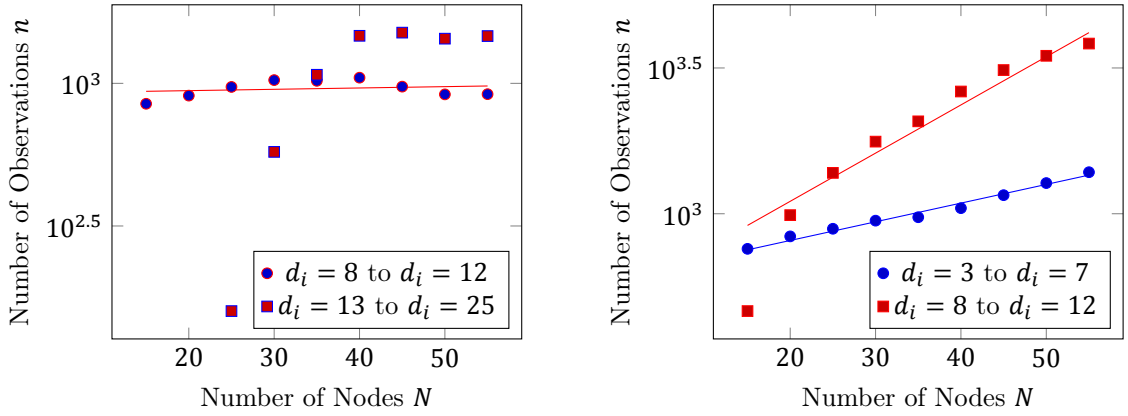
(a) Reconstruction of Erdős-Rényi directed graphs.

Error β_i	C	g	n(N=500)
0.05	540	0.010	$4.43 \cdot 10^7$
0.1	147	0.013	$3.94 \cdot 10^8$
0.15	71	0.014	$5.97 \cdot 10^8$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.8: Table for fitting (4.21) to error in β_i to get A and g values for different errors. The number of required observations is extrapolated to $N = 500$.

We can analyse the accuracy of estimates of the weighted degree β_i for nodes with low, mid or high degrees. For the Erdős-Rényi directed graphs, Figure 4.17a the number of observations did not increase exponentially as expected and levelled off suggesting that for high degrees, there was a convergence in the number of observations required. Figure 4.17b shows that the low degree nodes for Barabási-Albert graphs required fewer observations than the nodes with a mid-range degree, when considering an error in β_i of 0.05. Table 4.9 suggests for nodes with a low degree, the number of observations required for a network size of $N = 500$ would be around 10^6.



(a) Interpolation of the error in the total infection probability of a node β_i for mid and high degree nodes in Erdős-Rényi directed graphs.

(b) Interpolation of the error in the wighted sum of links for a node β_i for low and mid degree nodes in Barabási-Albert graphs.

Figure 4.17: Average error in the weighted degree β_i for a node with adjacency A_i. This suggests the weighted degree of nodes with a high degree can be estimated more accurately than nodes with a low degree.

Degree d_i	C	g	n(N=500)
8-12	922	0.0005	$1.58 \cdot 10^3$
13-25	-	-	-

(a) Reconstruction of Erdős-Rényi directed graphs.

Degree d_i	C	g	n(N=500)
3-7	602	0.006	$9.91 \cdot 10^5$
8-12	552	0.016	$4.75 \cdot 10^{10}$

(b) Reconstruction of Barabási-Albert graphs.

Table 4.9: Table for fitting (4.21) to error in estimates of the weighted sum of links connected to node i in networks of different sizes. C and g values are given for different ranges of degrees. The number of required observations is extrapolated to $N = 500$.

5

Network Reconstruction for the SI and SIR Epidemic Models

Conceptually, the network reconstruction method in Chapter 4 can be generalised to other models than the SIS epidemic process. Here, we consider the susceptible-infected (SI) process where a node cannot be cured once it is infected and susceptible-infected-removed (SIR) process where a node can be cured however is placed in the removed (R) compartment where it is no longer infectious. We demonstrate that maximum likelihood estimation allows for estimates of the adjacency matrix to be obtained for both processes.

We consider a similar simulation setting as for the SIS process with 900 runs for five network sizes in the range $N = 15$ to $N = 55$ and $N = 10$ to $N = 30$ for the SI and SIR models respectively. Each Erdős-Rényi directed and Barabási-Albert undirected graph is generated randomly for each run.

5.1. SI process

The SI process occurs in many real-world situations such as a epidemic of an infectious disease for which there is no known cure such as AIDS [8] where a person is likely to remain infective and alive for a significant period of time after infection.

