Improving network inference by overcoming statistical limitations

Gloria Cecchini

Bachelor degree in Mathematics at the University of Florence, Italy, 2011

Master degree in Mathematics at the University of Florence, Italy, 2015

A thesis presented for the double Research degree of

Doctor of Philosophy

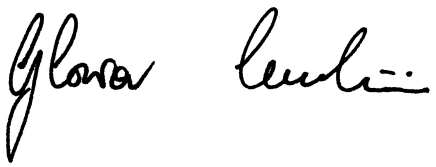
in Physics

at the University of Aberdeen
and at the University of Potsdam

Year of award 2018

Declaration

This thesis is entirely my own composition. It has not been accepted in any previous application for a degree. This thesis has been written as part of a degree undertaken at both Aberdeen and Potsdam Universities. Any personal data have been processed in accordance with the provisions of the Data Protection Act 1998. It is a record of my work and all verbatim extracts have been distinguished by quotation marks. My sources of information have been specifically acknowledged.



Summary

A reliable inference of networks from data is of key interest in many scientific fields. Several methods have been suggested in the literature to reliably determine links in a network. These techniques rely on statistical methods, typically controlling the number of false positive links, but not considering false negative links. In this thesis new methodologies to improve network inference are suggested. Initial analyses demonstrate the impact of false positive and false negative conclusions about the presence or absence of links on the resulting inferred network. Consequently, revealing the importance of making well-considered choices leads to suggest new approaches to enhance existing network reconstruction methods.

A simulation study, presented in Chapter 3, shows that different values to balance false positive and false negative conclusions about links should be used in order to reliably estimate network characteristics. The existence of *type I* and *type II errors* in the reconstructed network, also called biased network, is accepted. Consequently, an analytic method that describes the influence of these two errors on the network structure is explored. As a result of this analysis, an analytic formula of the density of the biased vertex degree distribution is found (Chapter 4).

In the inverse problem, the vertex degree distribution of the true underlying network is analytically reconstructed, assuming the probabilities of *type I* and *type II errors*. Chapters 4-5 show that the method is robust to incorrect estimates of α and β within reasonable limits. In Chapter 6, an iterative procedure to enhance this method is presented in the case of large errors on the estimates of α and β.

The investigations presented so far focus on the influence of false positive

and false negative links on the network characteristics. In Chapter 7, the analysis is reversed - the study focuses on the influence of network characteristics on the probability of *type I* and *type II errors*, in the case of networks of coupled oscillators. The probabilities of α and β are influenced by the shortest path length and the detour degree, respectively. These results have been used to improve the network reconstruction, when the true underlying network is not known a priori, introducing a novel and advanced concept of threshold.

Published online at the
Institutional Repository of the University of Potsdam:
https://doi.org/10.25932/publishup-42670
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-426705

Acknowledgement

I would like to thank my advisor Prof. Bjoern Schelter for his patience and support of my PhD study. His guidance helped me in all the time of research and writing of this thesis.

My sincere thanks goes also to Prof. Arkady Pikovsky for the support of my research and immense knowledge. I could not have imagined having a better mentor for my PhD.

A very special gratitude goes to Dr. Caroline Reid, a key character in this whole project. An amazing organiser, problem solver, and friend who made doing a PhD an easy task.

I would like to express my sincere gratitude to Helen and Suzi, office mates and friends. Thanks not only for helping me with English and bureaucratic stuff, but also because you made me feel welcome and made my stay in Aberdeen wonderful.

A very great appreciation goes out to all down at Research Fund and especially to all the PIs who created this project. This PhD has been an amazing experience both professionally and personally.

Last but not the least, I would like to thank my family that always supported me despite the geographical distance. Also, thanks for finding good excuses to travel and come to visit me.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642563.

Contents

Chapter 1

Introduction

Complex systems are of key interest in multiple scientific fields, ranging from medicine via physics, mathematics, engineering to economics [6, 9, 19]. These systems can be modelled, or represented as networks, where nodes are the elements of the system and links represent the interactions between them. Networks are ubiquitous in many fields of study [28]; examples include the Internet, airline connections, scientific collaborations and citations, trade market contacts, social relationships, cellular and ecological systems, transportation systems, power grids, and the human brain [1, 2, 3, 24, 25, 39, 47, 49, 62, 66, 68].

Brain connectivity analyses are one of the key topics in the Neurosciences. The complexity of the interaction of various regions of the brain necessitates studying the brain as a whole rather than just its individual parts. The application of network theory has facilitated these analyses, leading to a better understanding of the structure and function of the nervous system [60].

In a more theoretical framework in physics, networks are also studied to investigate synchronization phenomena of coupled oscillators as well as the analysis of chaotic behaviour and corresponding phenomena in dynamical systems [16, 38, 41, 54].

Defining the structural properties of networks, and specifically their characteristics, is of fundamental importance to understand the complex dynam-

ics of systems [40]. For instance, in the case of optimizing vaccination strategies, a wide understanding of network characteristics allows controlling the network dynamics, and consequently controlling the spread of diseases [18].

To capture particular features of a network, properties or characteristics have been introduced in the literature [50]. For instance, the node degree and the shortest path length are two of the most used characteristics. The node degree describes the number of links of a node, while the shortest path length is the minimal number of links separating two nodes [48]. The node degree distribution is a key property to study the general structure of the connectivity of a network, and it is predominantly used in this thesis for this reason.

Some complex systems can be directly observed such as power grids, or transportation systems. In these kinds of systems, the network topology is known a priori, and its characteristics can be investigated. Other complex systems cannot be directly observed, therefore network structure is not known a priori, and must be inferred indirectly. When the underlying network is not known a priori, reliably inferring network structure from data is crucial to represent the system accurately; this is known as the inverse problem. The functional network of the human brain, that is the brain network used to solve certain problems or manage certain tasks, is one prototypical example where the network has to be inferred. This can be achieved from observed electroencephalography or functional magnetic resonance imaging data. Understanding the functioning or malfunctioning of the human brain is of key interest, for example to treat brain-related diseases such as epilepsy, Parkinson's disease, or stroke [12, 52, 53].

When a network is to be inferred from data, typical analysis techniques provide a measure of connectivity strength for each link. Statistical methods are then used to decide whether these measures pass a certain threshold, and thereby provide a means to decide if the corresponding links are considered present. If a link is erroneously detected, this is called false positive link and it is referred to as a *type I error*. Likewise, an existing link that remains undetected is called false negative link and it is referred to as a *type II error*. The probability of detecting a false positive link is usually denoted by α,

while β refers to the probability that an existing link remains undetected.

Classical statistical methods aim to reconstruct with high certainty the presence of links, i.e., the analysis has high specificity, and the standard value of 0.05 for α is often chosen [17, 22, 23, 33, 35, 57, 59]. The decision to set $\alpha = 0.05$ does not take into account the probability of false negative links, since, usually, high specificity implies low sensitivity, meaning a high chance of missing links. Intuitively, there is an inverse relationship between the probabilities of *type I* and *type II errors*; hence, it is typically impossible to have both α and β equal to zero.

When the aim is to reconstruct with high certainty the presence of a given link, a low value for α must be used, and a high probability for false negative links must consequently be accepted. When instead the aim is to recover the general structure of the network, and reliably infer certain network characteristics, for instance the shortest path length, or node degree, balancing α and β is necessary.

This thesis focuses on the investigation of false positive and false negative links in inferred networks, and their effect on inferring accurate network characteristics. The analyses performed in this work have led to an improved reconstruction method by overcoming statistical limitations, namely *type I* and *type II errors* are taken into account in the inference thereby enhancing the results.

After a brief introduction of the mathematical background (Chapter 2), a simulation study investigating the impact of *type I* and *type II errors* on network topologies and characteristics is presented in Chapter 3. These results are analysed in a more theoretical framework and an analytic formula of the density of the biased node degree distribution is found in Chapter 4. Additionally, the inverse problem is studied - the density of the node degree distribution of the true underlying network is found as a function of the detected node degree distribution and the probabilities of *type I* and *type II errors*. A further analysis shows that this procedure is robust with respect to errors in α and β as they typically occur when they have to be estimated from data (Chapter 4). This implies that wrong estimates of *type I* and *type II errors*, within certain bounds, do not cause the reconstruction of the node

degree distribution to be rendered invalid.

Mathematical properties of the functional relationship between the true and the detected node degree distributions found in Chapter 4 are discussed in Chapter 5. The analysis in Chapter 4 is advanced in Chapter 6, suggesting an iterative procedure to reconstruct the node degree distribution in the case of uncertain estimates of α and β; this procedure is especially useful when large errors on the estimates of α and β are expected, and consequently the robustness, shown in Chapter 4, is not guaranteed. Lastly, in Chapter 7 networks of coupled oscillators, which are often used as models in various applications, are investigated. The investigation presented in the previous chapters is here reversed - the study focuses on the influence of network characteristics on the probability of *type I* and *type II errors*. These results are then applied when the underlying true network is not known a priori, to improve the network reconstruction, introducing a new concept of threshold.

Chapter 2

Mathematical background

This chapter is dedicated to mathematical definitions and results that form the basics of the work presented in the following chapters. The background knowledge presented here belongs to three different fields, namely probability theory, network theory, and test of hypothesis.

2.1 Probability theory

This section follows the ideas of [23, 63]. If not stated otherwise, the ideas and concepts have been taken from these books.

The field of probability theory refers to the study of random events. The set of all possible outcomes of an experiment, also referred to as realisations, is called *sample space* Ω, and an event is a subset of Ω. When different possible outcomes exist, the field of probability theory provides methods to quantify the likelihood of the realisations or in general events.

Operations and relationships from set theory can be used to study probability theory, such as the complement, the union, and the intersection. Consider the sample space Ω, the complement A^c of an event A is the set containing all the realisations of Ω that are not in A, i.e., $A^c = \Omega \setminus A$. Consider two or more events $A_1, \cdots, A_n$, the union $\bigcup_{i=1}^{n} A_i$ is the set of all the realisations contained in at least one of the events. The intersection $\bigcap_{i=1}^{n} A_i$ is the set of the realisations contained in all events. The events $A_1, \cdots, A_n$ are said to

be disjoint, or mutually exclusive, if the intersection of each pair of events is the null event, i.e., $A_i \cap A_j = \emptyset$ for every $i \neq j$.

Given a sample space Ω, the first aim of probability theory is to assign a measure to quantify the chance of each event in Ω to occur; this measure is called probability $\mathbb{P}$. To ensure consistency, the probability $\mathbb{P}$ has to satisfy the following axioms:

(i) $\mathbb{P}(\Omega) = 1$,

(ii) for every event A in Ω, $\mathbb{P}(A) \geq 0$, and

(iii) if $A_1, \cdots, A_n$ are disjoint events in Ω, then $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Some fundamental properties can be derived from the axioms described above, such as:

(i) $\mathbb{P}(\emptyset) = 0$,

(ii) for every event A in Ω, $0 \leq \mathbb{P}(A) \leq 1$,

(iii) if $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$,

(iv) for every event A in Ω, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, and

(v) for events A and B in Ω, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

In general, if Ω is finite and all the realisations are equally likely, the probability of an event A is the ratio of the cardinalities of A and the sample space, i.e., $\mathbb{P}(A) = |A|/|\Omega|$. Combinatorics is a useful tool to establish the cardinality of sets. An ordered subset is called a permutation, while an unordered one is called a combination. The number of permutations of k objects from a set of n elements is

$$P_{n,k} = \frac{n!}{(n-k)!} \tag{2.1}$$

without repetitions, and

$$P^r_{n,k} = n^k \tag{2.2}$$

with repetitions. The number of combinations of k objects from a given set of n elements is

$$C_{n,k} = \binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{2.3}$$

without repetitions, and

$$C^r_{n,k} = \binom{n+k-1}{k-1} = \frac{(n+k-1)!}{n!(k-1)!} \tag{2.4}$$

with repetitions.

An important concept in probability theory is independence. Generally, it is crucial to check whether two or more events are mutually independent, since the realisation of one event might influence the probability of the other. Two events A and B in Ω are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, and dependent otherwise, [42]. Generalising this concept, events $\{A_i\}_{i \in I}$ in Ω are independent if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i) \tag{2.5}$$

for every finite subset J of I.

If two events are dependent, then the probability of an event is conditioned by the realisation of the other. The conditional probability of event A given event B occurred, with $\mathbb{P}(B) > 0$, is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{2.6}$$

If two events A and B are independent, the probability of A should not be influenced by the occurrence of B, i.e., $\mathbb{P}(A|B) = \mathbb{P}(A)$.

One of the main concepts in probability theory is the so-called random variable. Random variables are a generalisation of the concept of events. It is in general possible to associate any outcome of an experiment to a real number; a random variable is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to each realisation in Ω. If the sample space Ω is either finite or countably infinite, the random variable is said to be discrete. If Ω is continuous, X is said to be continuous; a precise definition is beyond the

scope of this work.

In both the discrete and continuous case, the cumulative distribution function (CDF), or short distribution function, of a random variable X is defined as the function $F_X : \mathbb{R} \to [0, 1]$ such that for every realisation x of X

$$F_X(x) = \mathbb{P}(X \leq x), \tag{2.7}$$

is the probability that the random variable X takes values smaller than or equal to x.

The probability mass function (PMF), or short probability function of a discrete random variable X is the function $f_X : \mathbb{R} \to [0, 1]$ such that

$$f_X(x) = \mathbb{P}(X = x). \tag{2.8}$$

The analogous of PMF for a continuous random variable X is the probability density function (PDF), or short density. It is defined implicitly as the function $f_X : \mathbb{R} \to [0, 1]$ with

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{2.9}$$

In many situations, it is useful to consider more than one random variable, and it is convenient to introduce the so-called joint probability. Let X and Y be two discrete random variables, the joint probability mass function is

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) \tag{2.10}$$

and the respective marginal probability mass functions of X and Y are $f_X(x) = \sum_y f_{X,Y}(x, y)$ and $f_Y(y) = \sum_x f_{X,Y}(x, y)$.

In the case that X and Y are two continuous random variables, the joint probability density function $f_{X,Y}(x, y)$ is defined as

$$\mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(\bar{x}, \bar{y})d\bar{x}d\bar{y} \tag{2.11}$$

and the respective marginal probability density functions of X and Y are

$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy$ and $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx$ [31].

Another important property of random variables is their expected value. The expected value, or mean, of a random variable X is

$$\mathbb{E}(X) = \sum_x x f_X(x) \tag{2.12}$$

if X is discrete, and

$$\mathbb{E}(X) = \int x f_X(x)dx \tag{2.13}$$

if X is continuous. The notations μ, μ_X, or $\mathbb{E}(X)$ are usually used synonymously. One of the main properties of the expected value is linearity, i.e., $\mathbb{E}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i \mathbb{E}(X_i)$ for random variables $X_1, \cdots, X_n$ and $a_1, \cdots, a_n \in \mathbb{R}$. Also, if $X_1, \cdots, X_n$ are independent, then $\mathbb{E}(\prod_{i=1}^n X_i) = \prod_{i=1}^n \mathbb{E}(X_i)$.

To have a better insight about the general behaviour of a random variable, it is useful to understand the dispersion of the possible outcomes from the expectation. To measure this quantity, it is either the variance or its square root, the standard deviation, that is used. The variance

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] \tag{2.14}$$

of a random variable X is defined as the expected value of the quadratic difference of X and its expected value; the standard deviation

$$\sigma_X = \sqrt{\mathbb{V}(X)}, \tag{2.15}$$

is the square root of the variance and it is denoted by σ, σ_X, or $SD(X)$.

Two measures to quantify the joint variability of two random variables and the strength of their linear relationship are the covariance

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \tag{2.16}$$

and the correlation

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{2.17}$$

where X and Y are random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y. Note that if X and Y are independent random variables, then their covariance and correlation are zero.

In many scientific fields, some prototypical probability distributions are often used to describe various phenomena. The most frequently and widely used probability distributions are the uniform, Bernoulli, binomial, and normal distributions. The continuous random variable X is said to be *uniformly distributed* in $[a, b]$, if its PDF is

$$f(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}, \tag{2.18}$$

where $\mathbb{I}$ is the characteristic function. The expected value and variance of a uniform distribution are $\mathbb{E}(X) = (a+b)/2$ and $\sigma^2 = (b-a)^2/12$, respectively.

A discrete random variable X has *Bernoulli* distribution if there are only two possible realisations of X, say 0 and 1, also named failure and success. Call $p \in [0, 1]$ the probability $\mathbb{P}(X = 1) = p$, the Bernoulli PMF

$$f(x) = p^x(1-p)^{1-x} = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases} \tag{2.19}$$

is defined for $x \in \{0, 1\}$; its expected value is $\mathbb{E}(X) = p$, and its variance is $\sigma^2 = p(1-p)$.

The *binomial* distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ is the discrete probability distribution of the number of successes k of n independent trials which have Bernoulli distribution. It is denoted by $\mathcal{B}(n, p)$ and the corresponding PMF

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.20}$$

is defined for $k \in \mathbb{N}$. The expected value and variance of a binomial distribution are $\mathbb{E}(X) = np$ and $\sigma^2 = np(1-p)$.

A continuous random variable has *normal* distribution with mean μ and

SD σ, and it is denoted by $X \sim N(\mu, \sigma^2)$, if its PDF is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.21}$$

One of the main results of probability theory is the *Central Limit Theorem* [58]. It says that the distribution of the sum of independent random variables can be approximated with a normal distribution. This theorem is a key concept in probability theory since only few hypotheses about the distribution of the variables are required, which implies that it can be applied in many cases. The *Central Limit Theorem* states that even though the distribution may be far from being normal, yet for large sample size, the distribution of the standardized sample mean is approximately standard normal. Namely, given $X_1, \cdots, X_n$ independent and identically distributed random variables with finite mean μ and standard deviation σ, then for every $a \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq a\right) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \tag{2.22}$$

2.2 Network theory

In this section, an introduction to the fundamental theoretical tools used to describe and analyse networks is given. If not stated otherwise, the ideas and concepts have been taken from [48].

A network is defined as a set of nodes with links between them. In graph theory, a branch of mathematics that studies networks, a different notation is used: networks are called graphs, and nodes and links are called vertices and edges, respectively. In this thesis, the notation form network theory and graph theory are used synonymously.

The most common and efficient way to represent a network mathematically is the adjacency matrix. Consider a network G with n vertices labelled $1, \cdots, n$ with $n \in \mathbb{N}$. The adjacency matrix $\mathcal{A}$ of G is the $n \times n$ matrix with

elements

$$\mathcal{A}_{ij} = \begin{cases} 1 \text{ if there is link from node } i \text{ to node } j, \\ 0 \text{ otherwise.} \end{cases} \tag{2.23}$$

Networks can be directed or undirected. In an undirected network, connection of v_1 to v_2 implies the connection of v_2 to v_1. Note that this implies that the adjacency matrix is symmetric. In a directed network, this symmetry is broken, therefore if a path from v_1 to v_2 exists, a path from v_2 to v_1 does not necessarily exist. Therefore, the adjacency matrix is not necessarily symmetric.

To describe various real-world scenarios it is useful to define a network whose links are not only binary connections. Weighted networks have been introduced in the literature, and they are characterised by their adjacency matrix whose elements are real numbers.

In the literature, many other types of networks exist, e.g., simple, multigraph, or multi-layer. It is beyond the scope of this work to describe them in detail.

2.2.1 Network characteristics

To capture particular features of a network, a variety of measures or characteristics have been introduced [50]; here, some of the key network characteristics are described.

In an undirected network, the vertex degree describes the number of links of a node; if the vertex v has k edges attached, its vertex degree is $d_v = k$. For directed networks, the vertex degree is characterised by the *vertex in-degree* d_v^{in}, which is the number of edges pointing towards it, and the *vertex out-degree* d_v^{out}, which is the number of edges originating from it. In a network of n nodes, both the in-degree and the out-degree of a vertex are numbers between 0 and $n-1$; self-connections are not considered. Usually, in a directed network the *vertex degree* $d_v = d_v^{in} + d_v^{out}$ refers to the sum of the vertex in-degree and the vertex out-degree.

The frequency distribution of the vertex degrees is called *vertex degree*

distribution; it is an important property of the entire network, and a defining characteristic of the network structure.

For two randomly selected nodes v_1, v_2 in a network of n nodes, the shortest path length $\ell_{v_1 v_2}$ measures the number of links separating them if the shortest path is taken. For connected nodes v_1, v_2, when the oriented edge $v_1 \to v_2$ exists, the shortest path length is $\ell_{v_1 v_2} = 1$. The average path length

$$\gamma = \frac{1}{n(n-1)} \sum_{v_1 \neq v_2} \ell_{v_1 v_2} \tag{2.24}$$

gives a measure for the entire network, for $n > 1$. The efficiency

$$\epsilon = \frac{1}{n(n-1)} \sum_{v_1 \neq v_2} \frac{1}{\ell_{v_1 v_2}} \tag{2.25}$$

is defined as the sum of the inverse of the shortest path lengths. If a network is unconnected, i.e., a network that has two nodes for which the path between them does not exist, the shortest path is infinitely long for unconnected nodes, hence considering the average of the shortest path lengths is not meaningful. Efficiency for unconnected nodes will be zero, therefore a meaningful network average of the efficiency can be obtained.

Assortativity or assortative mixing by degree is a characteristic that expresses the preference for high-degree vertices to attach to other high-degree vertices, and low to low, respectively. Instead, in a network that shows disassortative mixing, high-degree vertices are preferencially connected to low-degree ones [45]. Assortativity can be measured as the correlation coefficient of degrees of pairs of connected nodes

$$\text{ass} = \frac{\sum_{ij} (\mathcal{A}_{ij} - d_i d_j / 2e^{tot}) d_i d_j}{\sum_{ij} (d_i \delta_{ij} - d_i d_j / 2e^{tot}) d_i d_j}, \tag{2.26}$$

where $\mathcal{A}$ is the adjacency matrix, d_i is the degree of the vertex i, e^{tot} is the total number of edges, and δ_{ij} is the Kronecker delta. When there is not assortative mixing, the terms in the sum at the numerator simplify and the correlation is zero; for assortative mixing, the correlation is positive, and it

goes to 1 when $d_i = d_j$ for every existing link $i \to j$; for disassortative mixing, the correlation is negative.

The tendency of a network to form tightly connected neighbourhoods can be measured by the clustering coefficient. There exist two definitions of the clustering coefficient as measures for the entire network - the average local clustering coefficient and the global clustering coefficient. The latter is defined as the ratio of three times the number of triangles to the number of pairs of adjacent edges in a network [10, 47]. For undirected networks, the local clustering coefficient $\mathcal{C}$ measures the probability that two vertices, that are connected to another vertex, are connected to each other. In other words, the local clustering coefficient $\mathcal{C}_v$ of a vertex v is the ratio between the number of triangles in the network with v as one vertex of the triangle, and the number of all pairs of vertices connected to v. The generalized definition of the local clustering coefficient of a vertex v for directed networks

$$\mathcal{C}_v = \frac{(\mathcal{A} + \mathcal{A}^T)^3_{vv}}{2[d_v(d_v - 1) - 2\mathcal{A}^2_{vv}]}, \tag{2.27}$$

where $\mathcal{A}$ is the adjacency matrix of a directed network, considers the directed triangles formed by v [28].

Using the idea of a local clustering coefficient, a new network characteristic, called the *detour degree*, is defined as part of the research for this thesis. For each edge e, the detour degree is the number of oriented triangles that have e as a side, and the other two sides oriented such that the combination of them creates a path with the same origin and end of e. Namely, for every oriented edge $v_1 \to v_2$ from node v_1 to node v_2, the *detour degree* $\Delta_{v_1 v_2}$ is the number of oriented paths of length 2 from v_1 to v_2. For example, in the case shown in Fig. 2.1, the detour degree is $\Delta_{v_1 v_2} = 2$, since there exist two directed paths of length 2 from v_1 to v_2 through k_1 and k_2. Since the edge between v_1 and k_3 is oriented towards v_1, a path from v_1 to v_2 through k_3 does not exist.

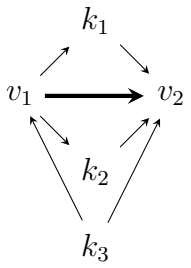


Figure 2.1: Example of detour degree $\Delta_{v_1 v_2} = 2$.

2.2.2 Network topologies

In many scientific fields, some prototypical networks that have some specific characteristics in common, but random in other aspects, are often used to describe various complex systems. These types of networks are called random networks. Here, some of the most frequently and widely used network topologies are presented.

Erdős-Rényi networks are random networks in which the set of nodes is fixed, and each pair of nodes is connected with independent probability p_c. The probability mass function of the node degree distribution of an Erdős-Rényi network

$$\mathbb{P}(d_v = k) = \binom{n-1}{k} p_c^k (1-p_c)^{n-1-k} \tag{2.28}$$

is a binomial distribution, where n is the number of nodes in the network.

Watts-Strogatz networks are also referred to as small-world networks. They are characterised by a high local connectivity with some long-range "short-cuts". Watts-Strogatz networks are built from a regular network, i.e., a network where every node has the same node degree. With probability p_r each link is rewired to another node randomly selected. The node degree distribution has probability mass function

$$\mathbb{P}(d_v = k) = \sum_{i=\max(2c-k,0)}^{\min(n-1-k,2c)} \binom{2c}{i} \left(\frac{p_r}{2}\right)^i \left(1-\frac{p_r}{2}\right)^{2c-i} e^{-cp_r} \frac{(cp_r)^{k-2c+i}}{(k-2c+i)!}, \tag{2.29}$$

in the assumption of the number of nodes $n \gg c$, where $2c$ is the node degree of every node in the initial regular network [44].

Barabási-Albert networks are constructed using a preferential attachment procedure. The main feature of these types of networks is that their node degree follows a power law; they are so-called scale free networks. They are constructed by adding nodes to an existing network. Each new node, with a certain number b of links attached to it, is connected to the network. The probability for one of these b links to be connected with any existing node is proportional to the degree of that node. The node degree distribution has probability mass function [5]

$$\mathbb{P}(d_v = k) = \frac{2b(b+1)}{k(k+1)(k+2)}. \tag{2.30}$$

2.3 Test of hypothesis

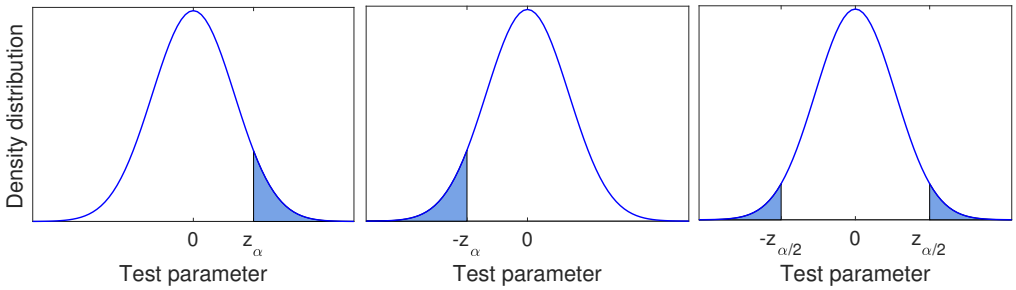
The ideas and concepts of this section have been taken from [23], if not stated otherwise.

A statistical hypothesis is a claim or assumption about a certain parameter or probability distribution. To verify if this assumption is true, a test needs to be performed. A test of hypothesis considers two complementary hypotheses; the null hypothesis H_0 is the initial assumption that is consider to be true, and the alternative hypothesis H_1 or H_a is the complement of H_0. There are only two possible conclusions from a test of hypothesis: reject H_0 or not reject H_0.

To decide whether the null hypothesis is rejected, the concept of $p-$value has been introduced. The $p-$value is the probability of finding values more extreme or equal to the observed results, when the null hypothesis is true.

The significance level α is the threshold for the $p-$value of the test, i.e. if the $p-$value of the test is smaller than or equal to the significance level, the null hypothesis is rejected. The standard value for the significance level is usually set to $\alpha = 0.05$, meaning that there is 5% chance of erroneously rejecting the null hypothesis, and the test is said to be statistically significant [17, 22, 23, 33, 35, 57, 59].

It is possible to perform three kind of hypothesis tests - upper-tailed, lower-tailed, and two-tailed; the difference between them resides in the choice



(a) Upper-tailed test (b) Lower-tailed test (c) Two-tailed test

Figure 2.2: Three kind of hypothesis tests: upper-tailed, lower-tailed, and two-tailed; the blue areas correspond to the rejection regions, and they are equivalent to α, for each case.

of the alternative hypothesis. Assume that the null hypothesis is $H_0 : \mu = \mu_0$, and the test statistics is $z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{N}}$, where $\overline{x}, \sigma,$ and N are the sample mean, the standard deviation, and the sample size, respectively. The alternative hypothesis can be formulated as shown in Table 2.1, where z_α is the value for z such that the rejection area corresponds to α.

Test	Alternative hypothesis	Rejection region	Figure
Upper-tailed	$\mu > \mu_0$	$z \geq z_\alpha$	Fig. 2.2a
Lower-tailed	$\mu < \mu_0$	$z \leq -z_\alpha$	Fig. 2.2b
Two-tailed	$\mu \neq \mu_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	Fig. 2.2c

Table 2.1: Three kind of hypothesis tests and corresponding rejection regions.

When a test of hypothesis is performed, H_0 might erroneously be rejected, or H_0 might erroneously not be rejected when it is false. These possible errors are

- *type I error* - rejecting the null hypothesis H_0 when it is true, and
- *type II error* - not rejecting H_0 when it is false.