In order to use ML estimation, we require an objective function that no longer includes the curing probability as for the SI process $\delta = 0$, additionally the constraint for the curing probability is no longer required. The optimisation problem (4.6) therefore reduces to

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmin}} \sum_{k=0}^{n-1} \sum_{i=1}^{N} -\log\left(\Pr\left[x_i[k+1] \middle| x[k], \theta\right]\right) \tag{5.1}$$

with the constraints

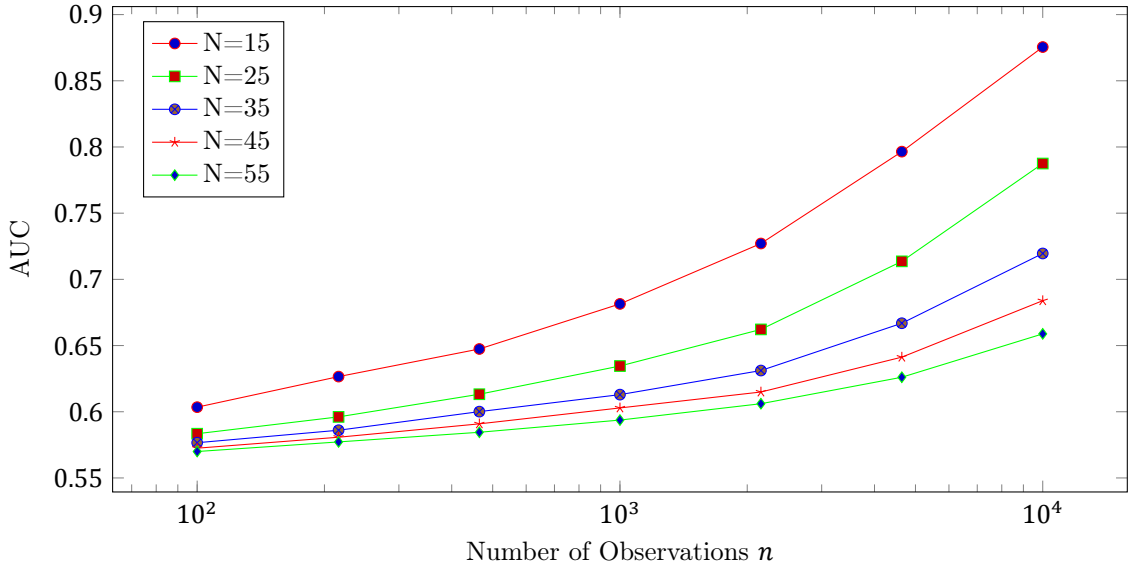
$$0 \leq \beta_{ij} \leq \beta_{\max} \; \forall \; i \neq j \tag{5.2}$$

The objective function for each node i is given as

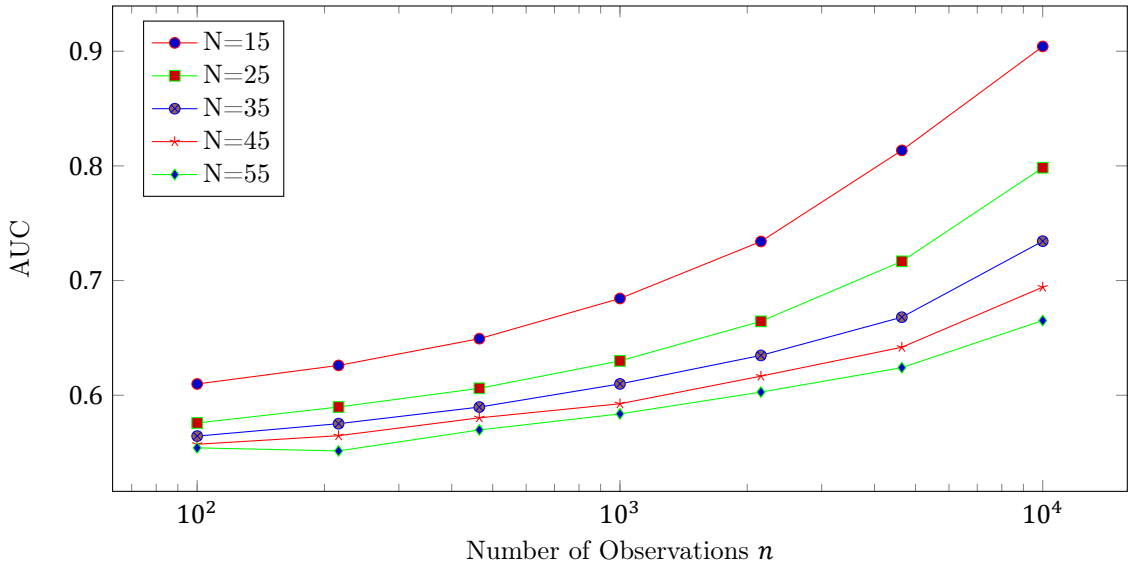
$$\Pr\left[x_i[k+1] \middle| x[k], \theta\right] = (1 - x_i[k])\left((1 - x_i[k+1]) + (2x_i[k+1] - 1)\left(\sum_{j=1}^{N} x_j[k]\beta_{ij}\right)\right) \tag{5.3}$$

To generate the observations we set a single node as infected and the rest as healthy. As nodes can only transition from healthy to infected, the state where all nodes are infected is an absorbing state in which case we reset the next state as having only a single random node as infected and the rest as healthy.

Figure 5.1 shows that we cannot expect a full reconstruction for $n \leq 10^4$, for both Erdős-Rényi directed and Barabási-Albert undirected graphs for network sizes of $N \geq 15$.



(a) AUC versus the number of observations n for a directed Erdős-Rényi graph.



(b) AUC versus the number of observations n for a Barabási-Albert graph.

Figure 5.1: Network reconstruction accuracy of estimates from observations of the SI process as a function of the number of observations n, for various network sizes, on a semi-logarithmic scale.

We can see that it is possible to fit the exponential (4.21) to the SI process in Figure 5.2 for a fixed AUC of 0.6 and 0.7. Tables 5.1a and 5.1b and show that it would be infeasible to reconstruct a network size of $N > 500$, for an AUC of 0.6 or greater.

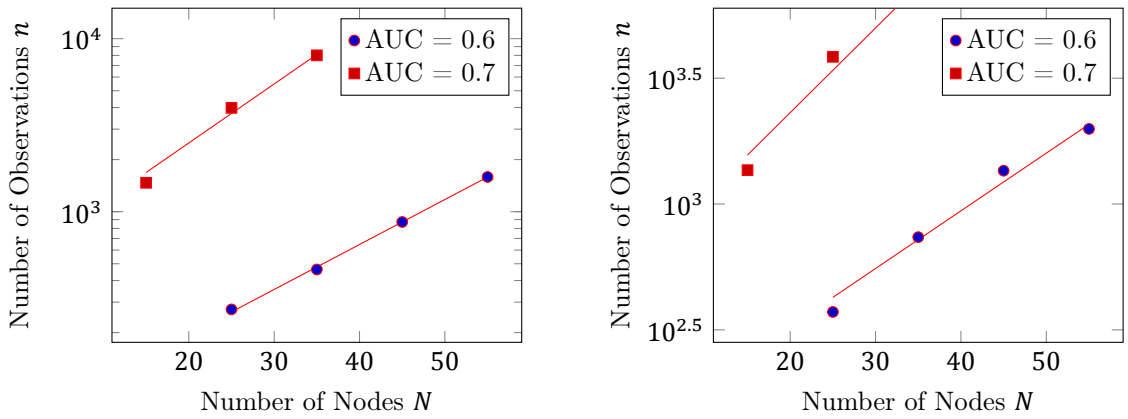
Error A	C	g	n(N=500)
0.6	114	0.023	$3.3 \cdot 10^{13}$
0.7	490	0.034	$3.3 \cdot 10^{19}$

(a) Reconstruction of Erdős-Rényi directed graphs.

Error A	C	g	n(N=500)
0.6	59	0.026	$5.8 \cdot 10^{14}$
0.7	519	0.034	$5.8 \cdot 10^{19}$

(b) Reconstruction of Barabási-Albert graphs.

Table 5.1: Table for fitting (4.21) to the calculated AUC to obtain C and g values for different errors. The number of required observations is extrapolated to $N = 500$.



(a) Interpolation of AUC for the Erdős-Rényi directed graphs. (b) Interpolation of AUC for Barabási-Albert graphs.

Figure 5.2: The number of observations n as a function of network size N for the SI process. Fitting an exponential curve suggests that the increase number of observations required increases exponentially as the number of nodes increases.

5.2. SIR process

The SIR process can be used to model diseases where there is a high mortality such as the plague [21]. The SIR model includes three compartments. The new removed compartment R is used to represent nodes that no longer participate in the spread of the disease, which in the case of a disease would represent deceased or immune individuals. Infection moves an individual from S to I, as is the case for the SI and SIS models. Curing moves an individual from I to R, where they can neither be infected by nor infect other nodes. In this model, the absorbing state is where there are no infected individuals, as a stable state is reached when no further infections or curing can occur. We consider several cascades where the initial state consists of a single, randomly selected infected node with the rest susceptible. The additional compartment requires an additional state for a node. By $x_i[k] = -1$ we indicate that node i at time k is in the removed state.