The *type I error* corresponds to a false positive conclusion, and the probability of this error to occur corresponds to the significance level α of the

test. The *type II error* corresponds to a false negative conclusion, and the probability of this error to occur is referred to as β.

Sensitivity and specificity are measures of the performance of a test. A test that has a low proportion of *type I errors* is said to have high specificity; sensitivity is related in the same way to *type II errors*.

When a test of hypothesis is used to establish if two variables are linearly correlated, a measure of linear correlation is introduced (Eq. (2.17)). When the correlation coefficients are estimated from a sample, they have a certain distribution that depends on the dimension of the sample and the true correlation coefficient. As shown in [37], for data that follow a bivariate normal distribution, the exact probability density function of the estimated correlation coefficients r for a sample of N data points with true correlation coefficient ρ is

$$f(N,\rho,r) = \frac{(N-2)\Gamma(N-1)(1-\rho^2)^{\frac{N-1}{2}}(1-r^2)^{\frac{N-4}{2}}}{\sqrt{2\pi}\ \Gamma(N-\frac{1}{2})(1-r\rho)^{N-\frac{3}{2}}} \quad {}_2F_1\left(\frac{1}{2},\frac{1}{2};\frac{2N-1}{2};\frac{r\rho+1}{2}\right), \tag{2.31}$$

where Γ is the gamma function and ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the Gaussian hypergeometric function. Note that this function is defined for $-1 \leq r \leq 1$ and $r\rho \neq 1$.

Consider a two-tailed test with null hypothesis of no correlation, i.e. the Pearson correlation coefficient is 0. The significance level α can be expressed as a function of the coefficient r_τ such that

$$\alpha = \int_{-1}^{-r_\tau} f(N,0,r)\ dr + \int_{r_\tau}^{1} f(N,0,r)\ dr, \tag{2.32}$$

meaning that the rejection region is obtained for coefficients $-1 \leq r \leq -r_\tau$ and $r_\tau \leq r \leq 1$, as shown in Fig. 2.3. Consequently, estimated coefficients $-r_\tau < r < r_\tau$ do not lead to a rejection of the null hypothesis. The probability of false negative conclusions

$$\beta = \int_{-r_\tau}^{r_\tau} f(N,\rho,r)\ dr \tag{2.33}$$

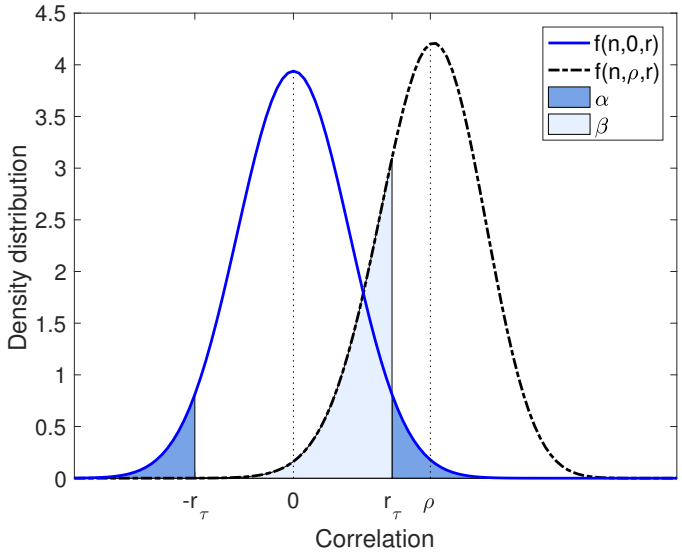


Figure 2.3: Density distribution of correlation coefficients r for a sample of $N = 100$ data points with true correlation coefficient $\rho = 0$ (solid blue line) and $\rho = 0.25$ (dashed black line), i.e. $f(100, 0, r)$ and $f(100, 0.25, r)$ as described by Eq. (2.31). The area in blue is the α value corresponding to $r_\tau = 0.18$, and the area in light blue is the respective value of β.

is calculated once r_τ is fixed, see Fig. 2.3.

Equations (2.32) and (2.33) show the dependence of α and β on the variables N, r_τ, ρ, i.e., $\alpha = \alpha(N, r_\tau)$ and $\beta = \beta(N, \rho, r_\tau)$. Therefore, $\beta(N, \rho, \alpha)$ is a function of α.

Figure 2.4 shows the relation between α and $1/\beta$ for $N = 100$ data points taken from a bivariate normal distribution. Different values for the true correlation coefficient are used, i.e., ρ varies from 0.3 to 0.45 in steps of 0.01.

2.4 Conclusion

In this chapter mathematical definitions and results belonging to three different scientific fields are presented. Probability theory, network theory, and test of hypothesis are the background knowledge that form the basis of the work presented in this thesis. Network is the topic of main interest of this thesis. Probability theory gives the basis of not only to define random networks,

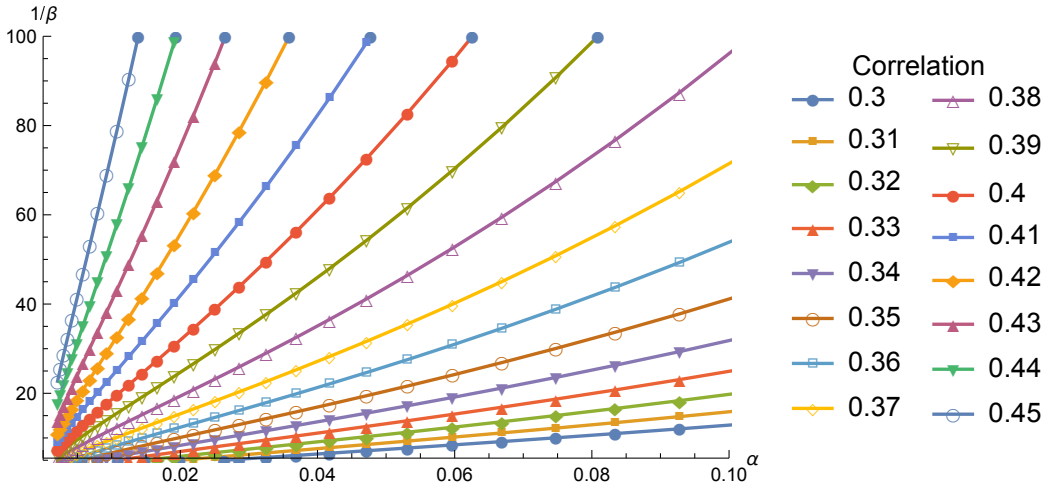


Figure 2.4: Relation between α and $1/\beta$ for correlation of 100 data points. Colours refer to different correlation coefficients, as indicated in the legend.

but also to express their characteristics, such as the vertex degree distribution. With the test of hypothesis, two fundamental concept are introduced: *type I* and *type II errors*.

In this thesis new methodologies to improve network reconstruction from data are suggested. These techniques rely on statistical methods, and take into account *type I* and *type II errors* in the reconstruction analysis.

Chapter 3

Improving network inference: the impact of false positive and false negative conclusions about the presence or absence of links

As part of the research for this thesis, the following manuscript has been published [15]. This chapter discusses the concepts presented in this publication.

3.1 Introduction

Recently, many research groups have focused on the inference of networks from data such as brain networks from observed electroencephalography or functional magnetic resonance imaging data [12, 52, 53, 61]. Particular emphasis is paid to the understanding of the normal functioning, e.g. healthy brain, as well as malfunctioning, e.g. diseased brain, of these networks. In the example of the brain, this promises to disclose information about how the brain processes signals and how alterations thereof cause specific diseases. A key hypothesis is that important characteristics are not specific to individual subjects but rather common in a given population. This is reflected by

the fact that brain networks, but also other networks, are typically classified into few main prototypic networks [46, 48], e.g., Erdős-Rényi [26, 27], Watts-Strogatz [64, 65], Barabási-Albert [4, 5] networks. These are three types of random networks, see Sec. 2.2.2. In this chapter, binary undirected networks of these three topologies are considered.

These prototypical models for networks are in turn characterised by few parameters; procedures have been described to generate these networks with their well-established characteristics [46, 48], see Sec. 2.2.1. Some of the key characteristics are the node degree distribution, the number of links, the global clustering coefficient, and the efficiency. In this chapter, these characteristics are considered since they are meaningful in random networks and give a global description in large networks [48].

In the *Inverse Problem*, the challenge is to infer the network topology from data. Two challenges are particularly relevant: (i) the reliable inference of links in the network once the nodes have been fixed [43, 69] and (ii) the successful usage of the characteristics above to uniquely determine the topology of network [7, 8].

Classical statistical methods to estimate links in a network aim to identify present links with high certainty, see Chapter 1. Typically the standard value of 0.05 for the probability of false positive links α is often chosen. The decision to set $\alpha = 0.05$ does not take into account the probability of false negative links β. The investigation in this chapter focuses on whether these common rules of *type I* and *type II errors* should be modified to achieve a more reliable inference of the correct topology of network. To this aim, their influence on the network topology and characteristic is analysed.

This chapter is structured as follows. Section 3.2 explains statistical errors and their influence on the network topology. A simulation study in the case of Erdős-Rényi, Watts-Strogatz and Barabási-Albert networks is presented in Sec. 3.3.

3.2 Inference reliability

Several methods have been suggested in the literature to address the challenge of reliable inference of links in the network. To determine the presence of links, these techniques usually rely on statistical inference [17, 22, 23, 33, 35, 57, 59].

Let G denote the true network. As a consequence of the choice of α and thereby β, leading to a non-zero probability of detecting false positive and false negative links, the detected network G^D will be a "mixture" of true links, false positive links, absent links and false negative links. Therefore, the number of detected links is generally different to the number of links of G. Also the node degree distribution, the global clustering coefficient and the efficiency are in general biased. To quantify the bias, a distance between distributions is used for each characteristic. Several distance measures are conceivable and have been investigated; for sake of simplicity and to make the arguments clearer, only the distance

$$\delta = |\mu_1 - \mu_2| \tag{3.1}$$

between two distributions is considered, as the modulus of the difference of the distribution's mean values. For example, the distance between the node degree distribution of G, which has mean μ_G, and the node degree distribution of G^D, which has mean μ_{G^D}, is $\delta = |\mu_G - \mu_{G^D}|$.

To investigate the relation between α and β, $N = 100$ data points taken from a bivariate normal distribution are considered, see Sec. 2.3. This choice is motivated by the fact that the Pearson correlation coefficient is used to establish if two variables are linearly correlated. To inspect in particular links with medium strength, the correlation ρ varies between 0.3 and 0.45 in steps of 0.01. Using Eq. (2.31), α and β are found in Eqs. (2.32)-(2.33). The relation between $1/\beta$ and α is shown in Fig. 2.4.

Visual inspection of Fig. 2.4 shows that a linear relationship is a good approximation. Fitting linear functions to the curves shows that their respective slopes vary between $0.1 \cdot 10^{-3}$ and $1.1 \cdot 10^{-3}$. These slopes will differ

if different parameters, such as the number of data points N, are chosen. The more data points are considered the more accurate the analysis. Note that the inverse proportionality of α and β implies that an infinite number of data points N is needed to have both α and β equal to zero.

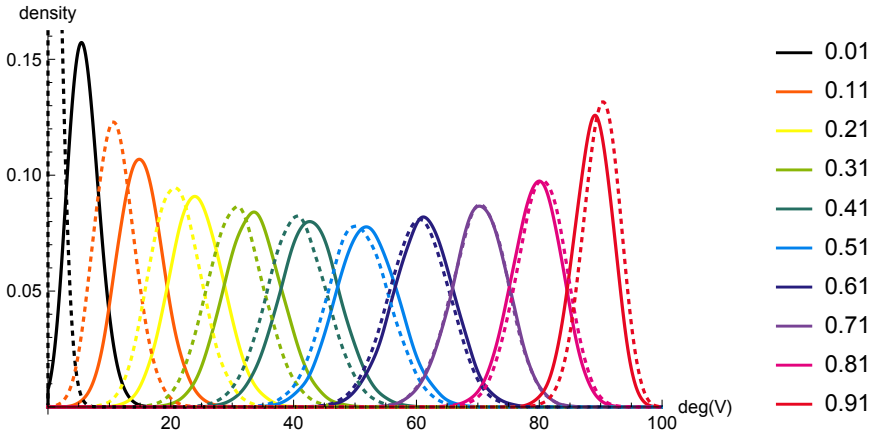
As an example of how the choice of α and consequently β affects the estimated network characteristics, Erdős-Renyi networks G_{p_c} are considered. The probability of connection p_c was varied between 0.01 and 0.99 in steps of 0.01. The detected networks $G^D_{p_c}$ were generated by artificially introducing false positive links with probability α and false negative links with probability β. The probability α varies between 0.005 and 0.1 in steps of 0.001, the relation between α and β was fixed by

$$\beta = \frac{10^{-3}}{\alpha}, \tag{3.2}$$

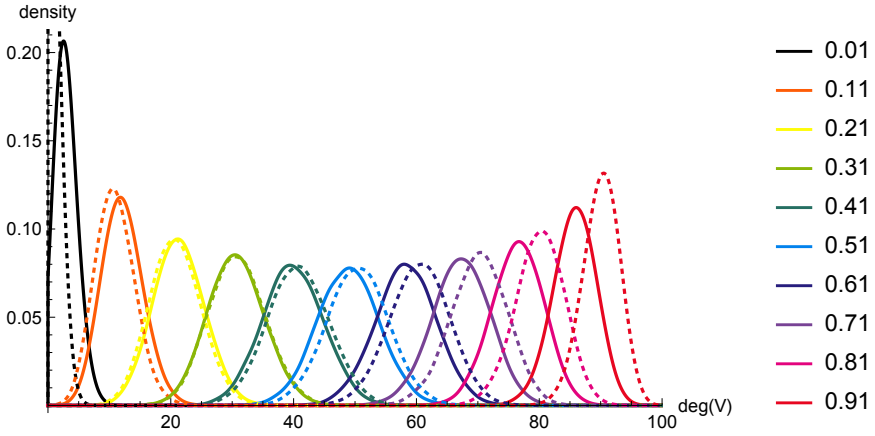
which represents a choice motivated above. Moreover, this choice corresponds to a method, which has relatively high sensitivity and specificity, i.e., $0.005 < \alpha, \beta < 0.2$. For each value of p_c and α, 200 networks with $n = 100$ nodes were generated. Figure 3.1 shows the true densities of the node degree derived from G_{p_c} (dashed lines) together with the average densities derived from the detected networks $G^D_{p_c}$ (solid lines). Results for $\alpha = 0.05$ and $\alpha = 0.02$ are shown. Different colours represent different Erdős-Renyi networks defined by the parameter p_c, for clarity, densities are plotted for p_c in steps of 0.1 only.

The distances (Eq. 3.1) between the true density and the detected density for each pair of p_c and α are shown in Fig. 3.2. For some values of p_c the distance is negligible, which means the detected node degree is almost identical to the true node degree. The optimal α, i.e. the one with the smallest distance between true density and detected density, depends on p_c.

To have a general result for the optimal choice of α when estimating a network characteristic of a given network topology, the sum over p_c is taken to marginalise out the influence of p_c for each α. This integrated quantity is



(a) Solid lines: $G^D_{p_c}$ with $\alpha = 0.05, \beta = 0.02$. Dotted lines: G_{p_c}



(b) Solid lines: $G^D_{p_c}$ with $\alpha = 0.02, \beta = 0.05$. Dotted lines: G_{p_c}

Figure 3.1: Densities of the node degree distributions for Erdős-Rényi networks of $n = 100$ nodes and different parameters $p_c = 0.01, ..., 0.91$ in steps of 0.1 represented by colour. The densities of the node degree distributions for the respective original networks G_{p_c} (dotted lines) and detected networks $G^D_{p_c}$ (solid lines) are shown.

called the total distance δ_{tot}, i.e.

$$\delta_{tot} = \sum_{p_c} \delta(p_c). \tag{3.3}$$

To identify the optimal choice of α, the interest is in finding where the minimum of the total distance δ_{tot} is located. Figure 3.3 shows δ_{tot} for the example

of the node degree of Erdős-Rényi networks. In this example, the minimum of δ_{tot} is located at $\alpha = 0.030$. This suggests that in order to optimally reconstruct the node degree of an Erdős-Rényi network $\alpha = 0.03$ should be chosen, which is close to the standard choice of $\alpha = 0.05$ but distinctively smaller.

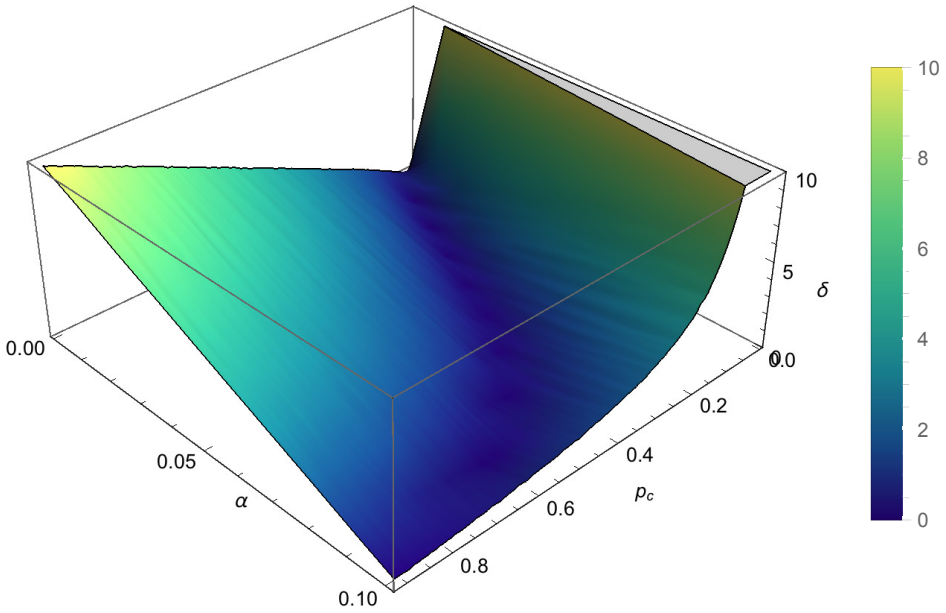


Figure 3.2: Distance δ between node degree distributions of Erdős-Rényi networks with 100 nodes depending on α and p_c. Distance δ is measured by calculating the difference between the mean of two corresponding distributions, Eq. (3.1). Colour code expresses distance values.

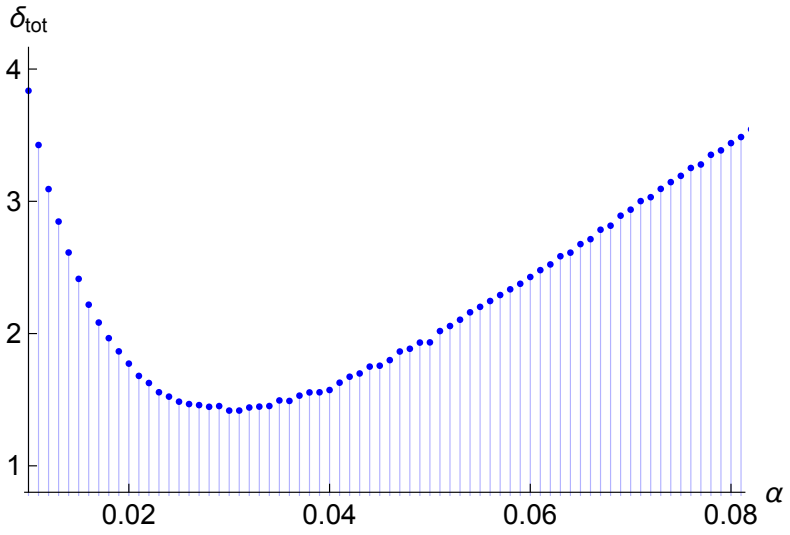


Figure 3.3: Total distances δ_{tot} between node degree distributions of Erdős-Rényi networks depending on α. The minimum is located at $\alpha = 0.030$.

3.3 Results

The analysis presented in the previous section is applied to Erdős-Rényi, with probability of connection p_c [Eq. (2.28)], Watts-Strogatz, with probability of rewiring p_r, and $c = 2$ [Eq. (2.29)], and Barabási-Albert networks, with parameter b [Eq. (2.30)].

For each network topology, four different network characteristics are investigated: node degree, number of links, global clustering coefficient and efficiency. The distance δ, which depends on both α and the parameter of the network topology (p_c, b or p_r) is presented as density plot for all the investigated characteristics and network topologies in Fig. 3.4. All 12 investigated scenarios show a dependence of the distance on the choice of α, suggesting that an optimum exists. For some scenarios, in particular node degree and number of links for Watts-Strogatz and Barabási-Albert networks, dependence of the distance on the parameter (p_r or b) is negligible. For other scenarios such as the global clustering coefficient in Watts-Strogatz networks the question arises if marginalising out the influence of p_r is distorting the results. Detailed results for each network topology are presented below.

For Erdős-Rényi networks of $n = 50$, $n = 100$, and $n = 250$ nodes, p_c varies from 0.01 to 0.99 in steps of 0.01. Figure 3.1 shows an example of some of these values in steps of 0.1. The results of the total distance δ_{tot} for the node degree of Erdős-Rényi networks are shown in Fig. 3.3. The minimum of δ_{tot} is located at $\alpha = 0.030$ ($\beta = 0.033$). For the remaining network characteristics, the Erdős-Rényi networks also show a clear minimum of the total distance in dependence on α. The specific values of α for the respective minimal total distances however vary; they are summarised in Table 3.1. The optimal α for efficiency is noticeably smaller than for the other network characteristics. Moreover, a broad range for p_c is used to cover the broad spectrum of Erdős-Rényi networks. Marginalising out the dependence of the distance δ on p_c may therefore be distorting the results (see also dependence on p_c in Fig. 3.4). For a specific application, narrowing the range of p_c to values relevant for the application is recommended.

The set of Barabási-Albert networks of $n = 50$, $n = 100$, and $n = 250$

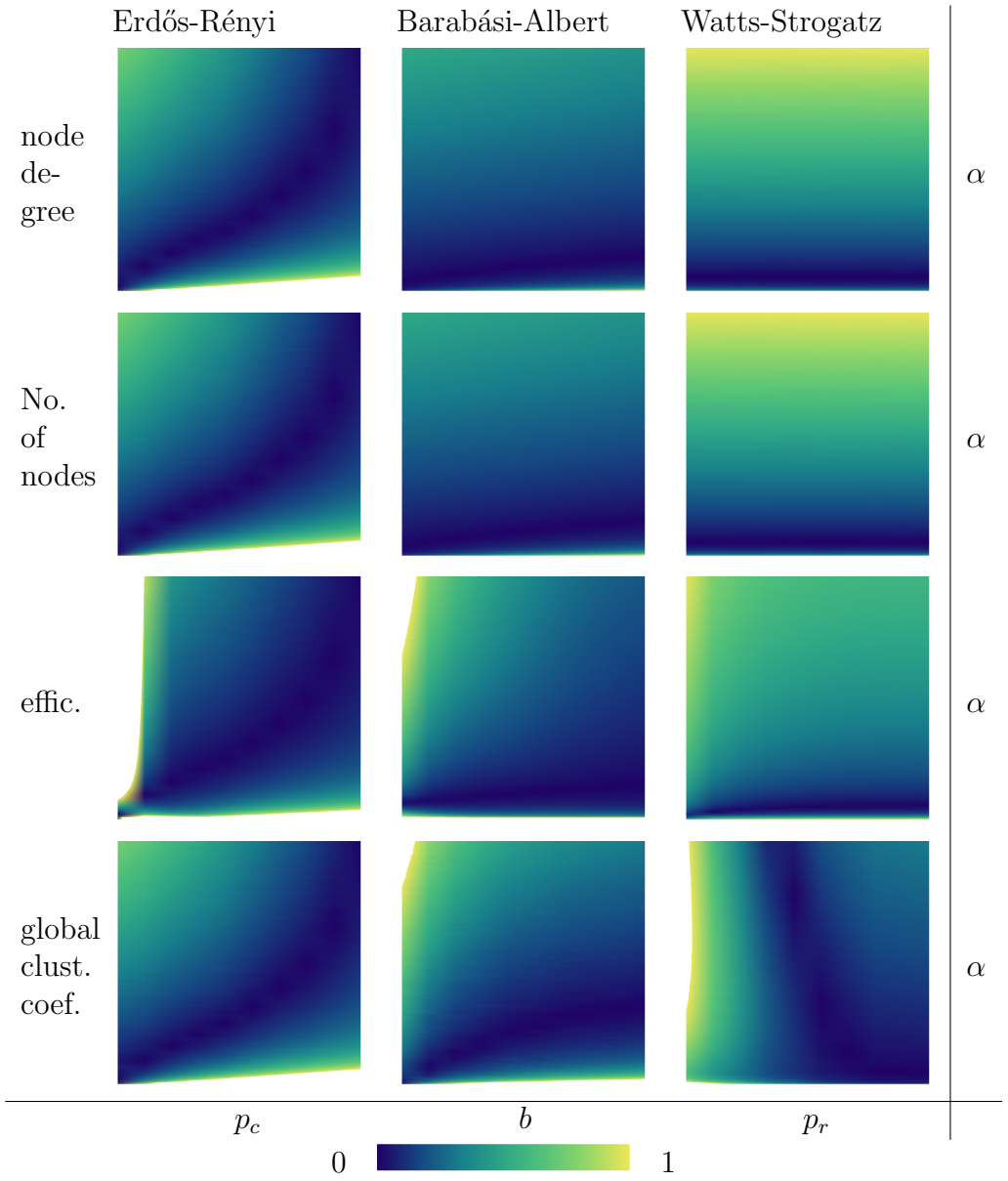


Figure 3.4: Distance δ for Erdős-Rényi, Barabási-Albert, and Watts-Strogatz networks with 100 nodes. Distance δ is calculated using Eq. (3.1), and it is normalised between 0 and 1. For each network topology the control parameter p_c varies from 0.01 to 0.99 in steps of 0.01, b from 1 to 10 in steps of 1, or p_r from 0.01 to 0.99 in steps of 0.01 on the $x-$axis, and the probability of false positive α from 0.005 to 0.1 in steps of 0.001 on the $y-$axis.

Network Topology	$n = 50$		$n = 100$		$n = 250$	
and Characteristic	α	β	α	β	α	β
Erdős-Rényi:						
node degree	0.031	0.032	0.030	0.033	0.031	0.032
number of links	0.031	0.032	0.030	0.033	0.031	0.032
global clustering coeff	0.035	0.029	0.031	0.032	0.031	0.032
efficiency	0.016	0.063	0.012	0.083	0.020	0.050
Barabási-Albert:						
node degree	0.018	0.056	0.012	0.083	0.007	0.143
number of links	0.018	0.056	0.012	0.083	0.007	0.143
global clustering coeff	0.024	0.042	0.021	0.048	0.019	0.053
efficiency	0.015	0.067	0.010	0.100	0.007	0.143
Watts-Strogatz:						
node degree	0.009	0.111	0.007	0.143	0.004	0.250
number of links	0.009	0.111	0.007	0.143	0.004	0.250
global clustering coeff	0.008	0.125	0.006	0.167	0.004	0.250
efficiency	0.009	0.111	0.006	0.167	0.004	0.250

Table 3.1: Table of α and β values for minimal total distances δ_{tot} of each network topology and characteristic.

nodes is chosen with parameters $b = 1$ to $b = 10$ varying in steps of 1. The total distances δ_{tot} for the node degree of networks with $n = 100$ nodes are shown in Fig. 3.5. The minimum is found for $\alpha = 0.012$ ($\beta = 0.083$), it is more pronounced than that for the Erdős-Rényi networks. Again, the other network characteristics and number of nodes all show a single minimum. The values for optimal α and β are summarised in Table 3.1. For this network topology a noticeably different optimal value for α was found for the clustering coefficient.

Finally, the set of Watts-Strogatz networks of $n = 50$, $n = 100$, and $n = 250$ nodes is chosen with parameter p_r varied between 0.01 and 0.99 in steps of 0.01. The distances between the distributions of the node degree, the number of links, the global clustering coefficient and the efficiency are analysed. The minimum of the total distance for the node degree of networks with $n = 100$ nodes is found for $\alpha = 0.007$ ($\beta = 0.143$). The total distances between node degree distributions for these networks are shown in Fig. 3.6.

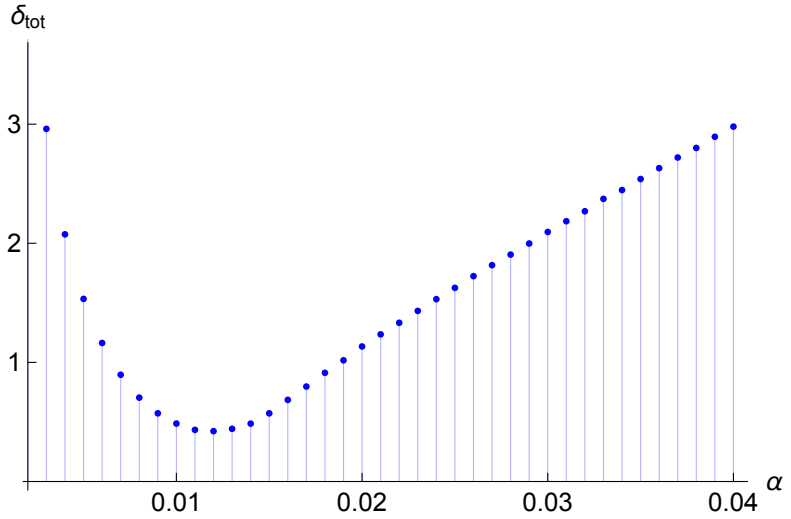


Figure 3.5: Total distance δ_{tot} between node degree distributions of Barabási-Albert networks of $n = 100$ nodes depending on α. The minimum is located at $\alpha = 0.012$ ($\beta = 0.083$).

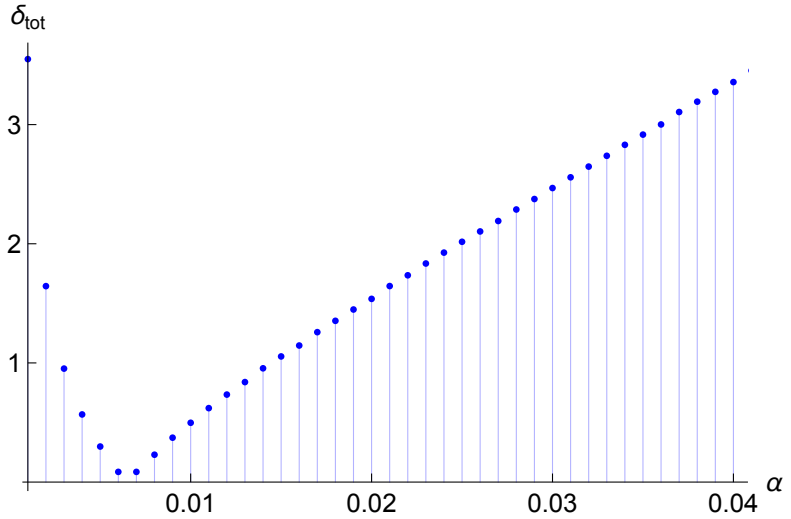


Figure 3.6: Total distance δ_{tot} between node degree distributions for Watts-Strogatz networks of $n = 100$ nodes depending on α. The minimum is located at $\alpha = 0.007$ ($\beta = 0.143$).

For all four characteristics and different values of n clear minima can be identified and the values for optimal α are similar (Table 3.1). The results for the efficiency however have to be interpreted with case as the distance shows a clear dependence on the parameter p_r (see Fig. 3.4).

3.4 Discussion

Three topologies of networks are considered, namely Erdős-Rényi, Watts-Strogatz, and Barabási-Albert. For each topology, and for a specific characteristic, e.g. the node degree distribution, the number of links, the efficiency or global clustering coefficient, the rate of false positive and false negative conclusions about links can be optimally chosen in order to have less biased reconstruction.

For Erdős-Rényi networks, the values for α identified with the method presented above, are close to standard choice of α of 0.05. Standard alpha values are suboptimal when the topology of network is different. For the set of Barabási-Albert networks, the values that yields the most reliable results of the node degree are $\alpha = 0.012$ and consequently $\beta = 0.083$. In this case, standard alpha values lead to a bigger distance between distributions of the node degree. The Watts-Strogatz networks yield the most reliable results for an even smaller value for $\alpha = 0.007$ and consequently $\beta = 0.154$. Moreover, for the optimal choice of α the corresponding β is rather high. This shows that the reliability of detecting individual false negative links in a network is less important than failing to recognise false positive links when network characteristics are estimated. Accepting a high rate of false negative links may thus be required when the aim is to infer a specific network characteristic.

This work shows that the standard choice of α of 0.05 is not optimal when the aim is to reconstruct the entire network topology. Moreover, α needs to be adjusted depending on specific network topologies and characteristics. For example, consider Erdős-Rényi networks with $p_c = 0.11$ and assume the relationship between α and β is Eq. 3.2. As result of 200 simulations, the mean of the node degree distribution of the original network G_{p_c} is 11 and

the mean for estimation using $\alpha = 0.05$ is 15. Choosing $\alpha = 0.03$ results in a mean of the node degree distribution of 13. The choice of $\alpha = 0.03$ is motivated by the assumption that the original network is known to be an Erdős-Rényi network with unknown parameter p_c, see Table 3.1. For the same study, when the aim is to infer the efficiency ϵ, the value is $\epsilon = 0.51$ for the true network, $\epsilon = 0.56$ for the one with $\alpha = 0.05$, and $\epsilon = 0.51$ when $\alpha = 0.012$. The more it is known about the network of interest, the more accurate the reconstruction is since the simulation study can be tuned accordingly.

As mentioned in Sec. 3.2, the relationship between α and β depends on the number of data points N, therefore the values of α and β leading to the minimal distance will change for different values of N. Nevertheless, the results will remain qualitatively the same.

The size of the network, i.e. the number of nodes, also influences the result. The number of false positive and false negative conclusions about the presence of links depends on the number of total links in the network. Keeping the same values of α and β and increasing, for example, the size of the network, leads to larger number of false positive and false negative detections of links. As shown in Table 3.1, for Erdős-Rényi networks the values of α and β leading to the minimal distance almost do not change. The reason is that the number of links increases proportionally with the number of nodes for each p_c. This does not happen for Barabási-Albert and Watts-Strogatz networks; the values of α leading to the minimal distance present a decreasing trend because of their constructions.

The node degree distribution, the number of links, the efficiency, and the global clustering coefficient have been considered as example characteristics to show that the results depend on the characteristic under investigation. Nevertheless, the approach described in this chapter can be readily applied to other characteristics, as well as other network topologies.

3.5 Conclusion

False conclusions about the presence of links in a network typically alter network characteristics, such as the node degree distribution, the number of links, the global clustering coefficient and the efficiency. Identification of the underlying network topology relies on these characteristics and is thus hindered by false conclusions about links as well. For these reasons, the analysis of false positive and false negative conclusions about links is of key importance.

In this chapter, assuming to know the underlying network topology, the influence of false positive and false negative conclusions about links in a network have been investigated. The values of α and β leading to minimal distance (difference in mean values) between the true network and the biased one change depending not only on the network topology, but also on the network characteristic of interest. Therefore, in the *Inverse Problem*, when the challenge is to infer the network topology from data, different values for α and β might be favourable when estimating different characteristics. In [15], the authors speculate that the simulation study can be used as an iterative procedure to achieve a better network reconstruction. Namely, when the network topology is not known a priori, various values for α can be chosen to perform the first iteration step of the network reconstruction. The result of this first step gives an idea of the network topology of interest. For the second iteration step the value for α can be adjusted according to the findings of the first step. This procedure can be iterated using the simulation study suggested in this chapter in each iteration step, ultimately leading to a reconstruction of the network tailored to its previously unknown network topology. This iterative procedure is presented in Chapter 6.

This result suggests various values for statistical inference could be considered within a simulation study to determine the optimal α for the network characteristic of interest. If several network characteristics are of interest, it may be useful to adjust the value of α for each characteristic.

The results presented in this chapter are analysed in a more theoretical framework in the next chapter.

Chapter 4

Analytical approach to network inference: investigating the degree distribution

As part of the research for this thesis, the following manuscript has been published [14]. This chapter discusses the concepts presented in this publication

4.1 Introduction

In this chapter, an analytical framework on network inference is presented; on the network level, it links the reconstructed network structure contaminated by *type I* and *type II errors* (Sec. 2.3) to the true underlying one. While the framework is rather general, the vertex degree distribution is used to derive the functional relationship between the reconstructed and true underlying network. This enables one to obtain superior estimates for the vertex degree distribution, see Sec. 2.2.1. It has been shown that including the vertex degrees into stochastic blockmodels improves their performance for statistical inference of group structure [36]. The functional relationship depends on the choice of *type I error*, *type II error* and the dimension of the network.

The chapter is structured as follows. In Sec. 4.2 a theoretical analysis of

the method is presented. Section 4.3 shows some cases where the method presented in Sec. 4.2 is applied.

4.2 Materials and methods

In Sec. 4.2.1, the influence of *type I* and *type II errors* on the network structure, i.e. false positive and false negative conclusions about links, is studied. In Sec. 4.2.2 different methods to solve the *Inverse Problem* are presented. Section 4.2.3 contains a brief description of the generalization to directed networks.

4.2.1 Networks change

In this section, only undirected networks are considered. Later (Sec. 4.2.3) a generalization to directed networks is presented.

Consider a network G with n nodes and vertex degree distribution defined by the probability function $\mathcal{P}$, i.e., $\mathcal{P}_i = \mathbb{P}(d = i)$ is the probability that the degree d is i, for $i = 0, \cdots, n-1$, see Sec. 2.1. Note that the degree of a vertex is between 0 and $n-1$, since each vertex can be connected to at most $n-1$ remaining vertices.

The focus of this section is to study the influence of *type I* and *type II errors* on the vertex degree distribution of a given network G. Let G' be the network detected when *type I* and *type II errors* occur. Therefore, α expresses the probability that a link absent in G is present in G' and β is the probability that a link present in G is no longer present in G'. Hence, the set of edges of G' is a combination of *true positive links* and *false positive links* of G. The vertex degree distribution of G' is characterised by the probability function $\mathcal{P}'$.

Consider a vertex and assume it has degree k, therefore there are k links connected to it and $n-1-k$ absent links. The aim is to evaluate the probability that this vertex has vertex degree k' in G'. The vertex degree

$$k' = j + i \tag{4.1}$$

is given by the sum of *true positive links* j and *false positive links* i; additionally, i and j have to satisfy

$$j \leq k \qquad \text{and} \tag{4.2a}$$

$$i \leq n-1-k. \tag{4.2b}$$

The condition described by Eq. (4.2a) guarantees that the number of false negative links is larger or equal than zero, and smaller or equal than the number of the original true positive links, i.e., $0 \leq k-j \leq k$. Likewise, the number of false positive links must be non-negative and smaller or equal than the number of the original non-present links, Eq. (4.2b).

The probability that a vertex has degree k' in G', knowing it has degree k in G is

$$\mathbb{P}(d'=k'|d=k) = \begin{cases} \displaystyle\sum_{i=0}^{k'} \binom{k}{k'-i}(1-\beta)^{k'-i}\beta^{k-k'+i}\binom{n-1-k}{i}\alpha^i(1-\alpha)^{n-1-k-i} \\ \qquad\qquad \text{if } k' \leq k \text{ and } k' \leq n-1-k \\ \displaystyle\sum_{i=0}^{k} \binom{k}{i}(1-\beta)^{i}\beta^{k-i}\binom{n-1-k}{k'-i}\alpha^{k'-i}(1-\alpha)^{n-1-k-k'+i} \\ \qquad\qquad \text{if } k < k' \leq n-1-k \\ \displaystyle\sum_{i=0}^{n-1-k'} \binom{k}{k-i}(1-\beta)^{k-i}\beta^{i}\binom{n-1-k}{k'-k+i}\alpha^{k'-k+i}(1-\alpha)^{n-1-k'-i} \\ \qquad\qquad \text{if } k' \geq k \text{ and } k' > n-1-k \\ \displaystyle\sum_{i=0}^{n-1-k} \binom{k}{k'-i}(1-\beta)^{k'-i}\beta^{k-k'+i}\binom{n-1-k}{i}\alpha^{i}(1-\alpha)^{n-1-k-i} \\ \qquad\qquad \text{if } n-1-k < k' < k. \end{cases} \tag{4.3}$$

The probability $\mathbb{P}(d'=k'|d=k)$ is a piecewise function for all combinations of i and j satisfying Eqs. (4.1) and (4.2). To obtain Eq. (4.3) consider,

as an example, the first case, i.e., $k' \leq k$ and $k' \leq n-1-k$.

The probability of having j *true positive links*, over all possible k original true positive links, is

$$\mathbb{P}(\text{no. true positive links} = j) = \binom{k}{j}(1-\beta)^j \beta^{k-j}, \tag{4.4}$$

which is the binomial distribution $\mathcal{B}(k, 1-\beta)$. Since $j = k' - i$ [Eq. (4.1)], Eq. 4.4 corresponds to the first part of the first case of Eq. 4.3. Similarly, the probability of having i *false positive links* is

$$\mathbb{P}(\text{no. false pos links} = i) = \binom{n-1-k}{i}\alpha^i(1-\alpha)^{n-1-k-i}, \tag{4.5}$$

which is the binomial distribution $\mathcal{B}(n-1-k, \alpha)$.

The first case of Eq. (4.3) can be obtained combining Eqs. (4.4) and (4.5), changing variable j according to Eq. (4.1), and considering all possible combinations of i and j. All the other cases can be derived in the same way following the conditions in Eq. (4.2). A more detailed argumentation can be found in Sec. 5.2.1.

The law of total probability

$$\mathbb{P}(d' = k') = \sum_{k=0}^{n-1} \mathbb{P}(d' = k' | d = k)\mathbb{P}(d = k) \tag{4.6}$$

for $k' \in \{0, \cdots, n-1\}$, is applied to obtain the matrix equation

$$\underbrace{\begin{bmatrix} \mathbb{P}(d'=0) \\ \vdots \\ \mathbb{P}(d'=n-1) \end{bmatrix}}_{=\mathcal{P}'} = \underbrace{\begin{bmatrix} \mathbb{P}(d'=0|d=0) & \cdots & \mathbb{P}(d'=0|d=n-1) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(d'=n-1|d=0) & \cdots & \mathbb{P}(d'=n-1|d=n-1) \end{bmatrix}}_{=A} \cdot \underbrace{\begin{bmatrix} \mathbb{P}(d=0) \\ \vdots \\ \mathbb{P}(d=n-1) \end{bmatrix}}_{=\mathcal{P}}, \tag{4.7}$$

i.e.,

$$\mathcal{P}' = A\mathcal{P}. \tag{4.8}$$

The matrix $A = A(n, \alpha, \beta)$ depends on n, α and β and has determinant

$$\det A = (1 - \alpha - \beta)^{\frac{n(n-1)}{2}}, \tag{4.9}$$

therefore it is invertible if and only if $\alpha \neq 1 - \beta$, see Sec. 5.4 for a proof. A complete analysis on the matrix A is presented in Chapter 5.

Assuming G is known, Eq. (4.8) characterises the influenced of *type I* and *type II errors* on the vertex degree distribution, and it allows to find the vertex degree distribution of the network G'.

4.2.2 Inference of networks' vertex degree distribution

Section 4.2.1 analyses the impact of *type I* and *type II errors* on the vertex degree distribution of a given network. Equation (4.8) allows to obtain $\mathcal{P}'$ from $\mathcal{P}$. This section focuses on the inverse problem, i.e., inverting Eq. (4.8), to infer the original vertex degree distribution from an observed one. When $\{\alpha, \beta\} \neq \{0, 0\}, \{1, 1\}$, since the convergence to zero of the determinant of A scales like $x^{\frac{n(n-1)}{2}}$ for $|x| < 1$ [Eq. (4.9)], numerical issues arise for relatively small n when inverting the matrix A to find $\mathcal{P}$ through $\mathcal{P} = A^{-1}\mathcal{P}'$. The cases for $\alpha, \beta = 0$ and $\alpha, \beta = 1$ are discussed in Sec. 5.3.

The *least squares method* is a standard approach to solve problems like Eq. (4.8). Although the matrix A is not singular, for reasonable parameter values for n, A is typically ill-conditioned, therefore the pseudoinverse of the truncated singular value decomposition of A is used.

The *singular value decomposition* of a matrix A is the factorization of the matrix into the product of $A = UWV^T$ where W is a diagonal matrix and the columns of the matrices U and V are orthonormal [67]. The elements $w_1, \cdots, w_n$ on the diagonal of W are called *singular values* of A and they are ordered such that $w_1 \geq w_2 \geq \cdots \geq w_r > w_{r+1} = \cdots = w_n = 0$, where r is the rank of A.

The singular value decomposition is a tool to compute the pseudoinverse of a matrix. If A has singular value decomposition $A = UWV^T$, its pseudoinverse A^+ is defined as $A^+ = VW^+U^T$, where W^+ is obtained from W replacing all the non-zero elements with their reciprocals.

The *truncated singular value decomposition* is a method for regularization of ill-posed least squares problems [32]. Once the singular value decomposition $A = UWV^T$ is found, the matrix W is truncated at, e.g., rank t such that only the first t *singular values* are considered; this matrix is usually called W_t. More precisely, W_t is a diagonal matrix with elements $w_1 \geq w_2 \geq \cdots \geq w_t > w_{t+1} = \cdots = w_n = 0$, with $t < r$. The truncated diagonal matrix W_t is used to find an approximation of the matrix A using its decomposition, i.e., $A_t = UW_tV^T$. The optimal value for t has been studied in [29, 30]. The matrix A_t is the closest approximation of A of rank t, [32]. The matrix W_t is used to calculate the pseudoinverse of A_t, i.e., $A_t^+ = VW_t^+U^T$, and therefore to solve Eq. (4.8), namely

$$\mathcal{P} = A_t^+ \, \mathcal{P}'. \tag{4.10}$$

4.2.3 Generalization for directed networks

For directed networks the vertex degree is characterised by the *vertex in-degree* and the *vertex out-degree* [48]. Usually, in a directed network the *vertex degree* is the sum of the vertex in-degree and the vertex out-degree (Sec. 2.2).

Both the in-degree and the out-degree of a vertex are numbers between 0 and $n-1$, if n is the number of vertices of the network (Sec. 2.2.1). Therefore, the analysis shown in Secs.4.2.1-4.2.2 remains valid if either the vertex in-degree or the vertex out-degree are considered instead of the vertex degree.

An undirected network with n nodes has at most $n(n-1)/2$ edges. A network with n nodes has at most $n(n-1)$ directed edges; a generalization for other characteristics is likely more complicated, and therefore requires a more in-depth analysis.

4.3 Simulation study

To demonstrate the abilities as well as limitations, the analysis presented in Sec. 4.2 is applied to some typical simulated networks. Note that this ap-

proach is derived analytically; simulation studies are predominantly needed to demonstrate its applicability in real-world examples and to check for numerical issues, etc. There might be practical issues, e.g., due to the dimension of the network, and with the aim to show how these challenges can be overcome, a simulation study is presented to explore the concrete applicability of this method.

Five network topologies that present different characteristics are presented so to have a spectrum of networks as wide as possible to which apply the analysis. Namely, consider Erdős-Rényi, Small-World, Scale-Free networks, a three-dimensional grid, and a network of randomly connected communities [48]. The probabilities α and β of *type I* and *type II errors* vary in the range $1\% - 10\%$ mimicking a typical analysis method that has high sensitivity and high specificity. Nevertheless, both lower and higher values for α and β can be chosen and the results obtained are qualitatively the same as the ones presented below.

Consider an Erdős-Rényi network G with 100 nodes and a probability of a connection of 0.2. The vertex degree has binomial distribution $\mathcal{B}(100, 0.2)$. Adding and removing links with probabilities $\alpha = 0.05$ and $\beta = 0.03$ results in a new network G'. The vertex degree distribution of G' is calculated empirically by counting the vertices' degrees. Applying the procedure explained above, the vertex degree distribution of the original network is estimated. Figure 4.1 shows the results using the cut-off for the truncated singular value decomposition method of 0.5, i.e., W_t contains only singular values greater than 0.5. The choice of t is motivated by smoothness and regularity of the solution obtained.

Figure 4.1 shows the histogram of the degrees of the vertices of the original network G, the density of the detected network G', the reconstructed vertex degree distribution of the original network $\mathcal{P}$ resulting from Eq. (4.10), and the result when a non-negative constraint is applied to the truncated singular value decomposition to avoid that numerical issues result in negative solutions. More precisely, the lsqnonlin Matlab function is used with lower bound condition $lb = \mathsf{zeros}(n)$; this function implements the *trust region reflective*

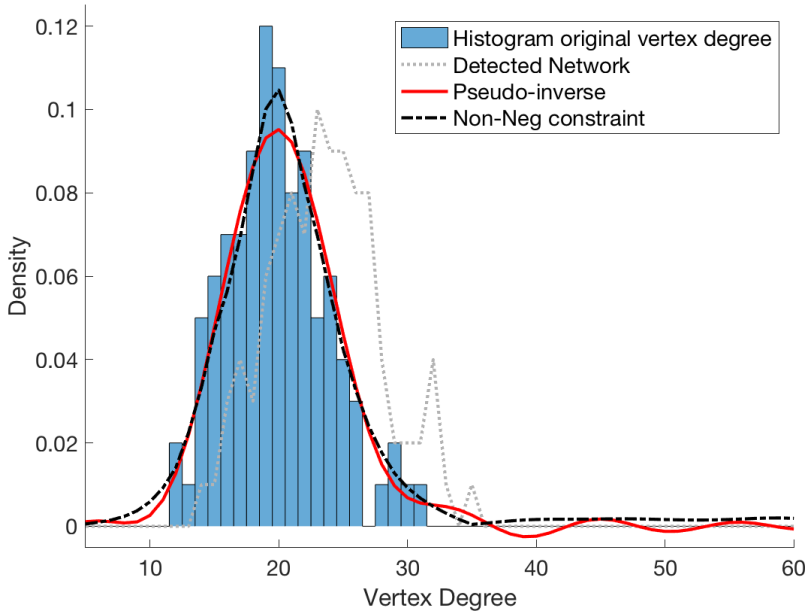


Figure 4.1: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, the result of network reconstruction using Eq. (4.10) knowing A and $\mathcal{P}'$, solid red line, and the result when a non-negative constraint is applied to the truncated singular value decomposition, black dashed line. The original network is an Erdős-Rényi network with 100 nodes and probability of connection 0.2.

algorithm [11, 20]. The density of G' is estimated by

$$\mathcal{P}'_i = \frac{\text{number of nodes with vertex degree} = i}{\text{number of nodes}} \tag{4.11}$$

its empirical distribution, and this is used to infer the original network.

Figure 4.2 shows the result when the original network G is a Small-World network. It is built from the regular network of 100 nodes, vertex degree 4, and probability of rewiring 0.4. The network G' is obtained by adding and removing links at random with probabilities $\alpha = 0.03$ and $\beta = 0.05$ respectively. The cut-off for the truncated singular value decomposition method is 0.33.

Figure 4.2 shows the histogram of the degrees of the vertices of the original network G, the density of the detected network G', the reconstructed vertex

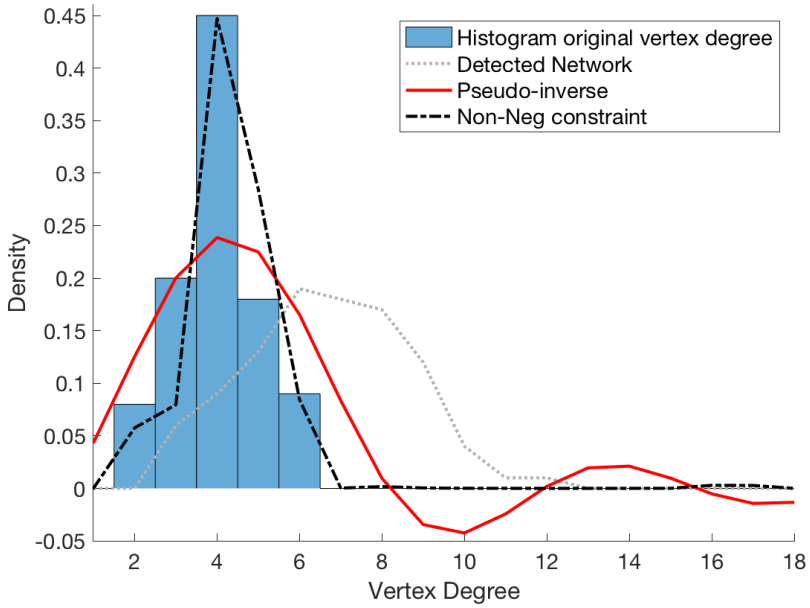


Figure 4.2: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, the result of network reconstruction using Eq. (4.10) knowing A and $\mathcal{P}'$, solid red line, and the result when a non-negative constraint is applied to the truncated singular value decomposition, black dashed line. The original network is a Small World network with 100 nodes and probability of rewiring 0.4.

degree distribution, and the result when a non-negative constraint is applied to the truncated singular value decomposition.

Figure 4.3 shows the result when the original network G is a Scale-Free network. It is built using a preferential attachment model for network growth. At each step a vertex, with a link attached to it, is added. The probability that the new vertex attaches to a given old one is proportional to its vertex degree. This procedure is repeated until the network has 100 nodes. The network G' is obtained by adding and removing links at random with probabilities $\alpha = 0.1$ and $\beta = 0.03$ respectively. The cut-off for the truncated singular value decomposition method is 0.4.

Figure 4.3 shows the histogram of the degrees of the vertices of the original network G, the density of the detected network G', the solution of Eq. (4.10), and the result when a non-negative constraint is applied to the truncated

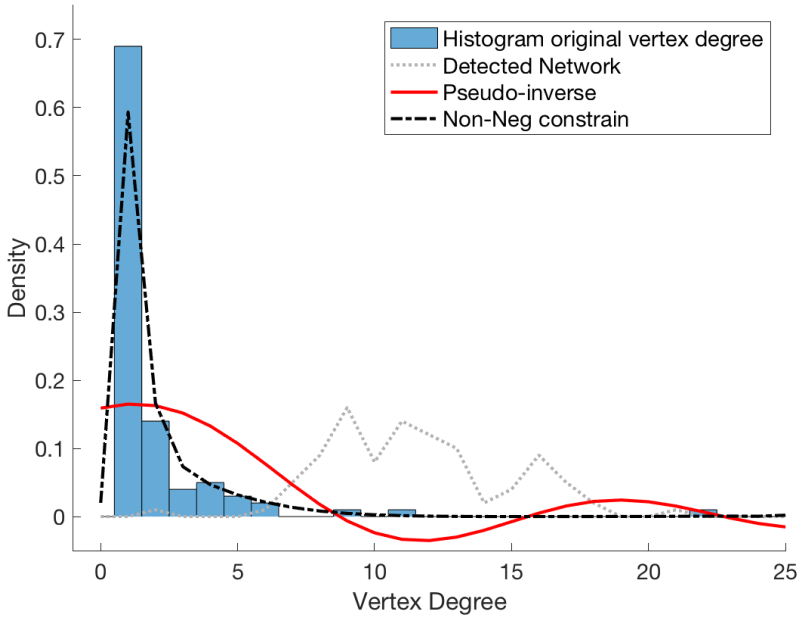


Figure 4.3: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, the result of network reconstruction using Eq.(4.10) knowing A and $\mathcal{P}'$, solid red line, and the result when a non-negative constraint is applied to the truncated singular value decomposition, black dashed line. The original network is a Scale-Free network with 100 nodes.

singular value decomposition.

The method presented in this chapter is now applied to another example - the original network G is a three-dimensional grid $4 \times 5 \times 5$; note that G has 100 nodes. The network G' is obtained by adding and removing links at random with probabilities $\alpha = 0.1$ and $\beta = 0.05$ respectively. The cut-off for the truncated singular value decomposition method is 0.38. Figure 4.4 shows the histogram of the degrees of the vertices of the original network G, the density of the detected network G', the solution of Eq. (4.10), and the result when a non-negative constraint is applied to the truncated singular value decomposition.

The last example presented is the case when G is a network of three randomly connected communities. It is built by constructing three Erdős-

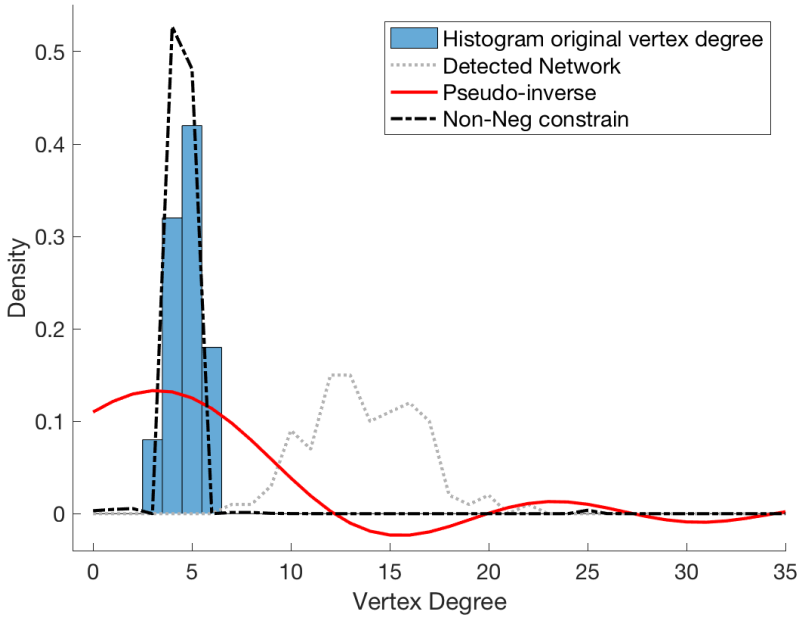


Figure 4.4: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, the result of network reconstruction using Eq.(4.10) knowing A and $\mathcal{P}'$, solid red line, and the result when a non-negative constraint is applied to the truncated singular value decomposition, black dashed line. The original network is a $4 \times 5 \times 5$ grid.

Rényi networks, with probability of connection 0.3, 0.6, and 0.9, and each with 33 nodes. Then, nodes from different communities are connected with probability 0.1. The network G' is obtained by adding and removing links at random with probabilities $\alpha = 0.05$ and $\beta = 0.03$ respectively. The cut-off for the truncated singular value decomposition method is 0.42. Figure 4.5 shows the histogram of the degrees of the vertices of the original network G, the density of the detected network G', the solution of Eq. (4.10), and the result when a non-negative constraint is applied to the truncated singular value decomposition.