For the SIR model, the ML-estimation problem becomes

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmin}} \sum_{k=0}^{n-1} \sum_{i=1}^{N} -\log\left(\Pr\left[x_i[k+1] \middle| x[k], \theta\right]\right) \tag{5.4}$$

with the constraints

$$0 \leq \beta_{ij} \leq \beta_{\max} \;\forall\; i \neq j \tag{5.5}$$

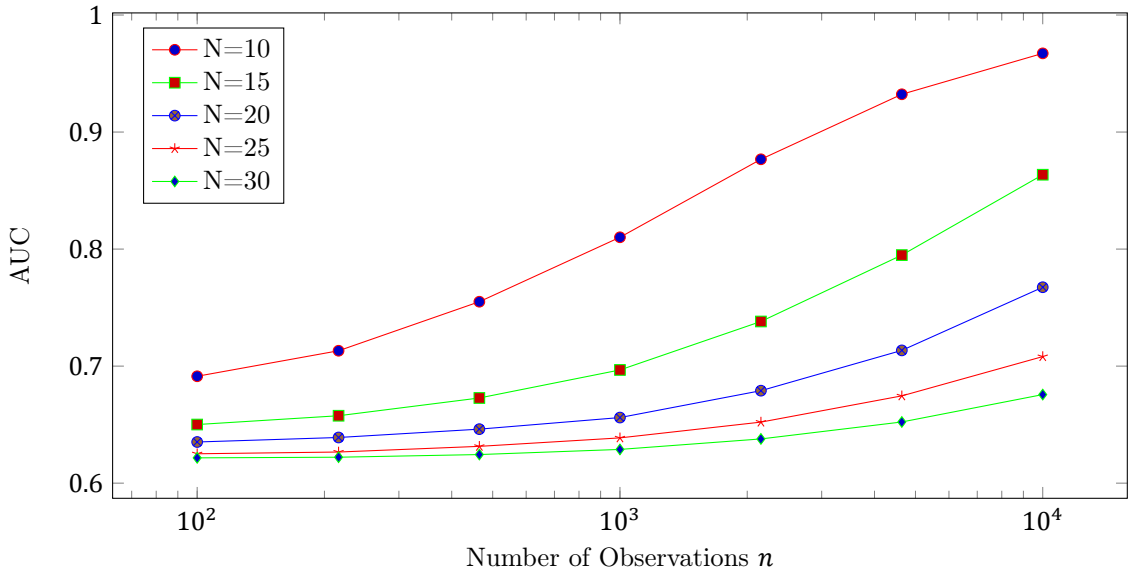
$$0 \leq \delta_i \leq \delta_{\max} \;\forall\; i \tag{5.6}$$

where the addends in the objective function (5.4) are given by

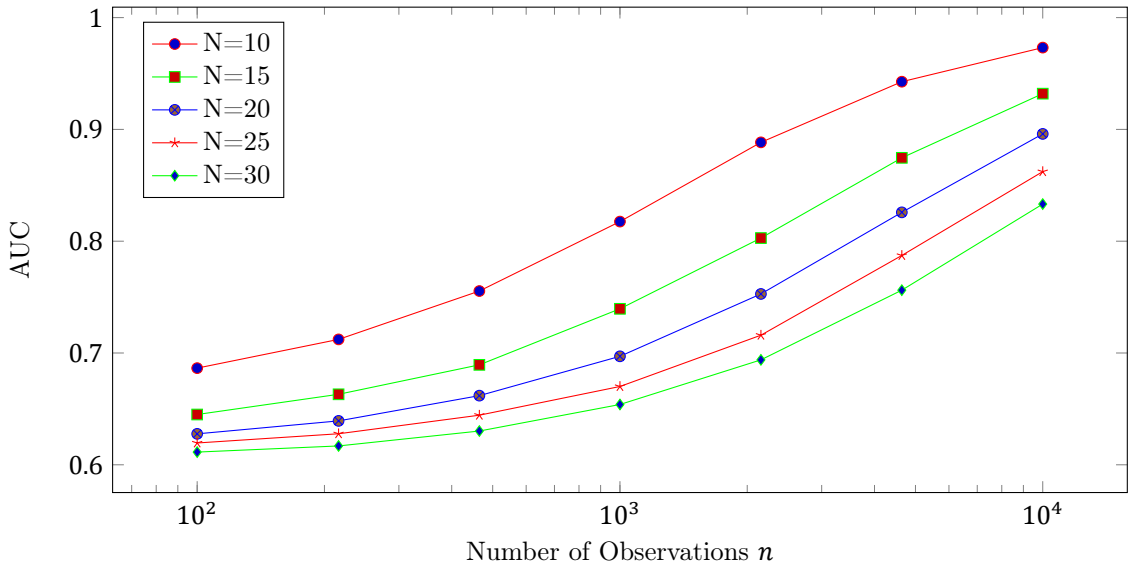
$$\begin{aligned} \Pr\left[x_i[k+1] \middle| x[k], \theta\right] = & \frac{x_i[k](x_i[k]+1)x_i[k+1]^2}{2}\left(\frac{(x_i[k+1]+1)}{2} - x_i[k+1]\delta_i\right) \\ & + (x_i[k]^2 - 1)(x_i[k+1]+1)\left(\frac{x_i[k+1]}{2} - 1\right) \\ & \cdot \left((1 - x_i[k+1]) + (2x_i[k+1]-1)\left(\sum_{j=1}^{N} x_j[k]\beta_{ij}\right)\right) \\ & + \frac{x_i[k+1](x_i[k+1]-1)x_i[k](x_i[k]-1)}{4} \end{aligned} \tag{5.7}$$

From Figure 5.3 we can see that it is possible to reconstruct the network from observations of the SIR process. However as demonstrated in Figure 5.4 we again require an exponentially increasing number

of observations. Tables 5.2a and 5.2b show how infeasible this becomes for a network size of $N = 500$ with over $n = 10^{20}$ observations required for an accurate reconstruction.