Another interesting aspect is the influence of *type I* and *type II errors* and the proposed method on the reconstruction of individual nodes and not just the correct distribution. This is particularly relevant for nodes that have

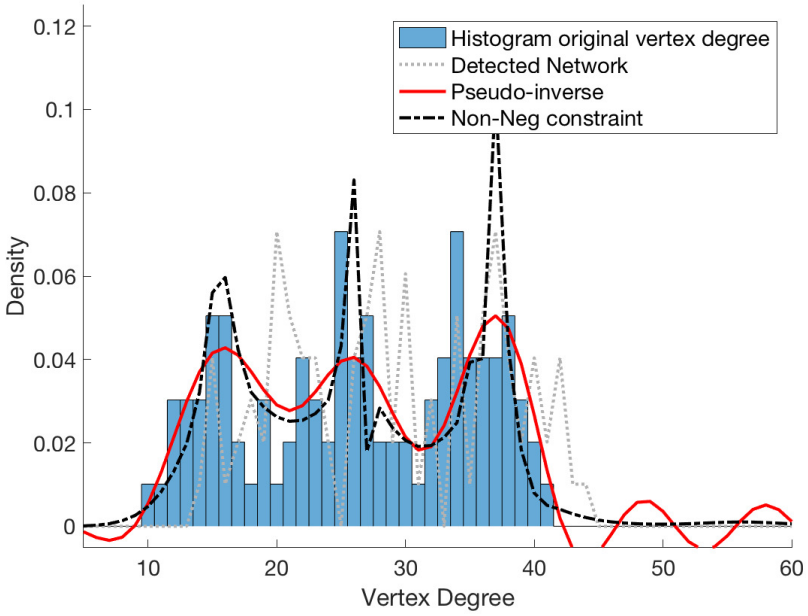


Figure 4.5: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, the result of network reconstruction using Eq.(4.10) knowing A and $\mathcal{P}'$, solid red line, and the result when a non-negative constraint is applied to the truncated singular value decomposition, black dashed line. A network of three randomly connected communities is used as original network.

a degree much larger than average, so-called hubs.

In the Scale-Free example, Fig. 4.3, the detected distribution appears to be smoother than the original, implying that a hub might have been converted to a non-hub. Analysing this in more detail, there is convincing evidence that this is not the case - hubs are correctly identified as hubs.

Consider a node d that has degree k in G that has n nodes. Due to *type I* and *type II errors*, this node in G' has degree d', a random variable with distribution shown in Eq. (4.3). Taking realisations of this random variable, and inverting the process using Eq. (4.10), allows us to compare individual degrees for a given node of the true network with the reconstructed one. Consider a network with $n = 100$ nodes, a node d with degree $k = 75$, probabilities of *type I* and *type II errors* of $\alpha = 0.05$ and $\beta = 0.03$,

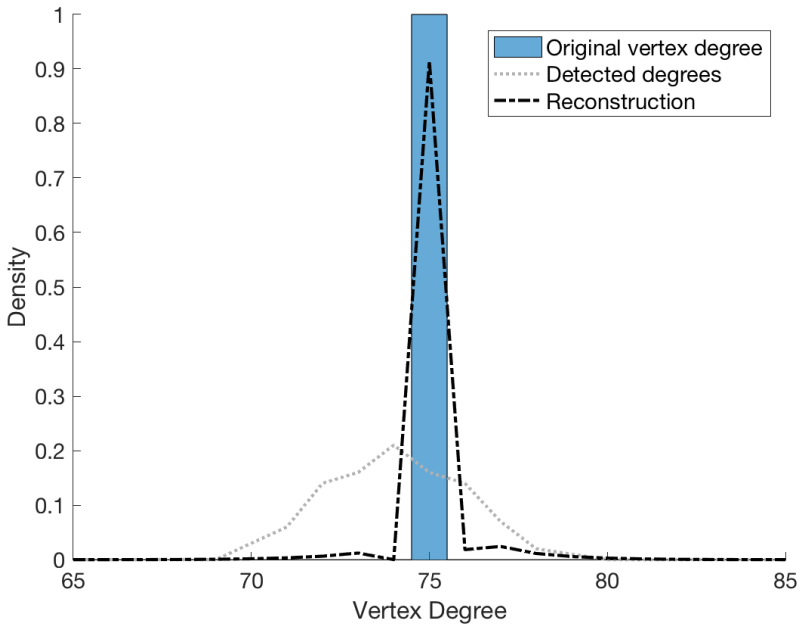


Figure 4.6: Reconstruction of the degree of a single node with original degree $k = 75$.

respectively, and simulate 100 realisations of the random variable described above. Figure 4.6 shows the reconstruction of the degree of d using these realisations. The result does not only show an improvement from the detected degrees k, but also illustrates the high accuracy of the reconstruction method.

Figure 4.7 shows the reconstruction of various degrees, i.e., k from 10 to 90 in steps of 10, using the same parameters $n = 100$, $\alpha = 0.05$, $\beta = 0.03$, and 100 realisations each. This again demonstrates that the method reliably reconstructs the correct degree for this individual node. Further simulations, not presented here, varying α and β between 0.01 and 0.1, show qualitatively the same results. In every case, the reconstruction is very robust, and this suggests that it is extremely unlikely that a hub is reconstructed as a non-hub. Moreover, the reconstruction works correctly not only on the general distribution, but also when it is applied to single nodes.

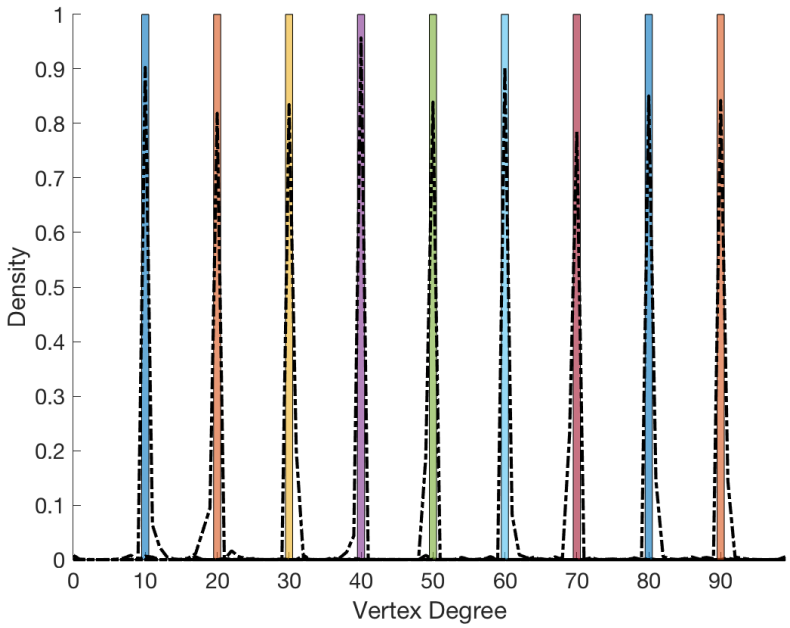


Figure 4.7: Reconstructions of the degree of single nodes with original degrees k from 10 to 90 in steps of 10.

4.4 Robustness of reconstruction

As stated above, the reconstruction method assumes the probabilities of *type I* and *type II errors* to be known a priori. While the *type I error* is controlled by statistical methods, the *type II error* must be inferred or reasonable assumptions from simulations, or prior studies, about the *type II error* must be available. To show the impact of violations of this and thereby the robustness of this method, the performance of the reconstruction is analysed when perturbations on α and β are introduced.

Figures 4.8-4.11 demonstrate the robustness of the approach for various examples. Figures 4.8-4.9 are used to show robustness with respect to β, while Figs. 4.10-4.11 show the robustness with respect to α. The perturbations are quantified in percentage using the parameter δ, e.g. the perturbations of β are expressed by $\beta+\delta\beta$. Note that, since $0 \leq \beta \leq 1$, the conditions for the perturbations are $-1 \leq \delta \leq 1/\beta - 1$; namely, called $\beta^p = \beta + \delta\beta$ the

perturbed β, then the conditions can be derived

$$\begin{aligned} 0 \leq & \beta^p \leq 1 \\ 0 \leq & \beta + \delta\beta \leq 1 \\ -1 \leq & \delta \leq 1/\beta - 1. \end{aligned} \tag{4.12}$$

Negative values for δ represent underestimated values for β and positive overestimated values for β. The same argument is used for the perturbation of α.

Figure 4.8 shows the reconstruction of a Scale-Free network with 100 nodes for the true value of $\beta = 0.03$, and also for various values of β deviating up to 1000% from the true value, i.e., $\beta^p = 0.33$. The cut-off for the truncated singular value decomposition method is 0.1 and the probability of *type I error* is $\alpha = 0.05$, assuming to control the family-wise error rate at this value, i.e., the probability of making at least one *type I error*; it is beyond the scope of this work to discuss cases in which the technique selected to reconstruct the network violates this assumption - however, below the results for different deviations from the true α are used to generate the plots to investigate its robustness. Figure 4.8 shows that this approach is robust to rather large perturbations of β, in both negative and positive directions. Up to $\delta = 500\%$, the bias of the reconstruction is negligible; only if $\delta = 1000\%$ or more deviates the reconstruction significantly from the true one, although it still performs better than the naïve approach of trusting the identified network structure.

Figure 4.9 shows the reconstruction of an Erdős-Rényi network with 100 nodes and probability of a connection of 0.2 for the true value of $\beta = 0.03$, and also for various values of β deviating up to $\delta = 400\%$ from the true value of β. The cut-off for the truncated singular value decomposition method is 0.55 and the probability of *type I error* is $\alpha = 0.05$. Also in this case, the method is robust to large perturbations of β, in both negative and positive directions. A deviation of more than 400% is needed for the method to fail and not to have an improvement over the naïve approach.

Figure 4.10 shows the reconstruction of an Erdős-Rényi network with 100 nodes and probability of a connection of 0.2 for the true value of $\alpha = 0.05$,

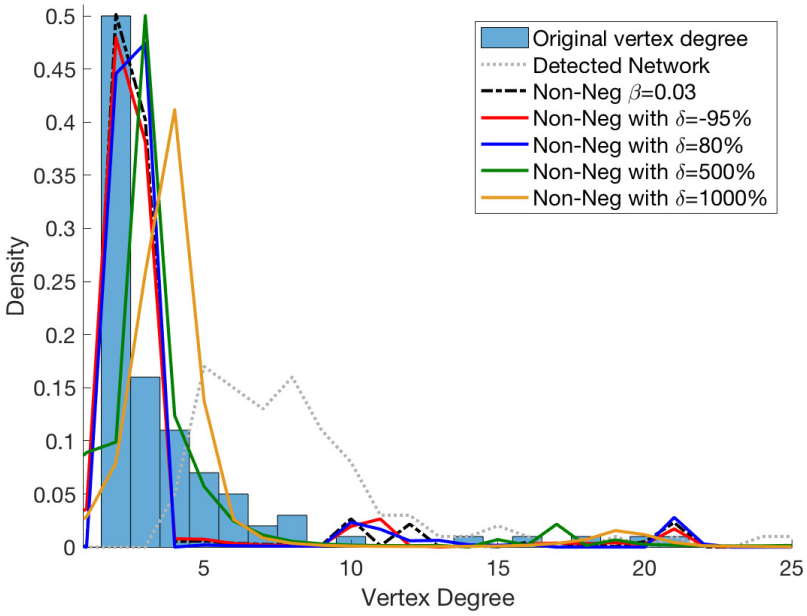


Figure 4.8: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, and the results when a non-negative constraint is applied to the truncated singular value decomposition using the true β, black dashed line, and perturbations from the true β, red, blue, yellow, and green solid lines. The original network is a Scale-Free network with 100 nodes.

and α deviating up to -85%. The cut-off for the truncated singular value decomposition method is 0.6 and the probability of *type II error* is $\beta = 0.03$. Figure 4.10 shows that the method is affected by relatively large perturbations of α. Namely, for $\delta < -85\%$ and $\delta > 50\%$, the reconstructions deviate significantly from the true one. The reason is that sparse networks are susceptible to perturbation of *type I error*. Figure 4.11 shows the reconstruction of a denser network, i.e., an Erdős-Rényi network with probability of a connection of 0.8, for the same true values of α and β. In this case, a deviation of 150% or more is needed for the method to fail. The comparison of Figs. 4.10 and 4.11 leads to the conclusion that dense networks are more robust to perturbations of *type I error* than sparse networks. This is intuitively motivated by the fact that the *type I error* affects links that are not present in

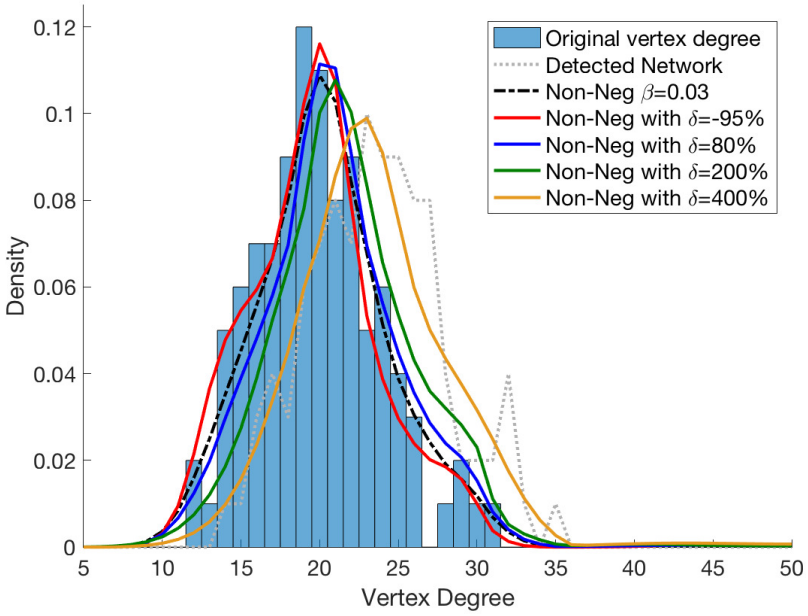


Figure 4.9: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, and the results when a non-negative constraint is applied to the truncated singular value decomposition using the true β, black dashed line, and perturbations from the true β, red, blue, yellow, and green solid lines. The original network is an Erdős-Rényi network with 100 nodes and probability of connection 0.2.

the network, and therefore it has a bigger influence on a sparse network.

The above results demonstrated that a rough estimate for α and β is sufficient to get an accurate reconstruction; the method is robust to relatively large perturbations of these two errors. Rough estimates of these parameters are typically available from simulation studies or prior knowledge about the system. Note again that the role of α and β are different; α is often controlled and can be obtained from known statistics of the techniques under the null hypothesis; β is more difficult as the true alternative would need to be known. Given the above simulations, the algorithm is more robust with respect to β than α, which aligns with the different role of these two errors. As mentioned at the beginning of Sec. 4.3, α and β vary in the range $1\% - 10\%$. Choosing either lower or higher values for the true α and β does not affect the general

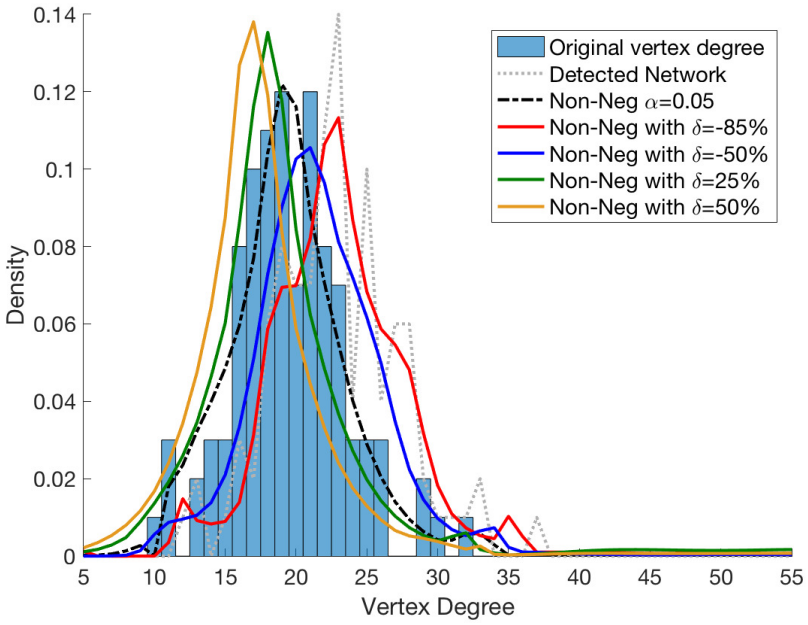


Figure 4.10: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, and the results when a non-negative constraint is applied to the truncated singular value decomposition using the true α, black dashed line, and perturbations from the true α, red, blue, yellow, and green solid lines. The original network is an Erdős-Rényi network with 100 nodes and probability of connection 0.2.

qualitatively result of the analysis, but it changes the range of perturbation that leads to the failure of the reconstruction method.

4.5 Conclusions

In this chapter, the impact of false positive and false negative conclusions about the presence or absence of links on the vertex degree distribution of a network is explored. Using an analytical approach, this dependence on the dimension of the network and the probabilities of *type I* and *type II errors* is investigated. Equation (4.8) describes the density of the vertex degree distribution of the biased network and thus allows to calculate the influence of false positive and false negative conclusions about links on any kind of

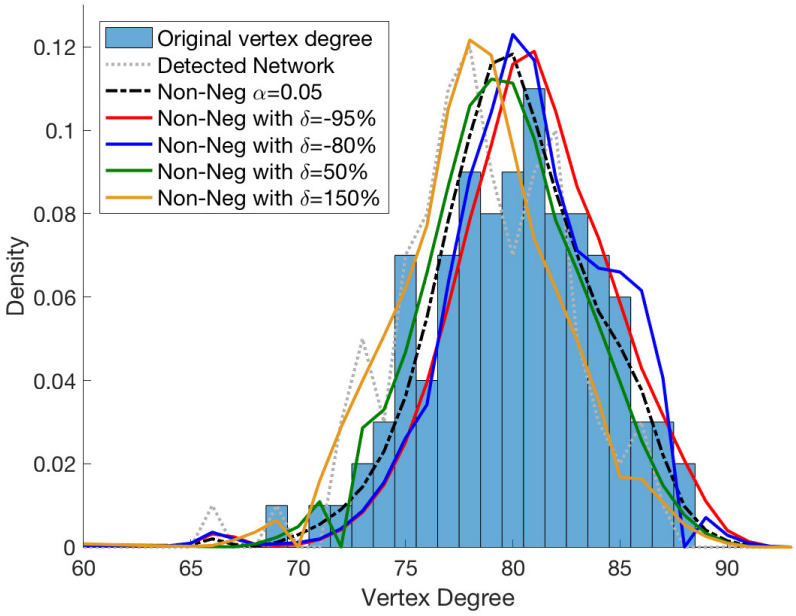


Figure 4.11: Density histogram of the original vertex degrees, blue bars, detected density vertex degree distribution, gray dotted line, and the results when a non-negative constraint is applied to the truncated singular value decomposition using the true α, black dashed line, and perturbations from the true α, red, blue, yellow, and green solid lines. The original network is an Erdős-Rényi network with 100 nodes and probability of connection 0.8.

network, assuming the probabilities of *type I* and *type II errors* are known.

In the inverse problem, the aim is to reconstruct the original network. Equation (4.10) enables to calculating analytically the vertex degree distribution of the original network if the biased one and the probabilities of *type I* and *type II errors* are given. When the dimension of the network is relatively large, numerical issues arise and consequently the truncated singular value decomposition is used to calculate the original network vertex degree distribution. Numerical simulations show that the vertex degree distribution is correctly recovered in all the cases discussed; the cases presented are designed to cover a variety of network topologies and therefore degree distributions.

The outcomes of this work are general results that enable to reconstruct analytically the vertex degree distribution of any network. The analytic for-

mula [Eq. (4.10)] that allows to find the original vertex degree distribution depends only on the detected vertex degree distribution and on the probabilities of *type I* and *type II errors*. This method is a powerful tool since the vertex degree distribution is a key characteristic of networks. Moreover, this method can be used to reconstruct individual node degrees to a very high accuracy. This should positively impact on various measures that can be derived from the networks. The proposed method should outperform standard approaches in terms of betweenness centrality, identification of hubs, and other network characteristics. This should be rigorously assessed in future research.

A limitation of this work is the assumption that the probabilities of *type I* and *type II errors* are known a priori. Nevertheless, the method is robust to relatively large perturbations of these two errors. Therefore, wrong estimates of *type I* and *type II errors*, within certain bounds, do not cause the reconstruction of be rendered invalid. Note again that in application the *type I error* is typically controlled, while an estimate for the *type II error* can only be obtained through prior experiments/knowledge or simulation studies. As shown in various simulations, the reconstruction method is robust to considerable deviations in β, which supports the usefulness of the technique over and above providing deeper insights into the role of these errors in network reconstruction; the approach is promising for real-world applications. Note though that it is always advisable to utilise simulation studies to characterise the advantageous and limitations in a concrete application at hand. Further analyses should study possible statistical approaches to infer these parameters employing Bayesian approaches or simulation studies. It is recommended to perform the latter to get an estimate of the *type II error* in particular.

Future studies should investigate analytically the influence of *type I* and *type II errors* on other network characteristics, e.g. the number of edges, the global clustering coefficient, and the efficiency, as shown through simulation studies in Chapter 3. As a consequence, more information about the original network can be found and, therefore, combining them all a better reconstruction of the network can be achieved.

Chapter 5

Poisson-binomial distribution: the case of the sum of two binomial distributions

5.1 Introduction

Chapter 4 presents the influence of *type I* and *type II errors* about the presence or absence of links on the vertex degree distribution, when a network is reconstructed. As shown in Eq. (4.8), the density of the biased vertex degree distribution is found using the matrix operator $A = A(n, \alpha, \beta)$, whose elements are the conditional probabilities $\mathbb{P}(d' = k'|d = k)$ defined in Eq. (4.3). The aim of this chapter is to study the properties of this matrix.

An important remark here is that the distribution, described by each element of the matrix A, is a special case of a Poisson binomial distribution. The Poisson binomial distribution is the distribution of a sum of independent Bernoulli random variables that are not necessarily identically distributed. This probability distribution has been first introduced by Poisson in 1837, and lately, it has been widely investigated [13, 21, 34, 70]. In this chapter, a special case of this distribution is analysed; the random variable $d'|d = k$ consists of a sum of $n - 1$ Bernoulli trials, k of those have probability of success $1 - \beta$ each, and each of the remaining $n - 1 - k$ trials has probability

of success α.

In Sec. 4.2.1, the probabilities $\mathbb{P}(d' = k'|d = k)$ are derived and briefly discussed. Section 5.2 is dedicated to analyse such probabilities in a more theoretical framework; those probabilities are analysed and a more accurate explanation of their properties is given (Secs. 5.2.2 - 5.2.3 - 5.3). The validity of Eq. (4.9) for the determinant of matrix A is proven is Sec. 5.4, and other properties of the matrix are discussed in Sec. 5.5.

5.2 Probabilities and matrix A

Consider a network G with n nodes and vertex degree distribution defined by the probability function $\mathcal{P}$, as shown in Chapter 4.2.1. Call G' the network detected when *type I* and *type II errors* occur, assume that α is the probability of a *type I error* and β is the probability of a *type II error*. Therefore, α expresses the probability a non-existing link in G is present in G' and β is the probability that an existing link in G is no longer present in G'. Hence, the set of edges of G' is a combination of true positive links and false positive links of G. The vertex degree distribution of G' is characterised by the probability function $\mathcal{P}'$.

5.2.1 Probability $\mathbb{P}(d' = k'|d = k)$

Consider a vertex and assume it has degree k, i.e., there are k links connected to it and $n - 1 - k$ absent links. The aim is to evaluate the probability that this vertex has vertex degree k' in G'. Figure 5.1 represents, in terms of random variables, the steps of the process of obtaining k'.

The random variable Z describes the degree k of the considered vertex; note that Z takes values in $\{0, \cdots, n-1\}$. From the set of k original links, j true positive links are taken with probability

$$\mathbb{P}(Y_1 = j) = \binom{k}{j}(1-\beta)^j \beta^{k-j}, \tag{5.1}$$

whereas β is the probability of a *type II error*.

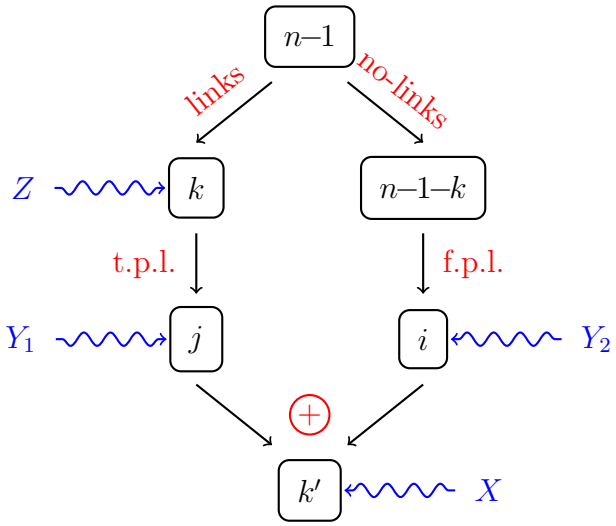


Figure 5.1: Schematic explanation of the construction of the probability $\mathbb{P}(d' = k'|d = k)$. In a network of n vertices, the vertex degree of a selected vertex is assumed to be k. From the existing k links, j true positive links (t.p.l.) are selected; from the remaining $n-1-k$ non-existing links, i false positive links (f.p.l.) are selected. The sum ($\oplus$) of j true positive links and i false positive links gives the vertex degree k'. Random variables are indicated in blue, and their realisations in black.

Similarly, since the probability to erroneously detect a link is α, and there are $n-1-k$ absent links, the probability of having i false positive links is

$$\mathbb{P}(Y_2 = i) = \binom{n-1-k}{i} \alpha^i (1-\alpha)^{n-1-k-i}. \tag{5.2}$$

Note that Eqs. (5.1) and (5.2) are the same as Eqs. (4.4) and (4.5), expressed in terms of random variables. The random variables Y_1 and Y_2 describe the number of true positive links and false positive links, respectively; they have distribution $Y_1 \sim \mathcal{B}(k, 1-\beta)$ and $Y_2 \sim \mathcal{B}(n-1-k, \alpha)$, following from Eqs. (5.1) and (5.2).

The number of true positive links cannot be larger than the number of originally existing links; likewise, the number of false positive links cannot exceed the number of originally non-existing links. Therefore, j and i have to satisfy $j \leq k$ and $i \leq n-1-k$, Eq. (4.2).

The vertex degree $k' = j + i$ is given by the sum of true positive links j

and false positive links i, Eq. (4.1), and it is described by the random variable X. Hence, $X = Y_1 + Y_2$ is the sum of two independent random variables. The random variable X takes value k' if $Y_2 = i$ and $Y_1 = k' - i$, for any $i \in \{0, \cdots, k'\}$ when $k' \leq k$ and $k' \leq n-1-k$, because of conditions $j \leq k$ and $i \leq n-1-k$ [Eq. (4.2)]. Therefore, the probability mass function of X given $Z = k$ is

$$\mathbb{P}(X{=}k'|Z{=}k) = \sum_{i=0}^{k'} \binom{k}{k'-i} (1-\beta)^{k'-i} \beta^{k-k'+i} \binom{n-1-k}{i} \alpha^i (1-\alpha)^{n-1-k-i} \quad (5.3)$$

if $k' \leq k$ and $k' \leq n-1-k$.

All the other cases, i.e., when $k' > k$ or/and $k' > n-1-k$, can be derived in the same way. The probability that a vertex has degree k' in G', knowing it has degree k in G is [Eq. (4.3)]

$$\mathbb{P}(X = k'|Z = k) = \begin{cases} \displaystyle\sum_{i=0}^{k'} \binom{k}{k'-i} (1-\beta)^{k'-i} \beta^{k-k'+i} \binom{n-1-k}{i} \alpha^i (1-\alpha)^{n-1-k-i} \\ \qquad\qquad \text{if } k' \leq k \text{ and } k' \leq n-1-k \\ \displaystyle\sum_{i=0}^{k} \binom{k}{i} (1-\beta)^{i} \beta^{k-i} \binom{n-1-k}{k'-i} \alpha^{k'-i} (1-\alpha)^{n-1-k-k'+i} \\ \qquad\qquad \text{if } k < k' \leq n-1-k \\ \displaystyle\sum_{i=0}^{n-1-k'} \binom{k}{k-i} (1-\beta)^{k-i} \beta^{i} \binom{n-1-k}{k'-k+i} \alpha^{k'-k+i} (1-\alpha)^{n-1-k'-i} \\ \qquad\qquad \text{if } k' \geq k \text{ and } k' > n-1-k \\ \displaystyle\sum_{i=0}^{n-1-k} \binom{k}{k'-i} (1-\beta)^{k'-i} \beta^{k-k'+i} \binom{n-1-k}{i} \alpha^i (1-\alpha)^{n-1-k-i} \\ \qquad\qquad \text{if } n-1-k < k' < k. \end{cases} \quad (5.4)$$

5.2.2 Compact Form

Equation (5.4) can be written in a more compact form

$$\mathbb{P}(d'{=}k'|d{=}k) = \sum_{i=\max\{0,k-k'\}}^{\min\{k,n-1-k'\}} \binom{k}{i} (1{-}\beta)^{k-i}\beta^{i} \binom{n{-}1{-}k}{k'{-}k{+}i} \alpha^{k'-k+i}(1{-}\alpha)^{n-1-k'-i}. \tag{5.5}$$

Equation (5.5) can be obtained from Eq. (5.4) considering through proper transformations. The first equation of Eq. (5.4) is valid for $k' \leq k$ and $k' \leq n-1-k$, therefore $\max\{0, k-k'\} = k-k'$ and $\min\{k, n-1-k'\} = k$. The substitution $x = k - k' + i$ of the sum increment i leads to Eq. (5.5), changing the name of the increment x back to i.

The second equation of Eq. (5.4) is valid for $k < k' \leq n-1-k$, therefore $\max\{0, k-k'\} = 0$ and $\min\{k, n-1-k'\} = k$. The substitution $x = k - i$ of the sum increment leads to Eq. (5.5).

The third equation of Eq. (5.4) is valid for $k' \geq k$ and $k' > n-1-k$, therefore $\max\{0, k-k'\} = 0$ and $\min\{k, n-1-k'\} = n-1-k'$. In this case there is no need of any substitution to obtain Eq. (5.5).

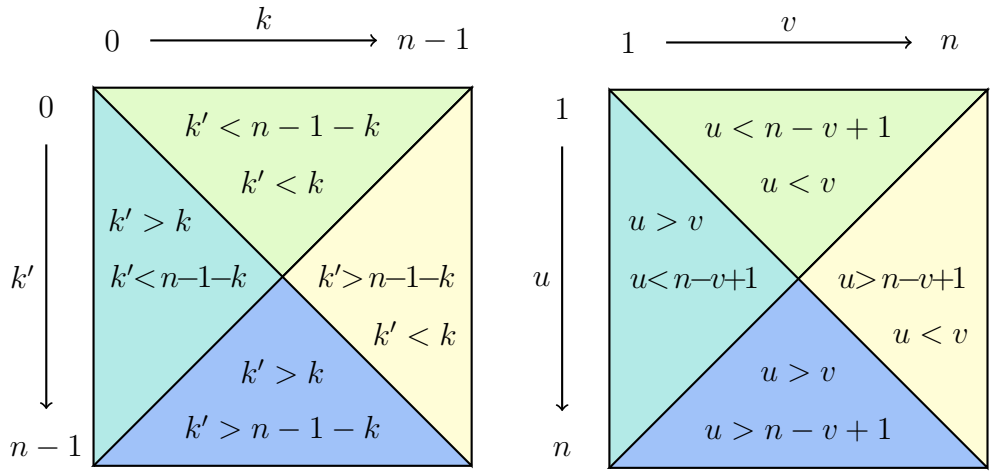
The fourth equation of Eq. (5.4) is valid for $n-1-k < k' < k$, therefore $\max\{0, k-k'\} = k-k'$ and $\min\{k, n-1-k'\} = n-1-k'$. The substitution $x = k - k' + i$ of the sum increment leads to Eq. (5.5).

5.2.3 Elements of matrix A

As shown in Chapter 4.2.1, the elements of the matrix A are the conditional probabilities defined by either Eq. (5.4) or (5.5). Since the element A_{uv} corresponds to the probability $\mathbb{P}(d' = k|d = k')$ for $u = k+1$ and $v = k'+1$, the matrix $A = A(n, \alpha, \beta)$ is an $n \times n$ matrix with elements

$$A_{uv} = \sum_{i=\max\{0,v-u\}}^{\min\{v-1,n-u\}} \binom{v-1}{i} (1-\beta)^{v-1-i}\beta^{i} \binom{n-v}{u-v+i} \alpha^{u-v+i}(1-\alpha)^{n-u-i}, \tag{5.6}$$

for $u, v \in \{1, \cdots, n\}$ and real numbers $\alpha, \beta \in [0, 1]$.



(a) Matrix A with elements defined by Eq. (5.4) or (5.5), and Eq. (4.7), using indices k', k.

(b) Matrix A with elements defined by Eq. (5.6), using indices u, v.

Figure 5.2: Block conditions of matrix A.

The cases $k' \lessgtr k$ and $k' \lessgtr n-1-k$ are not only the conditions for the piecewise function defined in Eq. (5.4), but they also define the lower and upper bound of summation in Eq. (5.5), and consequently of Eq. (5.6) with shifted variables u and v. Therefore, the matrix A is defined in four blocks as shown in Fig. 5.2, using both the notations k, k' and u, v. As demonstrated below in Sec. 5.3, the matrix A shows a structure with four blocks for extreme values of α and β, see Fig. 5.3.

5.3 Matrix A: special cases

In this section, special cases for the matrix A are discussed, i.e., when α and β take limit values of 0 and 1. Additionally, the case for $\beta = 1 - \alpha$ is also analysed, and it is the first to be examined.

The case for $\beta = 1 - \alpha$ is of key importance since it is the condition that makes the analysis in Chapter 4 impossible, in fact in this case the determinant of A is exactly zero, as shown below [Sec. 5.4]. Roughly, this happens because true and false positive links are indistinguishable. More precisely, refer to Fig. 5.1, when the random variables Y_1 and Y_2 have the

same probability of success, then it is not possible to reconstruct the random variable Z, since there is no distinction between the two sets made of k and $n-1-k$ links. Mathematically, when $\beta = 1-\alpha$, the Eq. (5.6) becomes

$$\begin{aligned} A_{uv}(n,\alpha,1{-}\alpha) &= \sum_{i=\max\{0,v-u\}}^{\min\{v,n-u\}} \binom{v{-}1}{i} \alpha^{v-1-i}(1{-}\alpha)^{i} \binom{n{-}v}{u{-}v{+}i} \alpha^{u-v+i}(1{-}\alpha)^{n-u-i} \\ &= \alpha^{u-1}(1-\alpha)^{n-u} \sum_{i=\max\{0,v-u\}}^{\min\{v,n-u\}} \binom{v-1}{i}\binom{n-v}{u-v+i} \\ &= \binom{n-1}{u-1}\alpha^{u-1}(1-\alpha)^{n-u} \quad . \end{aligned} \tag{5.7}$$

Note that $A_{uv}(n,\alpha,1-\alpha)$ does not depend on v but only on u, therefore each line has all identical elements, hence matrix A has structure

$$A(n,\alpha,1-\alpha) \simeq \begin{bmatrix} \bullet & \bullet & \bullet \\ \star & \star & \star \\ \circ & \circ & \circ \end{bmatrix}. \tag{5.8}$$

If α and β are both zero, then the random variables Y_1 and Y_2 take values k and 0 with probability 1, respectively; hence, $\mathbb{P}(X = k'|Z = k) = 1$ if $k' = k$, and 0 otherwise. Therefore, since $A_{uv} = \mathbb{P}(X = u+1|Z = v+1)$, the matrix

$$A(n,0,0) \simeq \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \tag{5.9}$$

is the identity.

If $\alpha, \beta = 1$, the random variables Y_1 and Y_2 take values 0 and $n-1-k$ with probability 1; hence, $\mathbb{P}(X = k'|Z = k) = 1$ if $k' = n-1-k$, and 0 otherwise. Therefore, the matrix

$$A(n,1,1) \simeq \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix}. \tag{5.10}$$

is anti-diagonal with all elements equal to one.

If $\alpha = 0$, then the random variable Y_2 takes value 0 with probability 1; hence, the random variable X is identical to Y_1, i.e., $\mathbb{P}(X = k'|Z = k) = \binom{k}{k'}(1-\beta)^{k'}\beta^{k-k'}$ if $k' \leq k$, and 0 otherwise. This equation expresses the probability that k' true links are chosen from the original k links; this is in

agreement with the condition that the probability $\mathbb{P}(X = k'|Z = k)$ is zero for $k' > k$. Note that it can also be obtained by Eq. (5.5) using $i = k - k'$ for the sum increment. Since $A_{uv} = \mathbb{P}(X = u + 1|Z = v + 1)$, this equation is equivalent to

$$A_{uv}(n, 0, \beta) = \begin{cases} \binom{v-1}{u-1}(1-\beta)^{u-1}\beta^{v-u} & u \leq v \\ 0 & u > v. \end{cases} \tag{5.11}$$

Notice that matrix

$$A(n, 0, \beta) \simeq \begin{bmatrix} \bullet & \cdots & \\ & \ddots & \\ & & \bullet \end{bmatrix}. \tag{5.12}$$

is upper triangular.

If $\alpha = 1$, then the random variable Y_2 takes value $n-1-k$ with probability 1; hence, $\mathbb{P}(X = k'|Z = k) = \binom{k}{n-1-k'}(1-\beta)^{k-n+1+k'}\beta^{n-1-k'}$ if $k' \geq n-1-k$, and 0 otherwise. This equation expresses the probability that $k'-(n-1-k)$ links are chosen from the original k true links; since $n - 1 - k$ false positive links have already been chosen, the probability $\mathbb{P}(X = k'|Z = k)$ must be zero for $k' < n - 1 - k$. Note that it can also be obtained by Eq. (5.5) using $i = n - 1 - k'$ for the sum increment. Therefore, this equation is equivalent to

$$A_{uv}(n, 1, \beta) = \begin{cases} \binom{v-1}{n-u}(1-\beta)^{v-1-n+u}\beta^{n-u} & u \geq n - v + 1 \\ 0 & u < n - v + 1. \end{cases} \tag{5.13}$$

Note that the matrix A has structure

$$A(n, 1, \beta) \simeq \begin{bmatrix} & \cdots & \bullet \\ & \iddots & \\ \bullet & & \end{bmatrix}. \tag{5.14}$$

If $\beta = 0$, then the random variable Y_1 takes value k with probability 1; hence, $\mathbb{P}(X = k'|Z = k) = \binom{n-1-k}{k'-k}\alpha^{k'-k}(1-\alpha)^{n-1-k'}$ if $k' \geq k$, and 0 otherwise. This equation expresses the probability that $k' - k$ false positive links are chosen from the original $n - 1 - k$ non present links; since k true links are already present, the probability $\mathbb{P}(X = k'|Z = k)$ must be zero for $k' < k$. Note that it can also be obtained by Eq. (5.5) using $i = 0$ for the

sum increment. This equation is equivalent to

$$A_{uv}(n,\alpha,0) = \begin{cases} \binom{n-v}{u-v}\alpha^{u-v}(1-\alpha)^{n-u} & u \geq v \\ 0 & u < v. \end{cases} \tag{5.15}$$

Notice that matrix

$$A(n,\alpha,0) \simeq \begin{bmatrix} \bullet & & \\ \bullet & \ddots & \\ \bullet & \bullet & \bullet \end{bmatrix}. \tag{5.16}$$

is lower triangular.

If $\beta = 1$, then the random variable Y_1 takes value 0 with probability 1; hence, the random variable X is identical to Y_2, i.e., $\mathbb{P}(X = k'|Z = k) = \binom{n-1-k}{k'}\alpha^{k'}(1-\alpha)^{n-1-k-k'}$ if $k' \leq n-1-k$, and 0 otherwise. This equation expresses the probability that k' false positive links are chosen from the original $n-1-k$ non present links; therefore, for $k' > n-1-k$ this probability is zero, since the number of false positive links cannot be larger than the original $n-1-k$ non-present links. Note that it can also be obtained by Eq. (5.5) using $i = k$ for the sum increment. This equation is equivalent to

$$A_{uv}(n,\alpha,1) = \begin{cases} \binom{n-v}{u}\alpha^{u}(1-\alpha)^{n-u-v+1} & u \leq n-v+1 \\ 0 & u > n-v+1. \end{cases} \tag{5.17}$$

Notice that the matrix A has structure

$$A(n,\alpha,1) \simeq \begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \\ \bullet & & \end{bmatrix}. \tag{5.18}$$

Figure 5.3 summarises the results obtained in this section. For extreme values of α and β, matrix A has distinct structures. These results not only underline the role of α and β, as shown in this section, but they also help in analytical calculations, such as the ones for the determinant in Sec. 5.4.

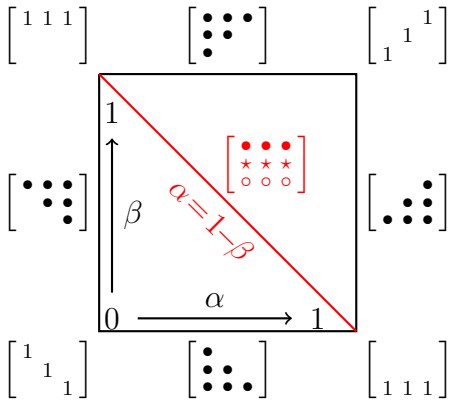


Figure 5.3: Matrix A: special cases. Matrix $A(n, \alpha, 0)$ middle below; $A(n, 1, \beta)$ middle right; $A(n, \alpha, 1)$ middle above; $A(n, 0, \beta)$ middle left; $A(n, 0, 0)$ bottom left; $A(n, 1, 0)$ bottom right; $A(n, 1, 1)$ top right; $A(n, 0, 1)$ top left; $A(n, \alpha, 1-\alpha)$ center.

5.4 Matrix A: determinant

This section is dedicated to prove that the matrix A has determinant

$$\det A = (1 - \alpha - \beta)^{\frac{n(n-1)}{2}}. \tag{5.19}$$

To prove this, some intermediate steps are needed. First, the limit cases for extreme values of α and β are analysed.

5.4.1 Limit cases

In this subsection, it is proven that the determinant of A satisfies Eq. (5.19), when $\beta = 1 - \alpha$ or $\alpha, \beta = 0, 1$, i.e., the special cases shown in Sec. 5.3.

If $\beta = 1 - \alpha$, Eqs. (5.7) and (5.8) show that each line of $A(n, \alpha, 1-\alpha)$ has all identical elements, i.e., each line is a multiple of vector $[1, \cdots, 1]$; hence, all the lines are linear depend and therefore the determinant is $\det A(n, \alpha, 1-\alpha) = 0$.

If $\alpha, \beta = 0$ then the matrix A is the identity and therefore the determinant is $\det A(n, 0, 0) = 1$, see Eq. (5.9). If $\alpha, \beta = 1$ then the matrix A is anti-diagonal with all elements equal to one, then the determinant is

$\det A(n,1,1) = (-1)^{\frac{n(n-1)}{2}}$, (Eq. (5.10)).

If $\alpha = 0, \beta \neq 0$, then the matrix $A(n,0,\beta)$ is upper triangular [Eqs. (5.11)-(5.12)] and therefore the determinant is the product of the elements on the diagonal $A_{uu}(n,0,\beta) = (1-\beta)^{u-1}$, i.e.,

$$\det A(n,0,\beta) = (1-\beta)^{\frac{n(n-1)}{2}}. \tag{5.20}$$

If $\alpha = 1, \beta \neq 0,1$, the matrix $A(n,1,\beta)$ has all zeros above the anti-diagonal [Eqs. (5.13)-(5.14)] and therefore the determinant is the product of the elements on the anti-diagonal $A_{uu}(n,1,\beta) = \beta^{u-1}$ and sign given by $(-1)^{\frac{n(n-1)}{2}}$, i.e.,

$$\det A(n,1,\beta) = (-\beta)^{\frac{n(n-1)}{2}}. \tag{5.21}$$

If $\beta = 0, \alpha \neq 0,1$, the matrix $A(n,\alpha,0)$ is lower triangular [Eqs. (5.15)-(5.16)] and therefore the determinant is the product of the elements on the diagonal $A_{uu}(n,\alpha,0) = (1-\alpha)^{n-u}$, i.e.,

$$\det A(n,\alpha,0) = (1-\alpha)^{\frac{n(n-1)}{2}}. \tag{5.22}$$

If $\beta = 1, \alpha \neq 0,1$, the matrix $A(n,\alpha,1)$ has all zeros below the anti-diagonal [Eqs. (5.17)-(5.18)] and therefore the determinant is the product of the elements on the anti-diagonal $A_{uu}(n,\alpha,0) = \alpha^{n-u}$ and sign given by $(-1)^{\frac{n(n-1)}{2}}$, i.e.,

$$\det A(n,\alpha,1) = (-\alpha)^{\frac{n(n-1)}{2}}. \tag{5.23}$$

This concludes the study of the determinant of A for the special cases presented in Sec. 5.3, proving that the determinant of A satisfies Eq. (5.19), when $\beta = 1-\alpha$ or $\alpha, \beta = 0,1$.

5.4.2 Transformations

To prove Eq. (5.19) for every dimension n of the matrix A, it is useful to describe it in terms of dimension $n-1$. In this way, a mathematical induction can be used as a technique for the proof. In this subsection, linear transformations to define A of dimension n in terms of dimension $n-1$ are

presented.

Future calculations become shorter if the transpose A^T of matrix A is considered. Considering A^T instead of A does not affect the calculation of the determinant since $\det A^T = \det A$.

Consider the matrix A, defined by Eq. (5.6), and call A^{T_n} the transpose of A of dimension n. The cases for $\alpha, \beta = 0, 1$ or $\beta = 1 - \alpha$ have already been discussed in Sec. 5.4.1; here, the case for $0 < \alpha, \beta < 1$ and $\beta \neq 1 - \alpha$ is considered. Note that Call $\overline{A^{T_n}}$ the matrix with elements

$$\overline{a_{ij}^n} = \begin{cases} a_{ij}^n \dfrac{(1-\alpha-\beta)^{n-1}}{(1-\alpha)^{n-1}} & i = 1 \\ \left(a_{ij}^n - \dfrac{a_{i1}^n a_{1j}^n}{a_{11}^n} \right) \dfrac{1-\alpha}{1-\alpha-\beta} & i = 2 \\ \left(a_{ij}^n - \dfrac{\beta}{1-\alpha} a_{i-1,j}^n \right) \dfrac{1-\alpha}{1-\alpha-\beta} & i = 3, \cdots, n \end{cases} \tag{5.24}$$

where a_{ij}^n are the elements of the matrix A^{T_n}. The aim is to prove that

$$\overline{A^{T_n}} = \left[\begin{array}{c|c} (1-\alpha+\beta)^{n-1} & \\ \hline 0 & A^{T_{n-1}} \end{array} \right]. \tag{5.25}$$

Before proving this identity, the case for $n = 3$ is presented to fully understand the nature of the transformations in Eq. (5.24).

The transpose of the matrix A for $n = 3$ is

$$A^{T_3} = \begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ (1-\alpha)\beta & (1-\alpha)(1-\beta)+\alpha\beta & \alpha(1-\beta) \\ \beta^2 & 2(1-\beta)\beta & (1-\beta)^2 \end{bmatrix}. \tag{5.26}$$

The aim is to manipulate the matrix applying linear transformations so that its determinant is not altered. First, one step of Gauss elimination is performed; this gives all zeros in the first column, apart from the first

element. After such calculation matrix A^{T_3} becomes

$$A^{T_3} \rightsquigarrow \begin{bmatrix} (1-\alpha)^2 & 2(1-\alpha)\alpha & \alpha^2 \\ 0 & 1-\alpha-\beta & -\frac{\alpha\beta}{1-\alpha}+\alpha \\ 0 & \frac{2(1-\alpha-\beta)\beta}{1-\alpha} & \frac{(1-2\alpha)\beta^2}{(1-\alpha)^2}-2\beta+1 \end{bmatrix}. \tag{5.27}$$

Secondly, multiply the first row with $\frac{(1-\alpha-\beta)^2}{(1-\alpha)^2}$ and the other two rows with $\frac{1-\alpha}{1-\alpha-\beta}$; then sum to the last row, the second row times $\frac{\beta}{1-\alpha}$. Note that these operations do not change the determinant; the matrix becomes

$$\overline{A^{T_3}} = \begin{bmatrix} (1-\alpha-\beta)^2 & \frac{2\alpha(1-\alpha-\beta)^2}{1-\alpha} & \frac{\alpha^2(1-\alpha-\beta)^2}{(1-\alpha)^2} \\ 0 & 1-\alpha & \alpha \\ 0 & \beta & 1-\beta \end{bmatrix}. \tag{5.28}$$

Note that the bottom right 2×2 submatrix is identical to A^{T_2}, i.e.,

$$A^{T_2} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}. \tag{5.29}$$

To get to the transformations in Eq. (5.24), following the steps presented in the example of $n = 3$, the general case of dimension n is inferred.

The first step, i.e., the Gauss elimination, corresponds to the transformations

$$\widehat{a^n_{ij}} = \begin{cases} a^n_{ij} & i = 1 \\ a^n_{ij} - \frac{a^n_{i1}}{a^n_{11}} a^n_{1j} & i = 2, \cdots, n \end{cases}. \tag{5.30}$$

Multiplying the first row with $\frac{(1-\alpha-\beta)^{n-1}}{(1-\alpha)^{n-1}}$ and the other rows with $\frac{1-\alpha}{1-\alpha-\beta}$ corresponds to

$$\widehat{\widehat{a^n_{ij}}} = \begin{cases} \frac{(1-\alpha-\beta)^{n-1}}{(1-\alpha)^{n-1}} \widehat{a^n_{ij}} & i = 1 \\ \frac{1-\alpha}{1-\alpha-\beta} \widehat{a^n_{ij}} & i = 2, \cdots, n \end{cases} = \begin{cases} \frac{(1-\alpha-\beta)^{n-1}}{(1-\alpha)^{n-1}} a^n_{ij} & i = 1 \\ \frac{1-\alpha}{1-\alpha-\beta} \left(a^n_{ij} - \frac{a^n_{i1}}{a^n_{11}} a^n_{1j} \right) & i = 2, \cdots, n. \end{cases} \tag{5.31}$$

Finally, except the first two rows, sum to each row its previous row times $\frac{\beta}{1-\alpha}$ leading to

$$\overline{a_{ij}^n} = \begin{cases} \widehat{\widehat{a_{ij}^n}} & i = 1,2 \\ \widehat{\widehat{a_{ij}^n}} - \frac{\beta}{1-\alpha}\widehat{\widehat{a_{i-1,j}^n}} & i = 3,\cdots,n \end{cases} \tag{5.32a}$$

$$= \begin{cases} \frac{(1-\alpha-\beta)^{n-1}}{(1-\alpha)^{n-1}} a_{ij}^n & i = 1 \\ \frac{1-\alpha}{1-\alpha-\beta}\left(a_{ij}^n - \frac{a_{i1}^n}{a_{11}^n}a_{1j}^n\right) & i = 2 \\ \frac{1-\alpha}{1-\alpha-\beta}\left[\left(a_{ij}^n - \frac{a_{i1}^n}{a_{11}^n}a_{1j}^n\right) - \frac{\beta}{1-\alpha}\left(a_{i-1,j}^n - \frac{a_{i-1,1}^n}{a_{11}^n}a_{1j}^n\right)\right] & i = 3,\cdots,n \end{cases} \tag{5.32b}$$

$$= \begin{cases} \cdots & i = 1 \\ \cdots & i = 2 \\ \frac{1-\alpha}{1-\alpha-\beta}\left[a_{ij}^n - \frac{\beta}{1-\alpha}a_{i-1,j}^n - \frac{a_{1j}^n}{a_{11}^n}\left(a_{i1}^n - \frac{\beta}{1-\alpha}a_{i-1,1}^n\right)\right] & i = 3,\cdots,n \end{cases} \tag{5.32c}$$

$$= \begin{cases} \cdots \\ \cdots \\ \frac{1-\alpha}{1-\alpha-\beta}\left[a_{ij}^n - \frac{\beta}{1-\alpha}a_{i-1,j}^n - \frac{a_{1j}^n}{a_{11}^n}\left((1-\alpha)^{n-i}\beta^{i-1} - \frac{\beta}{1-\alpha}(1-\alpha)^{n-i+1}\beta^{i-2}\right)\right] \end{cases} \tag{5.32d}$$

$$= \begin{cases} a_{ij}^n \dfrac{(1-\alpha-\beta)^{n-1}}{(1-\alpha)^{n-1}} & i = 1 \\ \left(a_{ij}^n - \dfrac{a_{i1}^n a_{1j}^n}{a_{11}^n}\right)\dfrac{1-\alpha}{1-\alpha-\beta} & i = 2 \\ \left(a_{ij}^n - \dfrac{\beta}{1-\alpha}a_{i-1,j}^n\right)\dfrac{1-\alpha}{1-\alpha-\beta} & i = 3,\cdots,n. \end{cases} \tag{5.32e}$$

Note that Eq. (5.32e) is identical to Eq. (5.24).

To prove Eq. (5.25), the identity $\overline{a_{i+1,j+1}^n} = a_{ij}^{n-1}$ has to be verified, i.e.,

$$\left(a_{2j}^n - \frac{a_{21}^n a_{1j}^n}{a_{11}^n}\right)\frac{1-\alpha}{1-\alpha-\beta} = a_{1j}^{n-1} \tag{5.33}$$

and

$$\left(a_{i+1,j+1}^n - \frac{\beta}{1-\alpha}a_{i,j+1}^n\right)\frac{1-\alpha}{1-\alpha-\beta} = a_{ij}^{n-1} \qquad \text{for } i = 2,\cdots n-1 \tag{5.34}$$

are valid. To achieve this, the calculation is split into six cases, i.e.,

$$i = j, \quad \begin{cases} j > i \\ j \geq n - i - 1 \end{cases}, \begin{cases} j > i \\ j < n - i - 1 \end{cases},$$

$$\begin{cases} j < i \\ j > n - i \end{cases}, \begin{cases} j < i \\ j < n - i \end{cases}, \begin{cases} j < i \\ j = n - i \end{cases}. \tag{5.35}$$

For each of these conditions, the identity $\overline{a^n_{i+1,j+1}} = a^{n-1}_{ij}$ has been verified using analytic computations in *Wolfram Mathematica 11.2.0.0*. The following pages show the code, and respective results, of such computations. The notation used in the code is the following: f[n,α,β,k',k] corresponds to $\mathbb{P}(d'{=}k'|d{=}k)$ [Eq. (5.5)]; c refers to the column of the matrix, i.e., $k+1$ or v; r refers to the row of the matrix, i.e., $k'+1$ or u. Therefore, the notation f[n, α, β, j, i] corresponds to the element $a^n_{i+1,j+1}$.