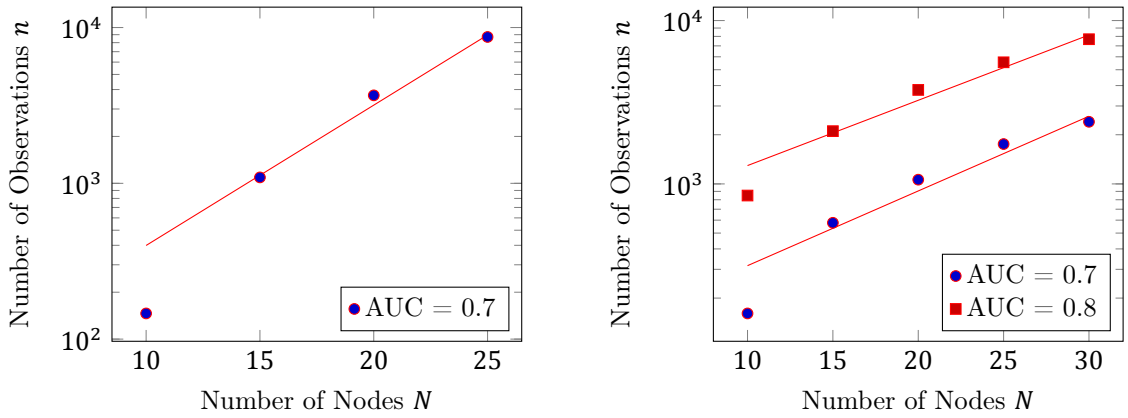


(a) AUC versus the number of observations n for a directed Erdős-Rényi graph.



(b) AUC versus the number of observations n for a Barabási-Albert graph.

Figure 5.3: Network reconstruction accuracy of estimates from observations of the SIR process as a function of the number of observations n, for various network sizes, on a semi-logarithmic scale.



(a) Interpolation of AUC for the Erdős-Rényi directed graphs. (b) Interpolation of AUC for Barabási-Albert graphs.

Figure 5.4: The number of observations n as a function of network size N for the SIR process. Fitting an exponential curve suggests that the increase number of observations required increases exponentially as the number of nodes increases.

Error A	C	g	n(N=500)
0.7	50.2	0.090	$5.4 \cdot 10^{46}$

(a) Reconstruction of Erdős-Rényi directed graphs.

Error A	C	g	n(N=500)
0.7	110	0.056	$8.1 \cdot 10^{24}$
0.8	516	0.040	$5.0 \cdot 10^{22}$

(b) Reconstruction of Barabási-Albert graphs.

Table 5.2: Table for fitting (4.21) to the calculated AUC to obtain C and g values for different errors. The number of required observations is extrapolated to $N = 500$.

6

Conclusion

A large part of the academic and data science community believe that the scale of Big Data will help to resolve the issues of network reconstruction. The idea being, that by tracking the propagation of information, such as posts, diseases or trends, the underlying properties of the network such as the adjacency, curing and infection rates can be obtained. This can be seen as part of the Big Data belief that there is enough data, in this case observations to allow for a reasonable estimate to be made of a complex system. The possibilities are vast if estimation were possible and would help in a wide array of issues from monitoring a group of individuals to uncover criminal gangs to helping remove the connections that allow diseases to propagate through a network.

We demonstrate that it is hardly possible in practice to completely reconstruct the infection probabilities, curing probabilities and the adjacency matrix A for the whole network for a given level of accuracy in a reasonable amount of time for a large number of nodes, that is greater than $N \geq 500$. We solve the maximum-likelihood estimation meaning that a better estimation does not exist, unless there is prior knowledge. Furthermore, we emphasise that we considered *idealised conditions* where all observations of all nodes are available and transitions are Markovian meaning reinfections for nodes are independent of previous infections for that node. We also consider the simplified setting of a static network where the adjacency of the nodes does not change over time, as it might in a real world network.

For larger networks where $N > 500$ one may wonder if it is possible to reconstruct parts of the network. However we observe a similar exponential increase in the number of required observations rendering network reconstruction infeasible, e.g. for social networks with millions of individuals. Estimates of infection probabilities and links for low-degree nodes with a degree between $d_i = 3$ and $d_i = 7$ and curing probabilities for high-degree nodes with a degree $d_i > 13$ are only possible if the number of observations grows as $n(N) \approx C10^{gn}$ where C is in the range $100 < C < 2000$ and $g < 0.01$.