Proof : $a^{n}_{i+1,j+1} - \beta / (1-\alpha)\, a^{n}_{i,j+1} = (1-\alpha-\beta) / (1-\alpha)\, a^{n-1}_{i,j}$

Definition

```
Clear["Global`*"]
```

```
f[n_, a_, b_, k1_, k_] := Sum[Binomial[k, i] (1 - b)^(k - i) b^i Binomial[n - 1 - k, k1 - k + i] a^(k1 - k + i) (1 - a)^(n - 1 - k1 - i), {i, Max[0, k - k1], Min[k, n - 1 - k1]}];
```

General assumptions

```
$Assumptions =
  n ∈ Integers && c ∈ Integers && r ∈ Integers && α ∈ Reals && β ∈ Reals && 0 < α < 1 && 0 < β < 1 && n > 3 && 1 ≤ r ≤ n - 1 && 1 ≤ c ≤ n - 1;
```

c = r

```
Assuming[c == r, FullSimplify[f[n, α, β, c, r] - f[n, α, β, c, r - 1] * β / (1 - α) - f[n - 1, α, β, c - 1, r - 1] * (1 - α - β) / (1 - α) == 0]]
```

```
True
```

c < r

```
(*c>n-r*)
Assuming[n - r < c < r,
 FullSimplify[f[n, α, β, c, r] - f[n, α, β, c, r - 1] * β / (1 - α) - f[n - 1, α, β, c - 1, r - 1] * (1 - α - β) / (1 - α) == 0]]
```

```
True
```

```
(*c<n-r*)
Assuming[c < r && c < n - r,
 FullSimplify[f[n, α, β, c, r] - f[n, α, β, c, r - 1] * β / (1 - α) - f[n - 1, α, β, c - 1, r - 1] * (1 - α - β) / (1 - α) == 0]]
```

True

```
(*c=n-r*)
Assuming[c < r && c == n - r,
 FullSimplify[f[n, α, β, c, r] - f[n, α, β, c, r - 1] * β / (1 - α) - f[n - 1, α, β, c - 1, r - 1] * (1 - α - β) / (1 - α) == 0]]
```

True

c > r

```
(*c>=n-r-1*)
Assuming[c > r, FullSimplify[f[n, α, β, c, r] - f[n, α, β, c, r - 1] * β / (1 - α) - f[n - 1, α, β, c - 1, r - 1] * (1 - α - β) / (1 - α)]]
```

$$\begin{cases} 0 & 1+c+r \geq n \\ -(1-\alpha)^{-2-c+n}\,\alpha^{c-r}\,(1-\beta)^{-1+r}\left(\alpha\,\beta\,\text{Binomial}[n-r,\,1+c-r]\,\text{Hypergeometric2F1}\left[1+c-n,\,1-r,\,2+c-r,\,\frac{\alpha\,\beta}{(-1+\alpha)\,(-1+\beta)}\right]+\right. & \text{True} \\ \quad \text{Binomial}[-1+n-r,\,c-r]\left(-(-1+\alpha+\beta)\,\text{Hypergeometric2F1}\left[1+c-n,\,1-r,\,1+c-r,\,\frac{\alpha\,\beta}{(-1+\alpha)\,(-1+\beta)}\right]+\right. & \\ \quad\quad \left.\left.(-1+\alpha+\beta-\alpha\,\beta)\,\text{Hypergeometric2F1}\left[1+c-n,\,-r,\,1+c-r,\,\frac{\alpha\,\beta}{(-1+\alpha)\,(-1+\beta)}\right]\right)\right) & \end{cases}$$

```
Assuming[1 + c + r ≥ n, FullSimplify[%16 == 0]]
```

True

```
(*c<n-r-1*)
```

```
Clear["Global`*"]
f[n_, a_, b_, k1_, k_] :=
  Piecewise[{{Sum[a^i (1 - b)^(k1 - i) (1 - a)^(-i - k + n - 1) b^(i + k - k1) Binomial[k, k1 - i] Binomial[-k + n - 1, i], {i, 0, k1}], k1 ≤ k && k1 ≤ -k + n - 1},
    {Sum[(1 - b)^i a^(k1 - i) b^(k - i) Binomial[k, i] (1 - a)^(i - k - k1 + n - 1) Binomial[-k + n - 1, k1 - i], {i, 0, k}], k < k1 ≤ -k + n - 1},
    {Sum[(1 - b)^(k - i) a^(k1 - k + i) b^i Binomial[k, k - i] (1 - a)^(n - 1 - k1 - i) Binomial[-k + n - 1, k1 - k + i], {i, 0, n - 1 - k1}], k1 ≥ k && k1 > -k + n - 1},
    {Sum[(1 - b)^(k1 - i) a^i b^(k - k1 + i) Binomial[k, k1 - i] (1 - a)^(n - 1 - k - i) Binomial[-k + n - 1, i], {i, 0, n - 1 - k}], n - 1 - k < k1 < k}}];

$Assumptions = n ∈ Integers && c ∈ Integers && α ∈ Reals && β ∈ Reals && r ∈ Integers && n > 5 && 1 ≤ r < c ≤ n - 1 && 0 < α < 1 && 0 < β < 1;

lineN[r_, c_] = FullSimplify[(f[n, α, β, c - 1, r - 1] - f[n, α, β, c - 1, 0] * f[n, α, β, 0, r - 1] / f[n, α, β, 0, 0]) * (1 - α) / (1 - α - β)];

lineNless1[r_, c_] = FullSimplify[f[n - 1, α, β, c - 1, r - 1]];

FullSimplify[lineN[r + 1, c + 1] - lineN[r, c + 1] * β / (1 - α) == lineNless1[r, c]]
```

$$\frac{1}{-1+\alpha+\beta}\left(\beta\left(\begin{cases}(1-\alpha)^{-c+n-r}\,\alpha^{c}\,\beta^{-1+r}\,\text{Binomial}[n-r,c]\,\text{Hypergeometric2F1}\left[-c,1-r,1-c+n-r,\frac{(-1+\alpha)\,(-1+\beta)}{\alpha\,\beta}\right] & c+r\le n\\ (1-\alpha)^{-1-c+n}\,\alpha^{1+c-r}\,(1-\beta)^{-1+r}\,\text{Binomial}[n-r,1+c-r]\,\text{Hypergeometric2F1}\left[1+c-n,1-r,2+c-r,\frac{\alpha\,\beta}{(-1+\alpha)\,(-1+\beta)}\right] & \text{True}\end{cases}\right)+\right.$$

$$\left.(-1+\alpha)\left(\begin{cases}(1-\alpha)^{-1-c+n-r}\,\alpha^{c}\,\beta^{r}\,\text{Binomial}[-1+n-r,c]\,\text{Hypergeometric2F1}\left[-c,-r,-c+n-r,\frac{(-1+\alpha)\,(-1+\beta)}{\alpha\,\beta}\right] & 1+c+r\le n\\ (1-\alpha)^{-1-c+n}\,\alpha^{c-r}\,(1-\beta)^{r}\,\text{Binomial}[-1+n-r,c-r]\,\text{Hypergeometric2F1}\left[1+c-n,-r,1+c-r,\frac{\alpha\,\beta}{(-1+\alpha)\,(-1+\beta)}\right] & \text{True}\end{cases}\right)\right) ==$$

$$\left(\begin{cases}(1-\alpha)^{-c+n-r}\,\alpha^{-1+c}\,\beta^{-1+r}\,\text{Binomial}[-1+n-r,-1+c]\,\text{Hypergeometric2F1}\left[1-c,1-r,1-c+n-r,\frac{(-1+\alpha)\,(-1+\beta)}{\alpha\,\beta}\right] & c+r\le n\\ (1-\alpha)^{-1-c+n}\,\alpha^{c-r}\,(1-\beta)^{-1+r}\,\text{Binomial}[-1+n-r,c-r]\,\text{Hypergeometric2F1}\left[1+c-n,1-r,1+c-r,\frac{\alpha\,\beta}{(-1+\alpha)\,(-1+\beta)}\right] & \text{True}\end{cases}\right)$$

```
Assuming[c < n - r - 1, FullSimplify[%6]]
```

True

5.4.3 Determinant: the proof

Using the results obtained so far in this chapter, it can now be proven that the $n \times n$ matrix A, defined by Eq. (5.6), has determinant $\det A = (1-\alpha-\beta)^{\frac{n(n-1)}{2}}$.

For $\alpha, \beta = 0, 1$ or $\beta = 1 - \alpha$, the proof is presented in Sec. 5.4.1; hence, assume $0 < \alpha, \beta < 1$ and $\beta \neq 1 - \alpha$. Since a matrix and its transposed have the same determinant, the matrix A^{T_n} is used for the mathematical induction.

The base of induction is $n = 2$; in this case the matrix is

$$A^{T_2} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{5.36}$$

and it has determinant $\det A^{T_2} = 1 - \alpha - \beta$.

The inductive step consists in assuming that $\det A^{T_{n-1}} = (1-\alpha-\beta)^{\frac{(n-1)(n-2)}{2}}$, i.e., the inductive hypothesis for dimension $n-1$, and proving the statement for dimension n.

Since $0 < \alpha, \beta < 1$ and $\beta \neq 1-\alpha$, the results in Sec. 5.4.2 can be applied. The transformations defined by Eq. (5.24) guarantee that the matrix $\overline{A^{T_n}}$ has the same determinant as A^{T_n}. According to Eq. (5.25), the determinant of A^{T_n} can be written as $\det A^{T_n} = (1 - \alpha - \beta)^{n-1} \det A^{T_{n-1}}$. The inductive hypothesis is now applied, i.e.

$$\begin{aligned} \det A &= \det A^{T_n} \\ &= \det \overline{A^{T_n}} \\ &= (1 - \alpha - \beta)^{n-1} \det A^{T_{n-1}} \\ &= (1 - \alpha - \beta)^{n-1} (1 - \alpha - \beta)^{\frac{(n-1)(n-2)}{2}} \\ &= (1 - \alpha - \beta)^{\frac{n(n-1)}{2}}, \end{aligned} \tag{5.37}$$

which concludes the proof.

5.5 Other properties

The matrix A is a left stochastic matrix, i.e., a real valued square matrix with each column summing up to 1. Namely, the $n \times n$ square matrix A has each of its entries defined as a probability, i.e., a real number between 0 and 1. The sum of column j is

$$\begin{aligned}\sum_{i=1}^{n} A_{ij} =& \sum_{i=1}^{n} \mathbb{P}(d' = i-1 | d = j-1) = \\ =& \sum_{i=0}^{n-1} \frac{\mathbb{P}(d = j-1 | d' = i)\mathbb{P}(d' = i)}{\mathbb{P}(d = j-1)} = \\ =& \frac{1}{\mathbb{P}(d = j-1)} \sum_{i=0}^{n-1} \mathbb{P}(d = j-1 | d' = i)\mathbb{P}(d' = i) = \\ =& \frac{1}{\mathbb{P}(d = j-1)} \mathbb{P}(d = j-1) = \\ =& \quad 1 \end{aligned} \tag{5.38}$$

where the second and the forth equalities are obtained using Bayes theorem and the law of total probability.

Note that, since A is a left stochastic matrix, A^T is a right stochastic matrix, i.e., each row has sum 1.

5.5.1 Eigenvectors and eigenvalues

Call $\mathcal{V}$ the matrix of eigenvectors of A^T, it can be shown that $\mathcal{V} = \{\mathcal{V}_{k'k}\}_{k',k \in \{0,\cdots,n-1\}}$ has elements

$$\mathcal{V}_{k'k} = \frac{1}{\binom{n-1}{k'}} \begin{cases} \displaystyle\sum_{i=0}^{k'} \binom{k}{i}\binom{n-1-k}{k'-i}\beta^{i-k}(-\alpha)^{k-i} \\ \qquad\qquad \text{if } k' \leq k \text{ and } k' \leq n-1-k \\ \displaystyle\sum_{i=0}^{n-1-k} \binom{k}{k'-i}\binom{n-1-k}{i}\beta^{k'-k-i}(-\alpha)^{k-k'+i} \\ \qquad\qquad \text{if } n-1-k < k' \leq k \\ \displaystyle\sum_{i=0}^{n-1-k'} \binom{k}{n-1-k'-i}\binom{n-1-k}{i}\beta^{k'-n+1+i}(-\alpha)^{n-1-k'-i} \\ \qquad\qquad \text{if } k' \geq n-1-k \text{ and } k' > k \\ \displaystyle\sum_{i=0}^{k} \binom{k}{i}\binom{n-1-k}{k'-i}\beta^{i-k}(-\alpha)^{k-i} \\ \qquad\qquad \text{if } k < k' < n-1-k. \end{cases} \tag{5.39}$$

Equation (5.39) can be proven to be valid for $n \leq 50$ using analytic computations in *Wolfram Mathematica 11.2.0.0*. For dimension of the matrix $n > 50$, better strategies to compute the determinant should be investigated to reduce computational time. The strategy in Sec. 5.4 to prove the determinant of A might be successful also in this case.

Equation (5.39) can be written in a compact form as

$$\mathcal{V}_{k'k} = \frac{1}{\binom{n-1}{k'}} \sum_{i=\max\{0,k-k'\}}^{\min\{k,n-1-k'\}} \binom{k}{i}\binom{n-1-k}{k'-k+i}\beta^{-i}(-\alpha)^{i}, \tag{5.40}$$

using Eq. (5.5); and also with the matrix indices $u, v \in \{1, \cdots, n\}$ in the form

$$\mathcal{V}_{uv} = \frac{1}{\binom{n-1}{u-1}} \sum_{i=\max\{0,v-u\}}^{\min\{v-1,n-u\}} \binom{v-1}{i}\binom{n-v}{u-v+i}\beta^{-i}(-\alpha)^{i}. \tag{5.41}$$

The inverse of the matrix $\mathcal{V}$ has elements

$$\mathcal{V}_{uv}^{-1} = (-1)^{u+v}\binom{n-1}{v-1}\binom{n-1}{u-1}\frac{\beta^{n-1}}{(\alpha+\beta)^{n-1}}\mathcal{V}_{uv} \tag{5.42}$$

for $u, v \in \{1, \cdots, n\}$. Also in this case, Eq. (5.42) can be demonstrated for $n \leq 50$, and has to be rigorously proven for a general n.

Still speculating about the validity of these equations to be true for a general n, it can be shown that the eigenvalues

$$\lambda_i = (1-\alpha-\beta)^{i-1} \tag{5.43}$$

of A^T are all distinct values, i.e. they have algebraic multiplicity 1, for $i \in \{1, \cdots, n\}$. Call Λ the matrix of the eigenvalues, i.e.,

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \tag{5.44}$$

the eigendecomposition of A^T is $A^T = \mathcal{V}\Lambda\mathcal{V}^{-1}$.

The formulation of the determinant of A does not contradict the validity of these equations, since it can be calculate as the product of all its eigenvalues, i.e.,

$$\det A = \prod_{i=1}^{n} \lambda_i \tag{5.45}$$

$$= \prod_{i=1}^{n} (1-\alpha-\beta)^{i-1} \tag{5.46}$$

$$= (1-\alpha-\beta)^{\frac{n(n-1)}{2}}, \tag{5.47}$$

that is the same result mathematically proven in Sec. 5.4. Note that the eigenvalues of A are the same as the ones of A^T.

As mentioned in Sec. 4.2.1, the matrix A is invertible if and only if $\det A \neq 0$, i.e., $\alpha \neq 1-\beta$. Also, for $0 < \alpha, \beta < 1$, the determinant of A converges to zero when n goes to infinity, since $|1-\alpha-\beta| < 1$.

5.5.2 Gauss elimination

Another interesting property of matrix A is the structure of the matrix when performing Gaussian elimination. Performing $n-1$ steps of Gauss elimination on the $n \times n$ matrix A, the diagonal matrix obtained has elements

$$d_i = (1-\alpha)^{n-2i+1}(1-\alpha-\beta)^{i-1}. \tag{5.48}$$

Also in this case, this formulation can be demonstrated for $n \leq 50$, and has to be rigorously proven for a general n.

As an example, the matrix A of dimension $n = 5$, after 4 steps of Gauss elimination is

$$\begin{bmatrix} (\alpha-1)^4 & 0 & 0 & 0 & 0 \\ 0 & -(\alpha-1)^2(\alpha+\beta-1) & 0 & 0 & 0 \\ 0 & 0 & (\alpha+\beta-1)^2 & 0 & 0 \\ 0 & 0 & 0 & -\frac{(\alpha+\beta-1)^3}{(\alpha-1)^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{(\alpha+\beta-1)^4}{(\alpha-1)^4} \end{bmatrix}. \tag{5.49}$$

This is an interesting property that might prove to be useful in future studies.

5.5.3 Mean and variance

In this subsection, mean and variance of the probability distribution $\mathbb{P}(d' = k'|d = k)$ are calculated. As described in the scheme in Fig. 5.1, Y_1 and Y_2 are the independent random variables with binomial distributions $\mathcal{B}(k, 1-\beta)$, [Eq. (4.4)], and $\mathcal{B}(n-1-k, \alpha)$, [Eq. (4.5)], respectively. The random variable $X = Y_1 + Y_2$ is given by the sum of Y_1 and Y_2.

The mean is a linear operator, therefore it follows that $\mathbb{E}(X) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$, hence

$$\mathbb{E}(X) = k(1-\beta) + (n-1-k)\alpha. \tag{5.50}$$

Since Y_1 and Y_2 are independent random variables, the variance of their sum

is $\mathbb{V}(X) = \mathbb{V}(Y_1) + \mathbb{V}(Y_2)$, therefore

$$\mathbb{V}(X) = k\beta(1-\beta) + (n-1-k)\alpha(1-\alpha). \tag{5.51}$$

5.6 Conclusion

In this chapter, a special case of a Poisson binomial distribution is analysed. The conditional probabilities of random variables $d'|d = k$ define the matrix A, introduced in Chapter 4. The random variables $d'|d = k$ consists of a sum of $n-1$ Bernoulli trials, k of those have probability of success $1-\beta$ each, and each of the remaining $n-1-k$ trials has probability of success α, whose probabilities correspond to special cases of Poisson binomial distributions.

In this chapter, several properties of the matrix A are discussed, and a proof for the determinant of the matrix A is presented, concluding the analysis presented in Chapter 4. Further analyses should focus on the analytical proofs of eigenvectors (Eq. 5.40) and eigenvalues (Eq. (5.43)) possibly applying the strategy used to prove the determinant of A (Sec. 5.4).

The results presented in this chapter should be generalised to the full case of the Poisson binomial distribution. The distribution of a sum of independent Bernoulli random variables that are not identically distributed should be investigated and analysed considering the properties disclosed in this chapter.

Chapter 6

Iterative procedure for network inference

6.1 Introduction

Chapter 3 describes the influence of false positive and false negative conclusions about links on the network structure. Several simulation results are presented and optimal values for each network topology and characteristic are shown. Section 3.5 speculates that the simulation study presented could be used as an iterative procedure to achieve a better network reconstruction. This chapter is dedicated to such a procedure. The results obtained in Chapter 4 are implemented in this procedure, since they provide an analytic framework to reconstruct the vertex degree distribution of the network.

Roughly, the iteration procedure consists of choosing various values for α to perform the iteration steps of the network reconstruction. For the first step, the standard value for α of 0.05 can be chosen as an example. The result of this first step gives a first estimate of the network topology of interest. For the second iteration step the value for α is adjusted according to the findings of the first step. This procedure is iterated, ultimately leading to a reconstruction of the network characteristic tailored to its previously unknown network topology.

6.2 Network inference: procedure

Chapter 4 shows that the vertex degree distribution of a network is influenced by false positive and false negative conclusions about the presence or absence of links. The analytic formula [Eq. (4.8)] shows the dependence of the biased vertex degree distribution on the network's dimension and the probabilities α and β of *type I* and *type II errors*, respectively. In Sec. 4.2.2, the density of the vertex degree distribution of the original network is found. Equation (4.10) $\mathcal{P} = A_k^+ \mathcal{P}'$ enables to calculating analytically the vertex degree distribution of the original network if the biased one and the probabilities of *type I* and *type II errors* are given; $\mathcal{P}$ and $\mathcal{P}'$ are the densities of the vertex degree distributions of original and biased networks. The matrix A_k^+ is the pseudoinverse of the truncated matrix A using the singular value decomposition, where A maps $\mathcal{P}$ into $\mathcal{P}'$, i.e. $\mathcal{P}' = A\mathcal{P}$. Note that A depends on n, α, β, where n is the number of vertices in the network.

As stated in Sec. 4.5, a limitation of this method is the assumption that the probabilities of *type I* and *type II errors* are known a priori. It has been shown that wrong estimates of these two errors, within certain bounds, do not cause the reconstruction of be rendered invalid. The iterative procedure presented in this chapter explores the way to adjust the estimates of these two errors, so to improve the reconstruction of the network of interest.

In Sec. 2.3, it is shown that α and β are reciprocally dependent. Their functional relationship depends on the nature of the problem taken into account. Consider, for example, a network of coupled oscillators, and assume to be able to detect the dynamic of every node. Perform an hypothesis test of no-correlation for every pair of nodes; if the $p-$value of the test is smaller than or equal to the significance level α, then the link between the corresponding nodes is considered to be present. In a real-world application, an estimate for the distribution of such correlation coefficients as a function of the number of data points N and the true correlation coefficient ρ should be available. Nevertheless, there exist cases in which ρ is not the same for every link; consider, e.g., a network of coupled oscillators where the initial coupling strengths are not the same for all the links. For sake of simplicity, and to

make the argument clearer, this chapter assumes that these correlation coefficients are identically distributed and follow the distribution in Eq. (2.31). In this case, it is reasonable to assume that it is possible to estimate ρ, given the distribution of the correlation coefficients. The relationship between α and β is described by Eqs. (2.32) and (2.33). Once α is chosen,

$$\beta = \beta(N, \rho, \alpha) \tag{6.1}$$

can be found as a function of $\alpha, N,$ and ρ.

Each step of the iterative procedure uses the information provided so far to reconstruct the vertex degree distribution of a network, when the biased one is given.

The first step of iteration consists itself of various parts. After acquiring data from a network, estimate the true correlation coefficient ρ. Fix the significance level to $\alpha = 0.05$ and perform the hypothesis test as explained above. The result of the test gives the so-called biased network, and therefore the vertex degree distribution $\mathcal{P}'$ can be calculated using Eq. (4.11). Knowing N and ρ, $\beta = \beta(N, \rho, \alpha)$ is found, since α has been fixed, and therefore the matrix A can be evaluated. All the ingredients needed to solve Eq. (4.10) are now available, hence the vertex degree distribution $\mathcal{P}$ of the original network is calculated, and this concludes the first iteration step. Table 6.1 summarises the procedure of the first iteration step.

Section 4.4 shows that wrong estimates of α and β, within certain bounds, do not cause the reconstruction $\mathcal{P} = A_k^+ \mathcal{P}'$ [Eq. (4.10)] of be rendered invalid. In addition, as shown in various simulations, the robustness on perturbations of α and β of the reconstruction method performs differently depending on the network topology of interest. Therefore, the procedure explained above and summarised in Table 6.1, can be enhanced using this result. Namely, the density of the vertex degree distribution $\mathcal{P}_0$, that is inferred at the last step of the procedure, is used to perform a robustness analysis varying the value of α, and consequently of β [Eq. (6.1)]. The value of α, that gives the most robust result, is used to iterate the procedure. A general step of iteration is summarised in Table 6.2.

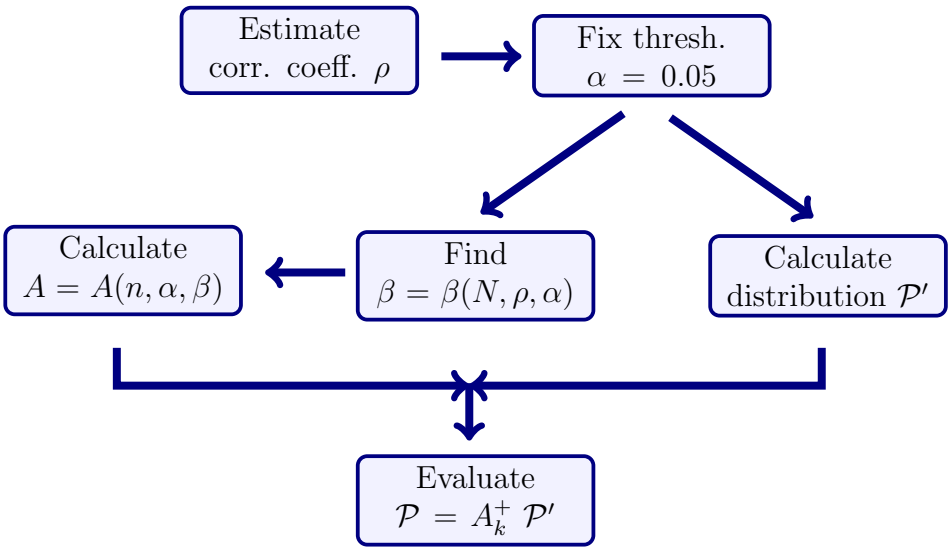


Table 6.1: Network Inference: procedure scheme. After taking data, a correlation analysis has to be performed. The value of the true correlation ρ is estimated. A value for the significance level $\alpha = 0.05$ is chosen as a threshold and all the values for which the null hypothesis is not rejected are discarded. The probability density function $\mathcal{P}'$ of the vertex degree distribution of the biased network is calculated. The value for β is found using Eq. (6.1), therefore also A is computed. Knowing $\mathcal{P}'$ and A, the probability density function $\mathcal{P}$ of the original vertex degree distribution is calculated using Eq. (4.10).

The robustness analysis consists of simulating several different biased networks $G'(\alpha)$ varying α, starting from a network G_0 with vertex degree distribution $\mathcal{P}_0$. For each $G'(\alpha)$, a robustness analysis is performed, as explained in Sec. 4.4. Namely, call $\mathcal{P}_0^*(\alpha, \delta)$ the density of the reconstructed vertex degree distribution, using as biased network $G'(\alpha)$, and a perturbation δ on α.

To quantify the bias between $\mathcal{P}_0$ and $\mathcal{P}_0^*(\alpha, \delta)$, the Kolmogorov-Smirnov distance is considered. The Kolmogorov-Smirnov test is a test based on the closeness of two distributions, which uses the Kolmogorov-Smirnov statistic, also called distance, to perform a test of hypothesis. The Kolmogorov-Smirnov distance is the largest difference between the two distributions [57].

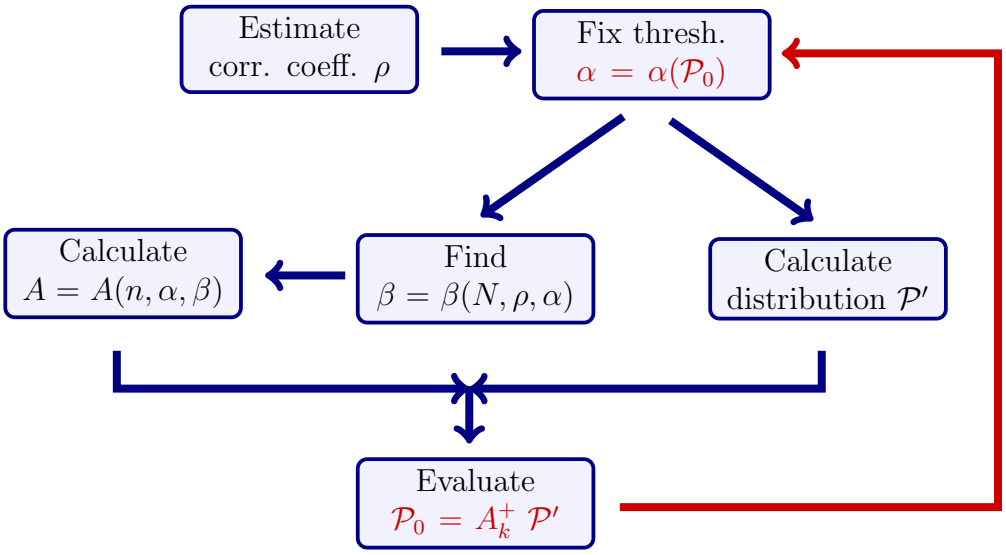


Table 6.2: Network Inference: procedure scheme with iteration. After taking data, a correlation analysis has to be performed. The value of the true correlation ρ is estimated. A value for the significance level α is chosen depending on $\mathcal{P}_0$ that is the vertex degree distribution inferred in the previous step. All the values for which the null hypothesis is not rejected are discarded. The probability density function $\mathcal{P}'$ of the vertex degree distribution of the biased network is calculated. The value for β is found using Eq. (6.1), therefore also A is computed. Knowing $\mathcal{P}'$ and A, the probability density function $\mathcal{P}$ of the original vertex degree distribution is calculated using Eq. (4.10).

The Kolmogorov-Smirnov distance

$$\mathcal{D}(\alpha, \delta) = \max |\mathcal{F}_0^*(\alpha, \delta) - \mathcal{F}_0|, \tag{6.2}$$

is the distance between $\mathcal{P}_0$ and $\mathcal{P}_0^*(\alpha, \delta)$, where $\mathcal{F}_0$ and $\mathcal{F}_0^*(\alpha, \delta)$ are the distributions of $\mathcal{P}_0$ and $\mathcal{P}_0^*(\alpha, \delta)$, respectively.

6.3 Simulation study, results and discussion

In this section a simulation study is presented to show the applicability of the method described in the previous section. Consider an Erdős-Rényi network with $n = 100$ nodes and probability of connection 0.2. Fix the true correla-

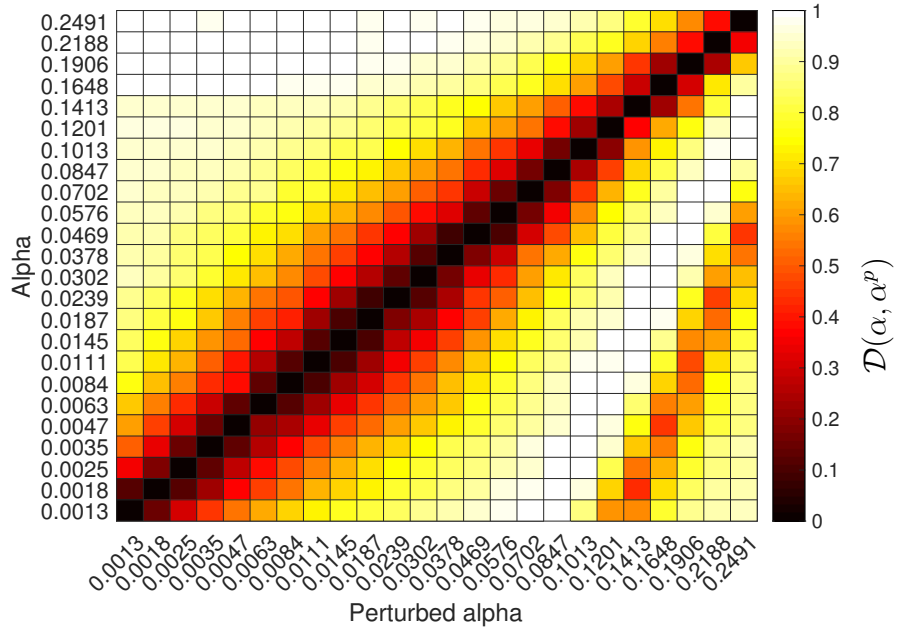


Figure 6.1: Kolmogorov-Smirnov distance $\mathcal{D}(\alpha, \alpha^p)$ obtained from $\mathcal{D}(\alpha, \delta)$, where $\alpha^p = \alpha + \delta\alpha$, i.e., the perturbed α (x-axis). An Erdős-Rényi network with $n = 100$ nodes and probability of connection 0.2 is used for the analysis.

tion coefficient $\rho = 0.3$ and the number of data points $N = 100$, following the density $f(100, 0.3, r)$ [Eq. (2.31)] a value for the correlation is assigned at random to each link; while for each absent link the value is chosen at random with distribution $f(100, 0, r)$. Fixing the significant level to $\alpha = 0.05$ and selecting all the links above this threshold, the biased or detected network is obtained. The procedure explained above is now applied to find the vertex degree distribution of the original network.

The result of the first iteration $\mathcal{P}_0 = A_k^+ \mathcal{P}'$ is used to perform a robustness analysis and the densities of the vertex degree distributions are calculated $\mathcal{P}_0^*(\alpha, \delta)$. Once the Kolmogorov-Smirnov distance $\mathcal{D}(\alpha, \delta)$ is evaluated, the value of α that gives the most robust results is chosen. Figure 6.1 shows the Kolmogorov-Smirnov distance; for simplicity, the x-axis represents the value of the perturbed α, i.e., $\alpha^p = \alpha + \delta\alpha$. In this case, various values of α give robust results, $\alpha = 0.0239$ is chosen and the procedure is iterated.

Figure 6.2 shows the results of three iteration steps. Note that the value of the true correlation coefficient $\rho = 0.3$ is assumed to be correctly inferred,

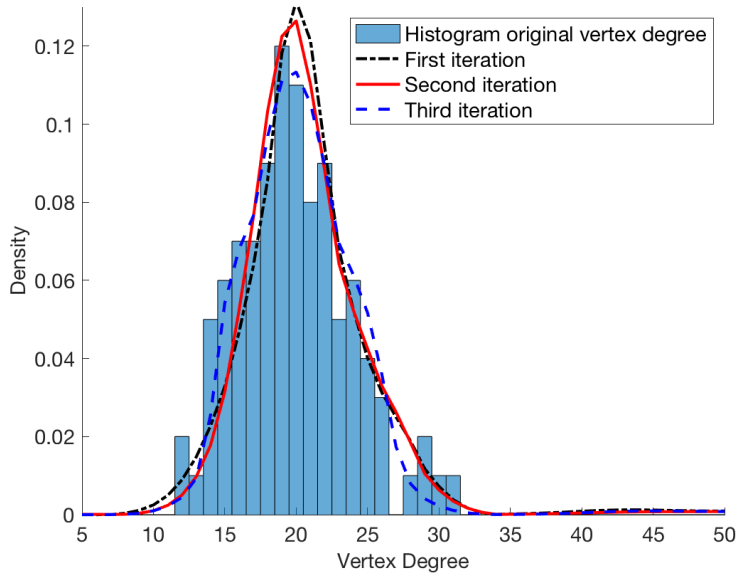


Figure 6.2: Results for three iteration steps of the reconstruction of an Erdős-Rényi network with $n = 100$ nodes and probability of connection 0.2, starting with correct inference of α and β.

and the theoretical distributions $f(100, 0.3, r)$ and $f(100, 0, r)$ are used to find β when the standard value of $\alpha = 0.05$ is fixed. Since a value for the correlation is assigned at random to each link, the actual value for α, and consequently β, will slightly differ form the theoretical values. As shown in Sec. 4.4, small perturbations of α and β do not cause the failure of the reconstruction method; in this example, this is reflected by the fact that the result of the first iteration gives already a correct reconstruction. The results for the other two iteration steps are shown here to demonstrate that the process converges, and once the correct reconstruction is achieved, the following results do not deviate from it. The difference of the three curves is due to the smoothness of the solutions.

Figure 6.3 shows the results of three iteration steps, when the estimate of α, and consequently β is initially wrong. Namely, the true correlation coefficient $\rho = 0.3$ is still assumed to be correctly inferred, $\alpha = 0.05$ is used to construct the detected network, but $\alpha = 0.07$ is used to evaluate $\mathcal{P}_0 = A_k^+ \mathcal{P}'$. Even if the result of the first iteration step does not provide a

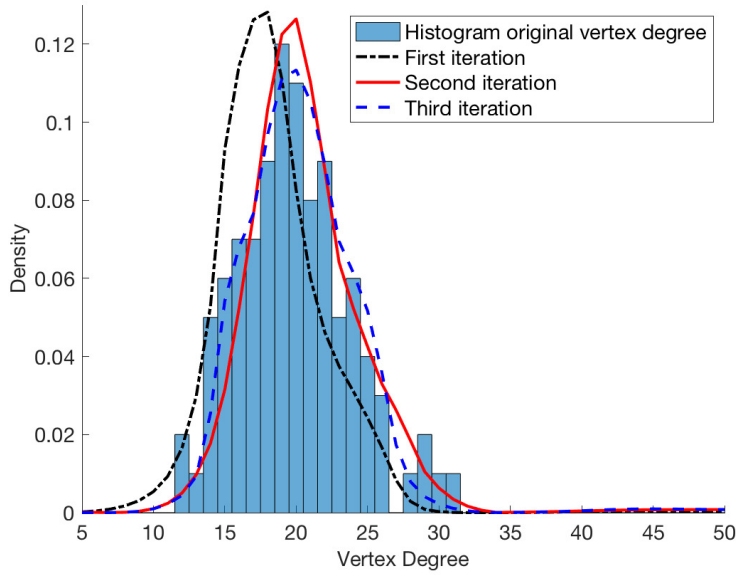


Figure 6.3: Results for three iteration steps of the reconstruction of an Erdős-Rényi network with $n = 100$ nodes and probability of connection 0.2, starting with wrong inference of α and β.

correct reconstruction, already at the second step this bias is corrected.

The method described in the previous section is now applied to a Scale-Free network of 100 nodes. Figure 6.4 shows the Kolmogorov-Smirnov distance $\mathcal{D}(\alpha, \alpha^p)$ obtained after the first iteration step. Interestingly, the values for α that give the most robust results are obtained for α as small as possible; intuitively, this means that high certainty about the presence of links is needed to recover the scale-freeness property of a network. Note that this is in agreement with the results shown in Fig. 3.4; in the case of Fig. 3.4, the relation between α and β is different from the case presented in this section, nevertheless, the general concept of keeping α as small as possible to correctly recover the scale-freeness property, remains valid.

Figure 6.5 shows the results of three iteration steps, when the estimate of α, and consequently β is initially wrong; the same values for ρ, true and wrong value of α are used. Also in this case, the second iteration step provides a correct reconstruction, which is improved at the third step.

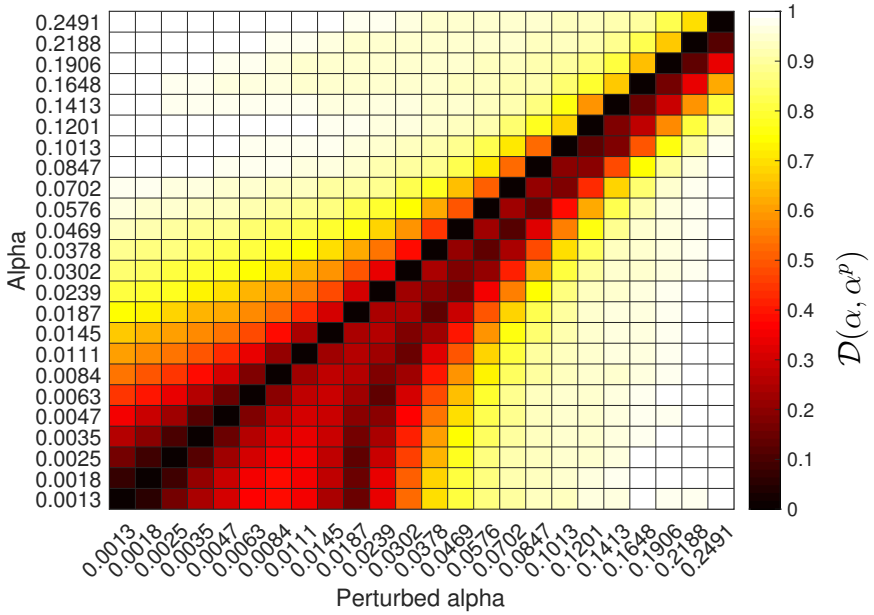


Figure 6.4: Kolmogorov-Smirnov distance $\mathcal{D}(\alpha, \alpha^p)$ obtained from $\mathcal{D}(\alpha, \delta)$, where $\alpha^p = \alpha + \delta\alpha$, i.e., the perturbed α (x-axis). A scale-free network of 100 nodes is used for the analysis.

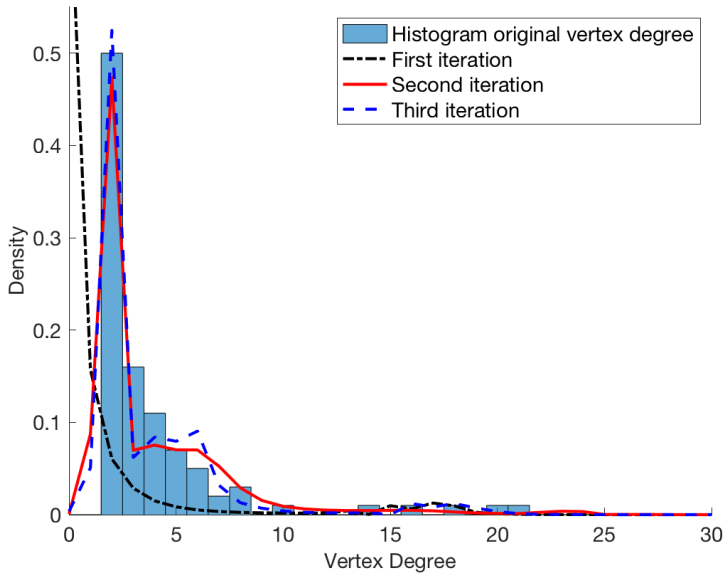


Figure 6.5: Results for three iteration steps of the reconstruction of a scale-free network with $n = 100$ nodes, starting with wrong inference of α and β.

6.4 Conclusion

The analysis presented in Chapter 4 is advanced in this chapter, and an iterative procedure to reconstruct the vertex degree distribution is suggested. This procedure should be used when the estimates of α and β might not be accurate and large errors on these estimates might occur; therefore, the robustness, shown in Chapter 4, is not guaranteed.

This iteration procedure allows to gain, at each step, a better insight on the network topology. Consequently, the value for α can be tuned to obtain a more robust reconstruction, thereby achieving a better result. It is shown, using some examples, that this method converges to the correct reconstruction of the vertex degree distribution. Furthermore, only few iteration steps are needed to achieve this goal.

Future investigations should apply this iterative procedure to various network topologies, to demonstrate that these results are not only case specific. Moreover, future studies should analyse the robustness of this procedure and find the size of perturbation needed to make the reconstruction fail even after various iteration steps. Finally, with the aim to make this procedure even more general, methods to infer ρ should be investigated. Furthermore, the case in which the inference of ρ is not accurate should be carefully studied.

Chapter 7

Impact of network characteristics on false conclusions about links

7.1 Introduction

The investigations presented so far in the thesis focus on the influence of false positive and false negative conclusions about links on the network structure. In Chapter 3, such an influence is analysed on network characteristics. It is shown that within the same network topology, the values for α and β leading to the least biased results change depending on the network characteristic of interest. In this chapter, the analysis is reversed - the study focuses on the influence of network characteristics on the probability of *type I* and *type II errors* to occur.

Assuming to know the underlying true network, a simulation study is performed in Sec. 7.2 to show the dependence of the probability of false positive and false negative links, and the shortest path length and the detour degree, respectively. These results are then applied in Sec. 7.3 when the underlying true network is not known a priori, to improve the network reconstruction.

7.2 Dependence of false conclusions on network characteristics

This section focuses on the study of the dependence of false positive and false negative conclusions about links from network characteristics. To this aim, two different oscillatory systems and reconstruction methods are considered, see Sec. 7.2.1.

The original network that has to be reconstructed is G; the subscriptions G_1 and G_2 are specified only when there is the need to differentiate the oscillatory system and reconstruction method used, see Sec. 7.2.1. Call S the matrix of all directed connection strengths inferred, i.e., S_{ij} is the strength of connection $i \to j$. When the aim is to reconstruct the original network G, i.e. to find the binary asymmetric adjacency matrix $\mathcal{A}$, from the observed connection strengths S, a threshold has to be chosen. If the connectivity measure passes a certain threshold, the link between the corresponding nodes is assumed to be present.

As shown in Chapter 2.3 for the case of Pearson correlation coefficients, also using the reconstruction methods presented in this section, the reconstructed coupling strengths have a certain distribution. This distribution is given by the sum of the distributions of the coupling strengths constructed from to the existing and not existing links. Consequently, *type I* and *type II errors* can occur when inferring the connections. Assuming to know the underlying true network, the interest is to check whether it is more likely to have a false conclusion depending on a specific local network characteristic. Sections 7.2.2 and 7.2.3 show that false positive conclusions are influenced by the shortest path length, while the detour degree influences false negative conclusions.

7.2.1 Reconstruction methods

The first case study consists of a directed network G_1 of coupled phase oscillators [56]

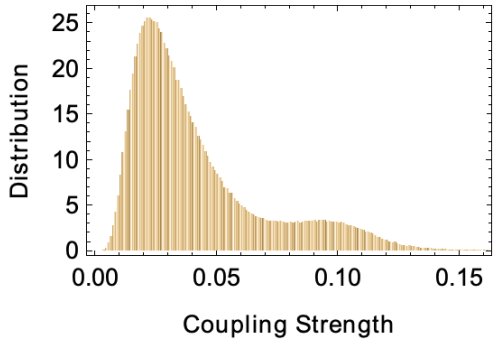
$$\dot{\phi}_k = \omega_k + \epsilon \sum_j \mathcal{A}_{kj} \sin(\phi_j - \phi_k - \Theta_{jk}). \tag{7.1}$$

The frequencies ω_k are uniformly distributed in the interval $(0.5, 1.5)$, the phase shifts Θ_{jk} are uniformly distributed in the interval $(0, 2\pi)$, and $\mathcal{A}$ is the adjacency matrix. The coupling constants are reconstructed using pairwise algorithm described in [55].

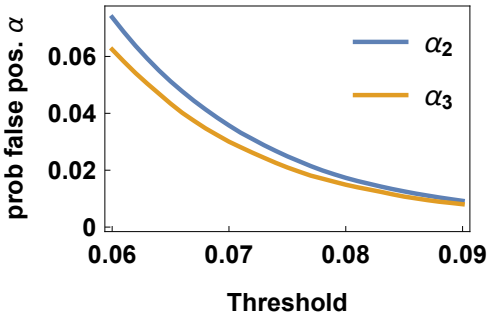
The second case study considered is a directed network of pulse-coupled neuronlike oscillators; such a network is indicated in this chapter with G_2. The model presented in [16] uses passive observation of pulse trains of all nodes to reconstruct networks of pulse-coupled neuronlike oscillators. It is assumed that units are described by their phase response curves and that their phases are instantaneously reset by incoming pulses. Using an iterative procedure, the properties of all nodes are recovered, namely their phase response curves and natural frequencies, as well as strengths of all directed connections. For the purpose of this thesis, only the reconstructed coupling strength are considered.

7.2.2 False positive and shortest path length

A false positive conclusion about a link $i \to j$ occurs when the reconstructed coupling strength passes the chosen threshold, but the directed link $i \to j$ does not exist. This means that there is a relatively strong strength of connection between node i and node j. This might happen because node i influences a third node k, which in turn influences node j, i.e., $i \to k$ and $k \to j$. Intuitively, this is likely to happen when the path between i and j is short. It is reasonable to assume that the strength of interaction decreases for each intermediate step. This concept corresponds to speculating that α is influenced by the shortest path length. Namely, a non-existing link $i \to j$ is erroneously considered present with probability α that depends on its shortest path length ℓ_{ij} (see Sec. 2.2 for the definition). In [43], the



(a) Histogram of the reconstructed coupling strengths



(b) Probabilities α_2 and α_3 as functions of the threshold

Figure 7.1: Reconstructed coupling strengths and relationship of α_k as a function of the threshold, for $N_{IT} = 100$ Erdős-Rényi networks G_1 with $n = 100$ vertices, probability of connection $p = 0.15$, $\epsilon = 0.3$, and $N_g = 500$.

authors state that this type of error is typical in a bivariate analysis. Call α_k the conditioned probability to observe a non-existing link $i \to j$, under the condition that its shortest path length is $\ell_{ij} = k$. As described before, it is more likely to have a false positive conclusion about a link when this non-existing link connects two nodes with shorter distance, i.e., if $k_1 < k_2$ then $\alpha_{k_1} > \alpha_{k_2}$.

A simulation study is made to demonstrate this speculation. Consider G to be an Erdős-Rényi network with $n = 100$ vertices and probability of connection $p = 0.15$. For each network, N_{IT} simulations are made to have enough statistical data. The same value ϵ for the coupling strength is used for all connections. A number N_g of data points is used to perform the reconstruction methods explained above.

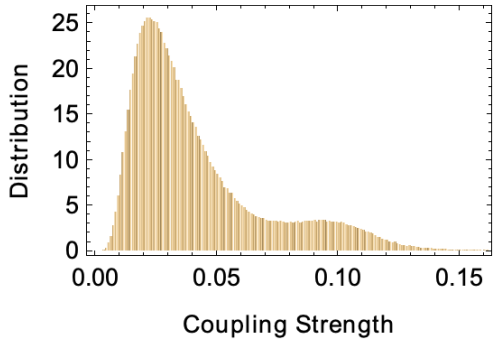
Figure 7.1 shows an example of $N_{IT} = 100$ simulations of a network, of type G_1, with original coupling strength $\epsilon = 0.3$; $N_g = 500$ data points are used for the reconstruction method. The histogram of the reconstructed coupling strengths [Fig. 7.1a] shows that the distributions of the true and absent links are not clearly separated, and therefore *type I errors* are expected for any value of the threshold. Figure 7.1b shows, as a function of the threshold, the conditioned probabilities α_2 and α_3 under the condition that the shortest path length is $\ell = 2$ and $\ell = 3$, respectively. Note that α_k for $k > 3$ does

not exist, due to the dimension and density of the network topologies. For every value of the threshold, the curve α_2 is always above α_3, as speculated above. Additionally, Fig 7.3 shows results for various values of ϵ and N_g. In all the cases shown, the relationship between the probability of a false positive conclusion about a link and its shortest path length is qualitatively the same.

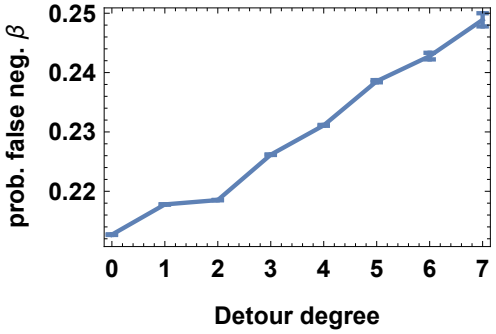
7.2.3 False negative and detour degree

A false negative conclusion about a link $i \to j$ occurs when the reconstructed coupling strength is below the chosen threshold, but the directed link $i \to j$ does exist. This might happen because node i not only influences node j directly, but it also influences it through many other short paths. Therefore, node j is influenced from all of the nodes involved, hence the interaction between i and j is corrupted. Intuitively, such effect is noticeable when the detours are rather short, and it becomes bigger when the number of detours increases. This concept corresponds to speculating that β depends on the detour degree (see Sec. 2.2 for the definition), and it is more likely to have a false negative conclusion about a link when it has a larger detour degree. The larger the detour degree $k_2 > k_1$ is, the more likely a *type II error* $\beta_{k_2} > \beta_{k_1}$ occurs, where β_k is the conditioned probability to miss an existing link $i \to j$, under the condition that its detour degree is $\Delta_{ij} = k$.

As before, a simulation study is made to demonstrate that this speculation is true. Consider the same example presented above, i.e., $N_{IT} = 100$ Erdős-Rényi networks G_1. Figure 7.2a is the histogram of the reconstructed coupling strengths (same as Fig. 7.1a). Figure 7.2b shows the conditioned probability β_k as a function of the detour degree Δ, for the value of the threshold equal to 0.08. Links with detour degree β_k for $k > 7$ do not exist in the network, due to lack of enough statistical data. As shown by the error bars, the larger the detour degree is, the fewer links with such detour degree are present in the network. Figure 7.2b presents an upward trend demonstrating that it is more likely to have a false negative conclusion about a link when it has a larger detour degree, as speculated above. Additionally,



(a) Histogram of the reconstructed coupling strengths



(b) Probability β_k as a function of the detour degree $\Delta = k$, for the value of the threshold equal to 0.08.

Figure 7.2: Reconstructed coupling strengths and β_k as a function of the detour degree Δ, for $N_{IT} = 100$ Erdős-Rényi networks G_1 with $n = 100$ vertices, probability of connection $p = 0.15$, $\epsilon = 0.3$, and $N_g = 500$.

Fig. 7.3 shows results for various values of ϵ, and N_g. In all the cases shown, the relationship between the probability of a false negative conclusions about a link and its detour degree is qualitatively the same.

7.2.4 Results

In Figs. 7.3 and 7.4, $N_{IT} = 100$ Erdős-Rényi networks, each with $n = 100$ vertices and probability of connection $p = 0.15$ are simulated. In Fig. 7.3, networks G_1 of coupled phase oscillators are considered, while networks G_2 of pulse-coupled neuronlike oscillators are considered in Fig. 7.4. Both cases show, for various values of ϵ and N_g, the histogram of the reconstructed coupling strengths, the relation of α_k and the threshold, and the relation of β_k and the detour degree Δ for a fixed value of the threshold.

The second column of Figs. 7.3 and 7.4 shows the relationship of α_k and the value of the threshold, as shown in Fig. 7.1. These plots demonstrate that for each value of the threshold, it is more likely to have a false positive conclusion about a link, when this non-existing link connects two nodes with shorter distance.

The third column of Figs. 7.3 and 7.4 shows the relationship of β_k and the detour degree Δ for a fixed value of the threshold. The value of the threshold

changes according to the histogram in the first column. These findings show that it is more likely to have a false negative conclusion about a link when it has a larger detour degree. However, there are two cases in which the relationship between β and Δ is not increasing monotonously; this happens in Fig. 7.3 for $\epsilon = 0.05$ and $N_g = 700, 800$. This is motivated by the fact that the initial value of the coupling strength $\epsilon = 0.05$ might be not strong enough to propagate on a path of length two and corrupt the strength of the direct connection. Namely, the direct connection has coupling strength $\epsilon = 0.05$, while the two-steps path has strength $\epsilon^2 = 0.0025$.

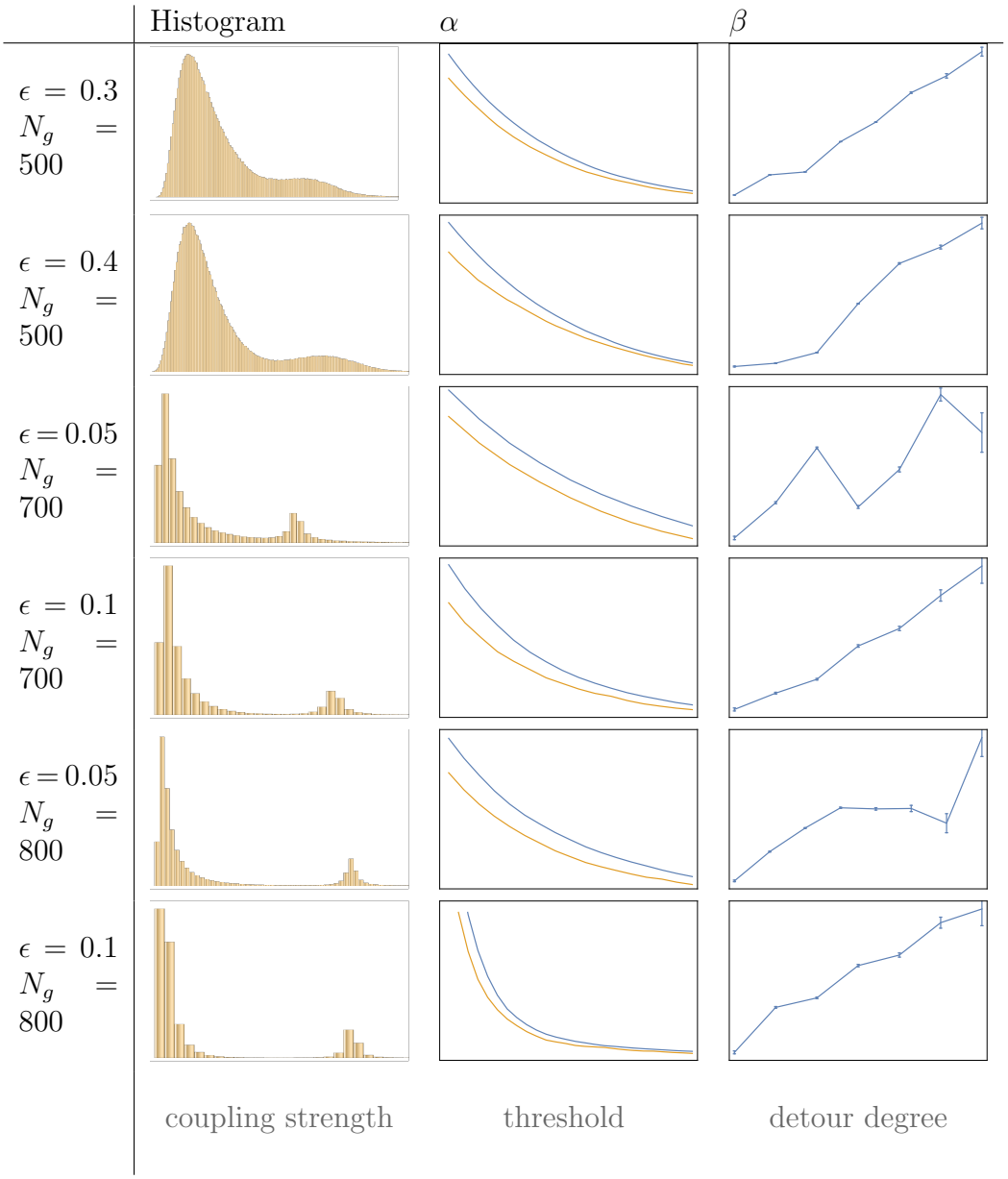


Figure 7.3: Results for $N_{IT} = 100$ Erdős-Rényi networks G_1 - network of coupled phase oscillators - each with $n = 100$ vertices and probability of connection $p = 0.15$. First column represents the histogram of the reconstructed coupling strengths; the second column shows the relation of α_k and the threshold; the third column presents the relation of β_k and the detour degree Δ for a fixed value of the threshold.

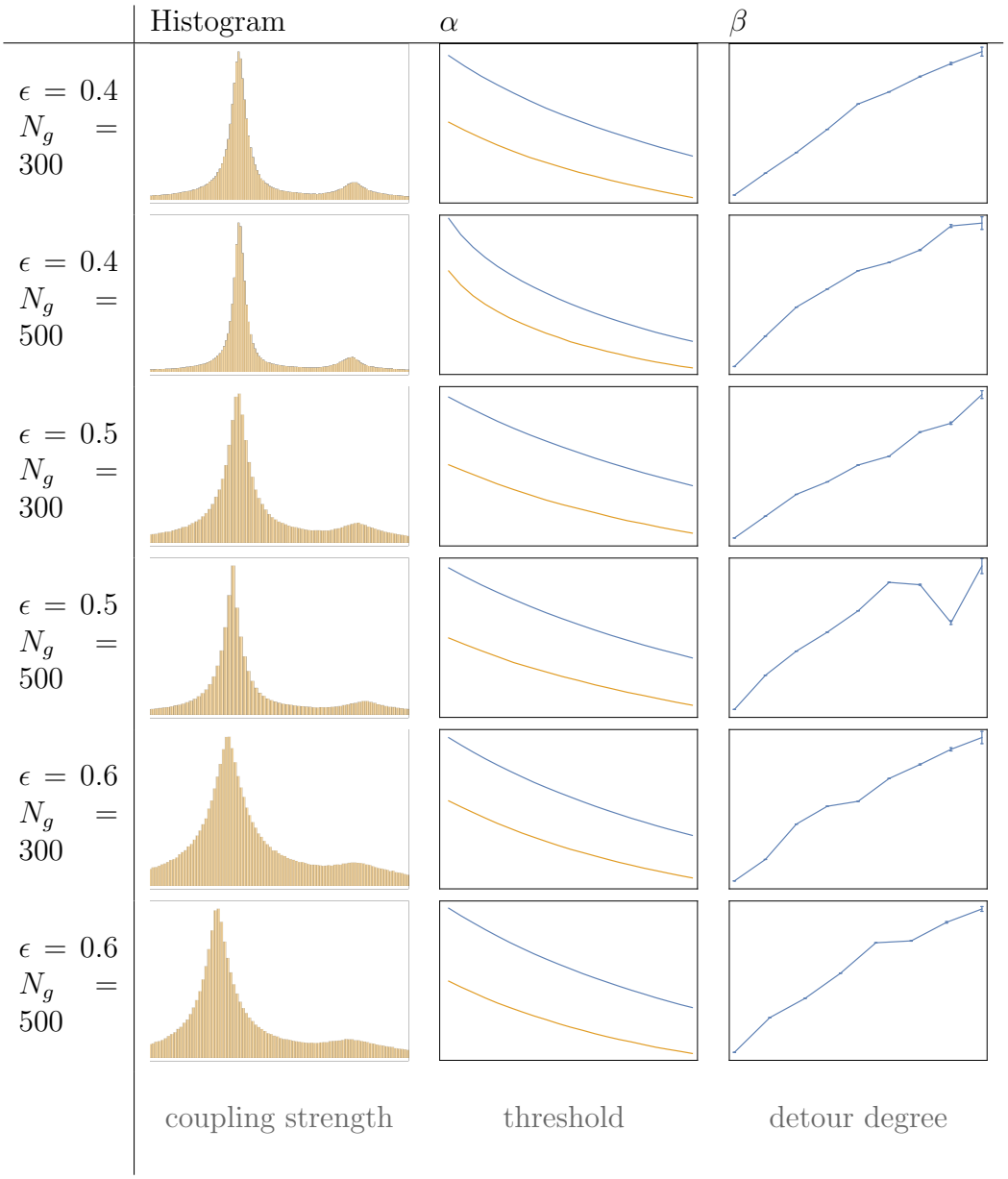


Figure 7.4: Results for $N_{IT} = 100$ Erdős-Rényi networks G_2 - network of pulse-coupled neuronlike oscillators - each with $n = 100$ vertices and probability of connection $p = 0.15$. First column represents the histogram of the reconstructed coupling strengths; the second column shows the relation of α_k and the threshold; the third column presents the relation of β_k and the detour degree Δ for a fixed value of the threshold.