Our results hold not only for the SIS epidemics, but also for SI and SIR epidemics. Hence, we believe that the reconstruction of large networks is infeasible for a general class of compartmental epidemic models, see for instance [38, 36].

6.1. Outlook

We show that the accuracy of an estimate of the curing and infection probabilities varies based on the degree of a nodes. To extend the analysis of partially reconstructed networks, a more graph-theoretic approach could establish if this relates to other graphs metrics such the closeness of a node and if this has an effect on the accuracy of an estimate. There could be further indicators as to whether an estimate of a node or link in a network is likely to be accurate or not.

Further research could analyse a network that varies over time to better simulate real-world conditions. The average degree over time could be considered to as this could have implications for the accuracy of an estimate as is the case for the degree in a static network.

A

Objective function derivation

Four different begin and end states combinations exist for a node i when a transition occurs from the viral state at time k, $x_i[k]$ to the viral state at time $k+1$, $x_i[k+1]$, resulting in the probabilities given by (A.2) to (A.5). These can be conditioned on the viral state at time k for the whole network $x[k]$ and the parameters θ.

For simplicity, we define

$$\sum_{j=1}^{n} x_j[k]\beta_{ij} = \chi_i[k] \tag{A.1}$$

The probability of infection at time $k+1$ for node i that is healthy at time k is

$$\Pr\left[x_i[k+1] = 1 | x_i[k] = 0, x[k], \theta\right] = \sum_{j=1}^{N} x_j[k]\beta_{ij} = \chi_i[k] \tag{A.2}$$

The probability of remaining infected at time $k+1$ for node i that is infected at time k is

$$\Pr\left[x_i[k+1] = 0 | x_i[k] = 1, x[k], \theta\right] = 1 - \delta_i \tag{A.3}$$

The probability of curing at time $k+1$ for node i that is infected at time k is

$$\Pr\left[x_i[k+1] = 0 | x_i[k] = 1, x[k], \theta\right] = \delta_i \tag{A.4}$$

The probability of remaining healthy at time $k+1$ for node i that is healthy at time k is

$$\Pr\left[x_i[k+1] = 0 | x_i[k] = 0, x[k], \theta\right] = 1 - \sum_{j=1}^{N} x_j[k]\beta_{ij} = 1 - \chi_i[k] \tag{A.5}$$

Not all of these transitions are possible between times k and $k+1$. The value of $x_i[k]$ and $x_i[k+1]$ can be used to select certain terms in the objective function meaning there is no need to separate the observations into sets. This gives the following equation for the probability of a transition for each node i from time k to time $k+1$

$$\begin{aligned}\Pr\left[x_i[k+1] \middle| x[k], \theta\right] =& x_i[k]\left(x_i[k+1] + (1 - 2x_i[k+1])\delta_i\right) \\ &+ (1 - x_i[k])\left((1 - x_i[k+1]) + (2x_i[k+1] - 1)\left(\sum_{j=1}^{N} x_i[k]\beta_{ij}\right)\right)\end{aligned} \tag{A.6}$$

B

Algorithms

We first generate the parameters for a given network using Algorithm 1. We then generate observations of the SIS process using the parameters θ and the network A using Algorithms 2 and 3.

In Algorithm 1 we need to firstly generate the adjacency matrix A with elements a_{ij}. The graph type is set as either a directed Erdős-Rényi graph in which case the graph parameters includes only the link probability p_{ER} or Barabási-Albert undirected graph where the graph parameters include the size of the initial complete graph $m0$ and the number of nodes that any further nodes will have to connect m.

We set the maximum curing rate for a node as $\tilde{\delta}_{\max} = N$ and the maximum infection rate for a link as $\tilde{\beta}_{\max} = 1$. This enables the curing probability δ_i and infection probability $\beta_i = \sum_{j=1}^{N} \beta_{ij}$ for a node to be a maximum of 1 for a sample time of $1/N$.

Algorithm 1 Generate Parameters

Input: Maximum curing rate $\tilde{\delta}_{\max}$, maximum infection rate $\tilde{\beta}_{\max}$, basic reproduction rate R_0, network size N, graph type, graph parameters
Output: curing probability vector δ, infection probability matrix B
A = generate adjacency matrix(N, graph type, graph parameters)
for $i = 1, \ldots, N$ do
    $\tilde{\delta}_i$ = generate random number in range $[0, \tilde{\delta}_{\max}]$
    for $j = 1, \ldots, N$ do
        $\tilde{\beta}_{ij}$ = generate random number in range $[0, a_{ij}\tilde{\beta}_{\max}]$
for $i = 1, \ldots, N$ do
    for $j = 1, \ldots, N$ do
        The effective spreading parameter for a link from node j to node i, $w_{ij} = \tilde{\beta}_{ij}/\tilde{\delta}_i$
$R_{0,curr}$ = largest eigenvalue of the effective spreading parameter matrix W
$\tilde{\delta} = \frac{R_{0,curr}}{R_0} \cdot \tilde{\delta}$
$T = \frac{1}{N}$
$\delta = \tilde{\delta} \cdot T$
$B = \tilde{B} \cdot T$

In Algorithm 2 the viral state x[k] of all nodes at time k is used to calculated the next viral state sequence $x[k+1]$. If all node are healthy for $x[k]$ the process is in an absorbing state so we set all the nodes to be infected.

Algorithm 2 Generate Observations

Input: curing probability vector δ, infection probability matrix B, number of observations n
Output: Viral state sequence $X[n]$
$x[0] =$ all-one vector u
for $k = 0, ..., n-1$ do
    if $x[0] = 0$ then
        $x[k+1] =$ u
    else
        $x[k+1] =$ calculate the next viral state $(x[k], \delta, B)$

In Algorithm 3 we generate a viral state at time $k+1$, $x[k+1]$ by using the curing probability δ_i for a node i if it is infected or the probability of infection $\chi_i[k] = \sum_{j=1}^{N} \left(\beta_{ij} x_i[k]\right)$ if it is healthy.

Algorithm 3 Calculate The Next Viral State

Input: curing probability vector δ, infection probability matrix B, current viral state $x[k]$
Output: next viral state $x[k+1]$
for $i = 1, .., N$ do
    $\chi_i[k] = \sum_{j=1}^{N} \left(\beta_{ij} x_i[k]\right)$
    if $x_i[k] == 1$ then
        if $\delta_i >$ random number in range $[0, 1]$ then
            $x_i[k+1] = 1$
        else
            $x_i[k+1] = 0$
    if $x_i[k] == 0$ then
        if $\chi_i >$ random number in range $[0, 1]$ then
            $x_i[k+1] = 0$
        else
            $x_i[k+1] = 1$

Bibliography

[1] S. Abdullah and X. Wu. "An Epidemic Model for News Spreading on Twitter". In: 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence. 2011, pp. 163–169.

[2] L. J. S. Allen and A. M. Burgin. "Comparison of deterministic and stochastic SIS and SIR models in discrete time". In: Mathematical Biosciences 163.1 (2000), pp. 1–33. ISSN: 0025-5564.

[3] S. K. Amirkolaei and S. Jalil. "Problems in Epidemic Inference on Complex Networks". PhD thesis. UCLA, 2019.

[4] A. Barabasi and R. Albert. "Emergence of Scaling in Random Networks". In: Science (New York, N.Y.) 286 (Nov. 1999), pp. 509–12.

[5] D. Bernoulli. "Essai d'une nouvelle analyse de la mortalite causee par la petite verole, et des avantages de l'inoculation pour la prevenir". In: Histoire de l'Acad., Roy. Sci. (Paris) avec Mem, (1760), pp. 1–45.

[6] D. Bucur. "Top influencers can be identified universally by combining classical centralities". In: Scientific Reports 10 (2020), p. 20550.

[7] R. H. Byrd, M. E. Hribar, and J. Nocedal. "An Interior Point Algorithm for Large-Scale Nonlinear Programming". In: SIAM Journal on Optimization 9.4 (1999), pp. 877–900.

[8] F. H. Chen. "A Susceptible-infected Epidemic Model with Voluntary Vaccinations". In: Journal of Mathematical Biology 53.2 (2006), pp. 253–272. ISSN: 1432-1416.

[9] M. Cinelli et al. "The limited reach of fake news on Twitter during 2019 European elections". In: PLOS ONE 15.6 (June 2020), pp. 1–13.

[10] G. Ciriello and C. Guerra. "A review on models and algorithms for motif discovery in protein-protein interaction networks". In: Briefings in Functional Genomics 7.2 (Apr. 2008), pp. 147–156. ISSN: 2041-2649.

[11] J. Clement. Number of daily active Facebook users worldwide as of 3rd quarter 2020. 2020. URL: `https://www.statista.com/statistics/346167/facebook-global-dau/[accessed:04/11/20]`.

[12] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. "On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations". In: Journal of Mathematical Biology 28.4 (1990), pp. 365–382. ISSN: 1432-1416.

[13] P. Erdös and A. Rényi. "On Random Graphs I". In: Publicationes Mathematicae Debrecen 6 (1959), p. 290.

[14] L. Euler. "Solutio problematis ad geometriam situs pertinentis". In: Commentarii Academiae Scientiarum Imperialis Petropolitanae 8 (1736), pp. 128–140.

[15] E. N. Gilbert. "Random Graphs". In: Ann. Math. Statist. 30.4 (Dec. 1959), pp. 1141–1144.

[16] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. "Inferring Networks of Diffusion and Influence". In: ACM Trans. Knowl. Discov. Data 5.4 (Feb. 2012), p. 30. ISSN: 1556-4681.

[17] Department of Health and Social Care. NHS Test and Trace: how it works. 2020. URL: `https://www.gov.uk/guidance/nhs-test-and-trace-how-it-works`.

[18] J. A. P. Heesterbeek. "A brief history of R 0 and a recipe for its calculation". In: Acta biotheoretica 50.3 (2002), pp. 189–204.

[19] H. W. Hethcote. "The Mathematics of Infectious Diseases". In: SIAM Review 42.4 (2000), pp. 599–653.

[20] F. S. Hillier and G. J. Lieberman. Introduction to Operations Research. Seventh. McGraw-Hill Science/Engineering/Math, 2001, pp. 1165–1166. ISBN: 0073211141.

[21] W. O. Kermack, A. G. McKendrick, and G. T. Walker. “A contribution to the mathematical theory of epidemics”. In: Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 115.772 (1927), pp. 700–721.

[22] I. Z. Kiss, J. C. Miller, and P. L. Simon. Mathematics of Epidemics on Networks: From Exact to Approximate Models. Interdisciplinary Applied Mathematics. Springer International Publishing, 2017. ISBN: 9783319508061.

[23] F. Liljeros et al. “The web of human sexual contacts”. In: Nature 411.6840 (2001), pp. 907–908. ISSN: 1476-4687.

[24] G. Marti et al. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. 2020.

[25] Mathworks. fmincon. 2021. URL: `https://uk.mathworks.com/help/optim/ug/fmincon.html`.

[26] S. Milgram. “The Small-World Problem”. In: Psychology Today 1.1 (1967).

[27] S. Myers and J. Leskovec. “On the Convexity of Latent Social Network Inference”. In: Advances in Neural Information Processing Systems. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc., 2010, pp. 1741–1749.

[28] C. Nowzari, V. Preciado, and G. Pappas. “Analysis and Control of Epidemics: A Survey of Spreading Processes on Complex Networks”. In: IEEE Control Systems 36 (May 2015).

[29] J. Oh et al. “National Response to COVID-19 in the Republic of Korea and Lessons Learned for Other Countries”. In: Health Systems & Reform 6.1 (2020). PMID: 32347772, e1753464.

[30] R. Pastor-Satorras and A. Vespignani. “Epidemic Spreading in Scale-Free Networks”. In: Phys. Rev. Lett. 86 (14 2001), pp. 3200–3203.

[31] R. Pastor-Satorras et al. “Epidemic processes in complex networks”. In: Reviews of Modern Physics 87.3 (2015), pp. 925979. ISSN: 1539-0756.

[32] T. P. Peixoto. “Network Reconstruction and Community Detection from Dynamics”. In: Phys. Rev. Lett. 123 (12 2019), p. 128301.

[33] S.T. Peters. Smallpox in the New World. Epidemic! Series. Benchmark Books, 2005, pp. 13–15. ISBN: 9780761416371.

[34] B. Prasse and P. Van Mieghem. “Exact Network Reconstruction from Complete SIS Nodal State Infection Information Seems Infeasible”. In: IEEE Transactions on Network Science and Engineering 6.4 (2019), pp. 748–759.

[35] B. Prasse and P. Van Mieghem. “Maximum-Likelihood Network Reconstruction for SIS Processes is NP-Hard”. In: ArXiv abs/1807.08630 (2018).

[36] B. Prasse and P. Van Mieghem. “Network Reconstruction and Prediction of Epidemic Outbreaks for General Group-Based Compartmental Epidemic Models”. In: IEEE Transactions on Network Science and Engineering 7.4 (2020), pp. 2755–2764.

[37] Rijksoverheid. Coronamelder app. 2020. URL: `https://coronamelder.nl/[accessed: 21/11/20]`.

[38] F. D. Sahneh, C. Scoglio, and P. Van Mieghem. “Generalized Epidemic Mean-Field Model for Spreading Processes Over Multilayer Complex Networks”. In: IEEE/ACM Transactions on Networking 21.5 (2013), pp. 1609–1620.

[39] J. Snow. On the Mode of Communication of Cholera. John Churchill, 1855.

[40] R. A. Stein. “Super-spreaders in infectious diseases”. In: International Journal of Infectious Diseases 15.8 (2011), e510 –e513. ISSN: 1201-9712.

[41] P. Van Mieghem. Performance Analysis of Complex Networks and Systems. Cambridge University Press, 2014. ISBN: 9781107415874.

[42] P. Van Mieghem and R. van de Bovenkamp. “Non-Markovian Infection Spread Dramatically Alters the Susceptible-Infected-Susceptible Epidemic Threshold in Networks”. In: Phys. Rev. Lett. 110 (10 2013), p. 108701.

[43] D. J. Watts and St. H. Strogatz. “Collective dynamics of ’small-world’ networks”. In: Nature 393.6684 (1998), pp. 440–442. ISSN: 0028-0836.