7.3 Coupling strength and network characteristics

In the previous section [Sec. 7.2], the dependence of α and β on the shortest path length and the detour degree, respectively, are demonstrated. In this section, the interest is to study properties of the distribution of the reconstructed coupling strengths as a function of the detour degree and the shortest path length. The aim here is to find common properties for these characteristics, with the goal of improving network inference.

7.3.1 Weighted network

In this section, it is considered detected network the one defined by the weighted adjacency matrix S; the elements of S are the reconstructed coupling strengths. Note that even when the underlying true network is not known a priori, the adjacency matrix S can be evaluated, and no further assumptions are needed. The detected network is all-to-all connected with weighted edges, without self-loops, i.e., edges that connect a node to itself. In Sec. 2.2, the definitions provided for the shortest path length and the detour degree are valid only for binary networks; a generalisation for weighted networks is needed.

The shortest path length from node i to node j through a weighted network is the minimum of the sum, over all the possible paths from i to j, of the contributions given by the weights, i.e.,

$$\ell_{ij} = \min \left(s_{ik_1}^{-1} + \cdots + s_{k_n j}^{-1} \right), \tag{7.2}$$

where s_{ij} are the elements of the adjacency matrix S, and therefore it corresponds to the weight of the link $i \to j$ [51]. Note that for binary networks, this definition is coherent with the one in Sec. 2.2. An existing link corresponds to weight 1 and an absent link to weight 0, the latter would lead to an infinite contribution in the sum. Therefore Eq. (7.2) reduces to the number of links separating i and j if the shortest path is taken.

The detour degree of the link $i \to j$ measures the contribution of all the possible 2-step paths from i to j. In weighted networks, such a contribution must consider the weight of the edges. Namely, the detour degree

$$\Delta_{ij} = \sum_k s_{ik} s_{kj} \tag{7.3}$$

is scaled by the product of the weights of the two edges that form the 2-step path; s_{ij} are the elements of the adjacency matrix S. For binary networks, this definition is coherent with the definition in Sec. 2.2, since for $s_{kh} \in \{0, 1\}$ Eq. (7.3) reduces to the total number of paths of length 2 from node i to node j. Note that, in both the binary and weighted cases, Eq. (7.3) can be expressed by the matrix form $\Delta = S^2$.

Using definitions in Eqs. (7.2) and (7.3), the distributions of the reconstructed coupling strengths as a function of the shortest path length and the detour degree are analysed in Secs. 7.3.2 and 7.3.3, respectively.

7.3.2 Coupling strength and shortest path length

Consider, like the examples presented in Secs. 7.2.2 and 7.2.3, $N_{IT} = 100$ simulations of Erdős-Rényi networks of type G_1, with original coupling strength $\epsilon = 0.4$, and $N_g = 500$ data points used for the reconstruction method. Figure 7.5 shows the distribution of the reconstructed coupling strengths as a function of the shortest path length, or distance. All the values for the reconstructed coupling strengths lie in the part of the plane delimited by the curve $y = 1/x$; this is motivated by the fact that the shortest path length between nodes i and j must be equal to or smaller than the inverse of the reconstructed coupling strength s_{ij} by definition, see Eq. (7.2).

Using the true underlying network for the colour coding, Fig. 7.6 shows the same distribution as the one in Fig. 7.5, displaying in blue the original links and the absent links in orange. Interestingly, the reconstructed coupling strengths of the original links lie on the $1/x$ curve almost entirely (98.9%); while the values associated to the absent links have a double-distribution: on the curve and below. This information can thus be used when the underlying

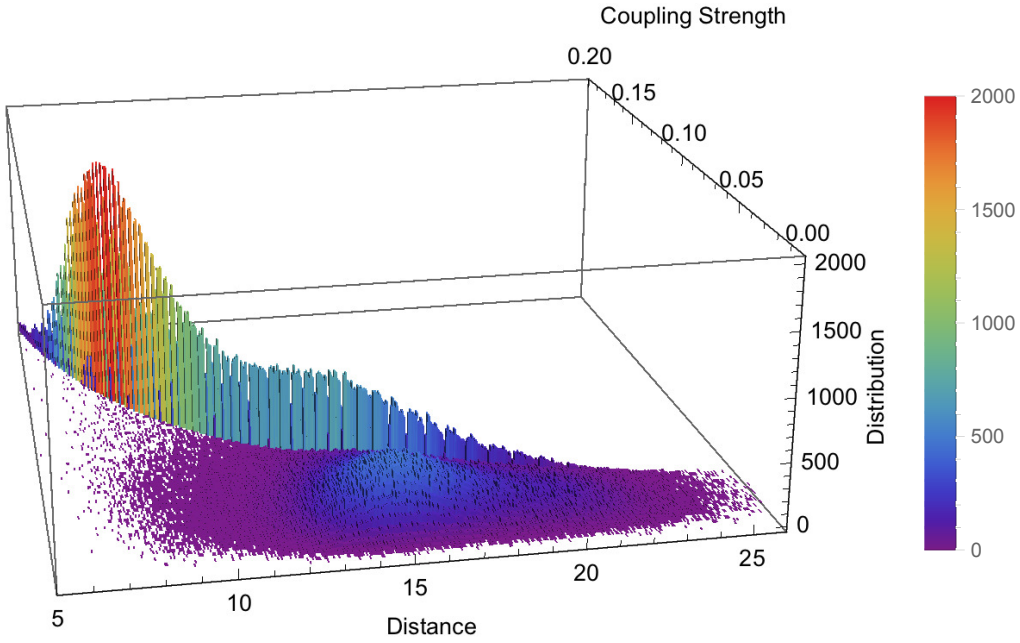


Figure 7.5: Histogram of the reconstructed coupling strengths as a function of the shortest path length (distance). Colour code expresses the frequency.

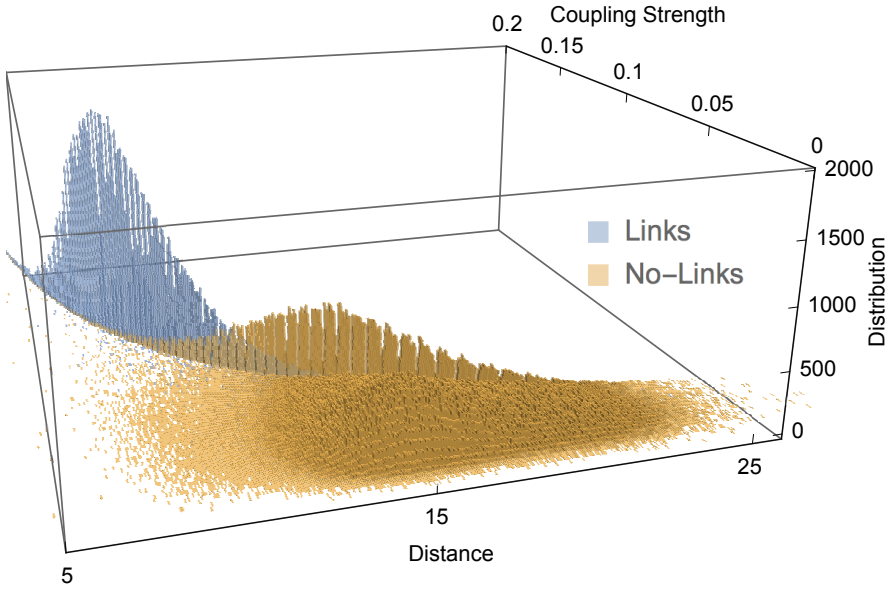


Figure 7.6: Histogram of the reconstructed coupling strengths as a function of the distance for the true and absent original links.

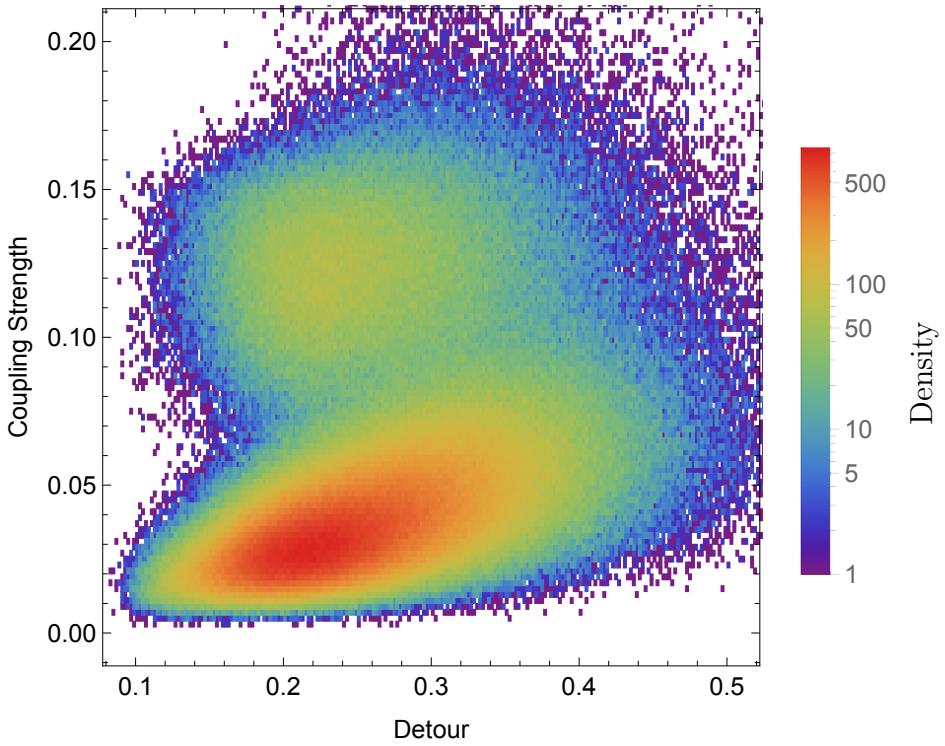


Figure 7.7: Density histogram for the reconstructed coupling strengths as a function of the detour degree. Colour code expresses the frequency in the logarithmic scale.

true network is not known a priori to improve the network inference, as shown in Sec. 7.3.4.

7.3.3 Coupling strength and detour degree

As above, consider $N_{IT} = 100$ simulations of Erdős-Rényi networks of type G_1, with $\epsilon = 0.4$ and $N_g = 500$. Figure 7.7 shows the reconstructed coupling strengths as a function of the detour degree; the logarithmic scale is used to express the density. From a visual inspection, a double-distribution structure emerges from this plot.

Figure 7.8 shows the scatter plot of the reconstructed coupling strengths as a function of the detour degree, for the original links in blue and the absent links in orange. Note that Fig. 7.7 coincides to the density histogram of the points plotted in Fig. 7.8. As expected the points corresponding to

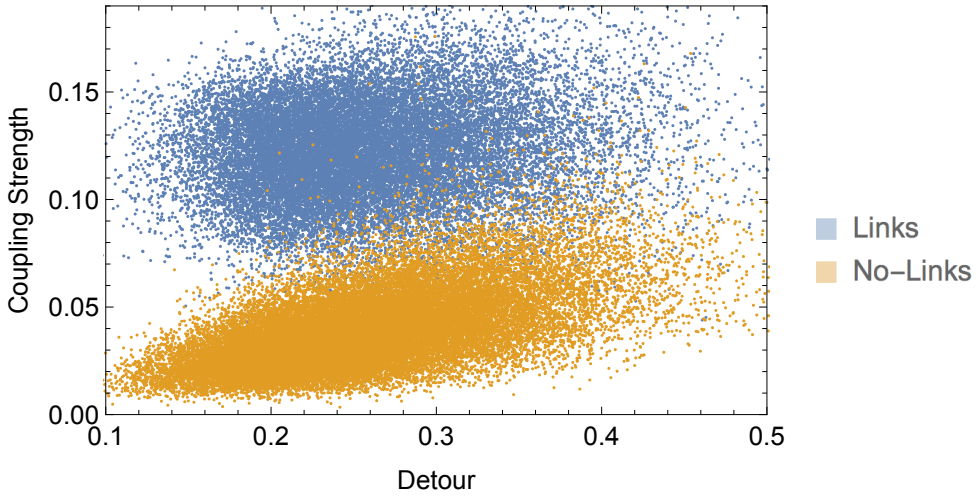


Figure 7.8: Reconstructed coupling strengths as a function of the detour degree for the original true and absent links

the strongest coupling strengths, i.e. the ones lying in the upper part of the plot, are originated by the original links. Furthermore, a dependence on the detour degree emerges in both true and absent original links. The double-distribution structure, arising from Fig. 7.7, corresponds indeed to the distributions of true and absent original links. Therefore, for these types of networks, when the underlying true network is not known a priori, a different rule to decide the value of the threshold should be chosen, as shown in Sec. 7.3.4.

7.3.4 Advanced threshold

The results presented in Secs. 7.3.2 and 7.3.3 show the dependence of the reconstructed coupling strengths and two network characteristics - the shortest path length and the detour degree. These results suggest that network reconstructions might benefit from different strategies related to the definition of the threshold. The naïve choice consists in selecting a value for the coupling strength, and all the reconstructed coupling strengths larger than this value are considered to be present, the rest are discarded. In this section, two novel choices for the threshold are presented.

Considering the analysis in Sec. 7.3.2, the first suggested choice for the

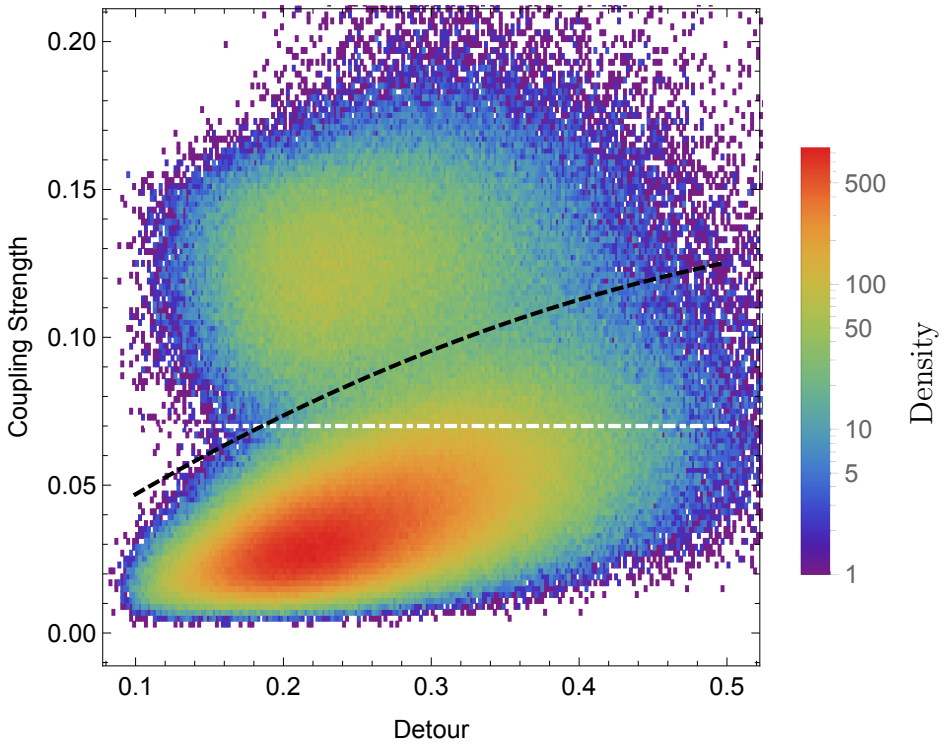
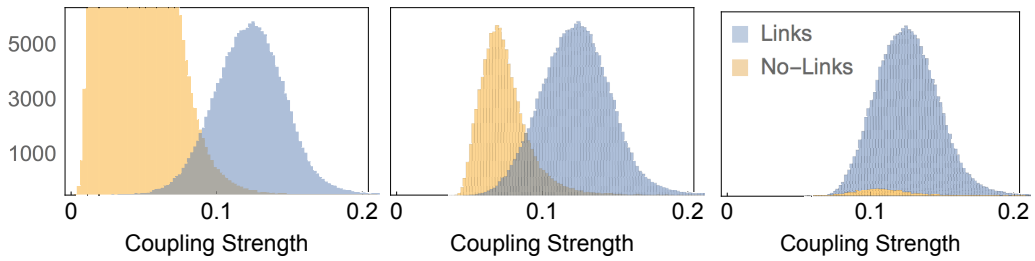


Figure 7.9: Naïve threshold (white dotted-dashed line) and detour-threshold (black dashed line), and density histogram for the reconstructed coupling strengths as a function of the detour degree. Colour code expresses the frequency in the logarithmic scale.

threshold consists of discarding all the reconstructed coupling strengths that do not lie on the $y = 1/x$ curve, where x and y are the shortest path length and the reconstructed coupling strength, respectively. As above, consider $N_{IT} = 100$ simulations of Erdős-Rényi networks of type G_1, with $\epsilon = 0.4$ and $N_g = 500$. Take all links whose reconstructed coupling strength is the reciprocal of the shortest path length, i.e., the points lying on the curve $1/x$. When making this choice for the threshold, the probability of false positive links is 0.093, and the probability of false negative links is 0.011. If instead the naïve threshold of 0.07 is used, i.e., keeping only the coupling strengths larger than 0.07, the probabilities are $\alpha = 0.059$ and $\beta = 0.014$. Note that the probability of detecting a false positive link is slightly increased with the new choice of the threshold, while the probability of false negative links is



(a) Original (b) Path-threshold (c) Detour-threshold

Figure 7.10: Histogram of the reconstructed coupling strengths for the true and absent original links in three cases: (a) all the reconstructed coupling strengths; (b) when considering the thresholds based on the shortest path length; (c) the thresholds based on the detour degree.

decreased.

The second suggested choice for the threshold is based on the analysis of Sec. 7.3.3. As shown, the dependence of the reconstructed coupling strengths as a function of the detour degree, presents a double-distribution structure. Intuitively, the new threshold is defined as the curve that corresponds to the minimum between the two bulges of the histogram in Fig. 7.7. Figure 7.9 shows the naïve and new thresholds. When making the new choice for the threshold, the probability of false positive links is 0.006, and the probability of false negative links is 0.083. Note that the probability of detecting a false negative link is increased with the new choice of the threshold, while the probability of false positive links is decreased.

Figure 7.10 shows the histograms of the reconstructed coupling strengths for the true and absent links in three situations: all the coupling strengths, and when using the thresholds based on the shortest path length (path-threshold) and based on the detour degree (detour-threshold). Figure 7.10b shows that the path-threshold performs better then any naïve choice for the threshold when the aim is to minimise the number of false negative. Even if a relatively large number of false positive links is still present, the naïve choice for the threshold would have let to an even larger number, when keeping the same number of false negative. The goal achieved with the detour-threshold proves to be evident from Fig. 7.10c; the number of remaining false positive

Type of	$\epsilon = 0.4$		$\epsilon = 0.3$		$\epsilon = 0.2$		$\epsilon = 0.1$	
threshold	α	β	α	β	α	β	α	β
naïve	0.059	0.014	0.024	0.155	0.054	0.232	0.119	0.358
path	0.093	0.011	0.128	0.021	0.161	0.051	0.159	0.172
detour	0.006	0.083	0.008	0.186	0.024	0.256	0.084	0.338

Table 7.1: Table of proportion of false positive (α) and false negative (β) links for three choices of the threshold (naïve, path-threshold, and detour-threshold) and network of type G_1 (see Sec. 7.2.1) with initial coupling strength ϵ.

links appears to be insignificant with respect to the total.

Table 7.1 summarises the results for $N_{IT} = 100$ simulations of Erdős-Rényi networks of type G_1, with $N_g = 500$ and initial coupling strength ϵ. In all four scenarios, the result that minimises α is obtained using the detour-threshold, and the result that minimises β is obtained using the path-threshold.

7.4 Conclusion

Assuming to know the underlying true network, numerical simulations presented in Sec. 7.2 show that local network characteristics influence the probability of false negative and false positive conclusions about links to occur. In particular, *type I* and *type II errors* are influenced by the shortest path length and the detour degree, respectively [Secs.7.2.2 - 7.2.3].

When the underlying true network is not known a priori, the interest is to check whether common rules apply for the reconstructed coupling strengths as a function of the detour degree and the shortest path length. In Sec. 7.3 the dependence between these quantities is analysed and common rules are found. These relations suggest a novel approach to for the choice of the threshold. In Sec. 7.3.4 two new advanced choices for the threshold are presented, and each one of them leads to minimise the proportion of false positive or false negative conclusions about links. These two new choices are a first attempt to advanced thresholding and the results obtained are promising for further

studies. Future analysis should investigate how to combine these thresholds, in a four dimensional space, with the aim to find a unique threshold that improve both probabilities of false positive and false negative conclusions. This should incorporate both more statistics and more understanding of the roles of the influence of the shortest path length and detour degree on α and β.

Further studies should also investigate if additional network characteristics play a role in reconstruction analyses, and therefore could be used to improve network inference. Furthermore, different reconstruction methods should be studied to check whether the common rules found in this chapter apply to a wider range of cases.

Chapter 8

Conclusions

A reliable inference of networks from data is of key interest in many scientific fields. Several methods have been suggested in the literature to reliably determine links in a network. These techniques rely on statistical methods, typically controlling the number of false positive links, but not considering false negative links. In this thesis new methodologies to improve network inference are suggested. Initial analyses demonstrate the impact of false positive and false negative conclusions about the presence or absence of links on the resulting inferred network. Consequently, revealing the importance of making well-considered choices leads to suggest new approaches to enhance existing network reconstruction methods.

A simulation study, presented in Chapter 3, shows that different values to balance false positive and false negative conclusions about links should be used in order to reliably estimate network characteristics. The existence of *type I* and *type II errors* in the reconstructed network, also called biased network, is accepted. Consequently, an analytic method that describes the influence of these two errors on the network structure is explored. As a result of this analysis, an analytic formula of the density of the biased vertex degree distribution is found (Chapters 4- 5).

In the inverse problem, the vertex degree distribution of the true underlying network is analytically reconstructed, assuming the probabilities of *type I* and *type II errors*. A further analysis shows that this procedure is robust

with respect to errors in α and β as they typically occur when they have to be estimated from data. This implies that wrong estimates, within reasonable limits, do not cause the reconstruction of the node degree distribution to be rendered invalid. In Chapter 6, an iterative procedure to enhance this method is presented in the case of large errors on the estimates of α and β.

The investigations presented so far focus on the influence of false positive and false negative links on the network characteristics. In Chapter 7, the analysis is reversed - the study focuses on the influence of network characteristics on the probability of *type I* and *type II errors*, in the case of networks of coupled oscillators. The probabilities of α and β are influenced by the shortest path length and the detour degree, respectively. These results have been used to improve the network reconstruction, when the true underlying network is not known a priori, introducing a novel and advanced concept of threshold.

Future studies should investigate the influence of *type I* and *type II errors* on other network characteristics, in order to enhance network inference. The identification of functional relationships of various biased network characteristics and their counterparts from the true network structure could result in a better identification of the exact underlying true network, when the single results are combined. The correct reconstruction of network characteristics discloses important properties of the general network structure, leading to the identification of key factors, such as small world behaviour or scale-freeness.

In the case of networks of coupled oscillators, future investigations should explore the existence of relations of other network characteristics and the probabilities of α and β. Further types of thresholds, like the ones described in this thesis, can be introduced to optimise the balance of *type I* and *type II errors*. The combination of these results leads to create one high-dimensional threshold that considers the relations of α and β, and all the network characteristics investigated. Future studies should also examine other dynamical systems to find possible relations of network characteristics and the probabilities of α and β.

In a more general framework, the work reported in this thesis, should be tested in various complex systems, when a network is to be inferred. This

ranges from networks of oscillators with chaotic dynamics to applications to the Neurosciences. Another application of the methodologies investigated in this thesis is to time-variant networks, which are a topic of interest in numerous fields. Time-variant networks are networks with connections that change over time. These networks are usually studied in terms of multilayer networks, namely each time step corresponds to a layer of the network. The methodologies described in this thesis should be applied to each single layer, and therefore generalised for multilayer networks.

Another problem of key interest is when only a subset of all the nodes has been observed. Using the approach presented in this work might help to improve network reconstruction also in the case of undetected nodes. The correct reconstruction of the vertex degree distribution of the subset of nodes that has been observed, can be used to calculate the degrees of the undetected nodes. It is reasonable to assume that the degrees of the unobserved nodes follow the same degree distribution as the observed ones.

Bibliography

[1] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74, 2002.

[2] J. R. Banavar, F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo. Topology of the fittest transportation network. *Phys. Rev. Lett.*, 84, 2000.

[3] Y. Bar-Yam. *Dynamics of complex systems.* Boulder, CO : Westview Press, 2003.

[4] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[5] A. L. Barabási and M. Pósfai. *Network science.* Cambridge University Press, Cambridge, 2016.

[6] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical Processes on Complex Networks.* Cambridge University Press, Cambridge, 2008.

[7] S. Bialonski, M. T. Horstmann, and K. Lehnertz. From brain to earth and climate systems: Small-world interaction networks or not? *Chaos*, 20, 2010.

[8] S. Bialonski, M. Wendler, and K. Lehnertz. Unraveling spurious properties of interaction networks with tailored random networks. *PLoS ONE*, 6, 2011.

[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Phys. Rep.*, 424, 2006.

[10] B. Bollobás and O. M. Riordan. *Mathematical results on scale-free random graphs*, chapter 1. Wiley-Blackwell, 2005.

[11] M. A. Branch, T. F. Coleman, and Y. Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Sci. Comput.*, 21, 1999.

[12] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10, 2009.

[13] K. Butler and M. A. Stephens. The distribution of a sum of independent binomial random variables. *Methodol. Comput. Appl.*, 19, 2017.

[14] G. Cecchini and B. Schelter. Analytical approach to network inference: Investigating degree distribution. *Phys. Rev. E*, 98, 2018.

[15] G. Cecchini, M. Thiel, B. Schelter, and L. Sommerlade. Improving network inference: The impact of false positive and false negative conclusions about the presence or absence of links. *J. Neurosci. Meth.*, 307, 2018.

[16] R. Cestnik and M. Rosenblum. Reconstructing networks of pulse-coupled oscillators from spike trains. *Phys. Rev. E*, 96, 2017.

[17] M. Chavez, M. Valencia, V. Latora, and J. Martinerie. Complex networks: new trends for the analysis of brain connectivity. *Int. J. Bifurcat. Chaos*, 20, 2010.

[18] P. Clusella, P. Grassberger, F. J. Pérez-Reche, and A. Politi. Immunization and targeted destruction of networks using explosive percolation. *Phys. Rev. Lett.*, 117, 2016.

[19] R. Cohen and S. Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, Cambridge, 2010.

[20] T. F. Coleman and Yuying Li. On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Math. Program.*, 67, 1994.

[21] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning poisson binomial distributions. *Algorithmica*, 72, 2015.

[22] F. De Vico Fallani, J. Richiardi, M. Chavez, and S. Achard. Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philos. T. R. Soc. B*, 369, 2014.

[23] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, 8th edition, 2011. ISBN-13: 978-0-538-73352-6.

[24] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

[25] V. M. Eguíluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian. Scale-free brain functional networks. *Phys. Rev. Lett.*, 94, 2005.

[26] P. Erdős and A. Rényi. On random graphs. *Publ. Math-Debrecen*, 6, 1959.

[27] P. Erdős and A. Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 1960.

[28] G. Fagiolo. Clustering in complex directed networks. *Phys. Rev. E*, 76, 2007.

[29] M. Frank and J. M. Buhmann. Selecting the rank of truncated SVD by Maximum Approximation Capacity. 2011.

[30] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE T. Inform. Theory*, 60, 2014.

[31] M. Grinstead, C. and L. Snell, J. *Introduction to Probability*. American Mathematical Society, 1997.

[32] P. C. Hansen. The truncatedsvd as a method for regularization. *BIT*, 27, 1987.

[33] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci.*, 104, 2007.

[34] Y. Hong. On computing the distribution function for the poisson binomial distribution. *Comput. Stat. Data An.*, 59, 2013.

[35] M. Jalili and M. G. Knyazeva. Constructing brain functional networks from eeg: partial and unpartial correlations. *J. Integr. Neurosci.*, 10, 2011.

[36] B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83, 2011.

[37] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics*. Mathematics of Statistics. D. Van Nostrand Company, second edi edition, 1951.

[38] B. Kralemann, A. Pikovsky, and M. Rosenblum. Reconstructing effective phase connectivity of oscillator networks from observations. *New J. Phys.*, 16, 2014.

[39] M. Kurant and P. Thiran. Extraction and analysis of traffic and topologies of transportation networks. *Phys. Rev. E*, 74, 2006.

[40] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87, 2001.

[41] S. Li, F. Li, W. Liu, and M. Zhan. Network reconstruction by linear dynamics. *Physica A*, 404, 2014.

[42] G. S. Lo. *A Course on Elementary Probability Theory*. 2018.

[43] W. Mader, M. Mader, J. Timmer, M. Thiel, and B. Schelter. Networks: On the relation of bi- and multivariate measures. *Sci. Rep.*, 5, 2015.

[44] M. B. C. Menezes, S. Kim, and R. Huang. Constructing a watts-strogatz network from a small-world network with symmetric degree distribution. *PLoS ONE*, 12, 2017.

[45] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89, 2002.

[46] M. E. J. Newman. Random graphs as models of networks. In Stefan Bornholdt and Hans Georg Schuster, editors, *Handbook of Graphs and Networks: From the Genome to the Internet*. 2002.

[47] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45, 2003.

[48] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[49] G. Ódor and B. Hartmann. Heterogeneity effects in power grid network models. *Phys. Rev. E*, 98, 2018.

[50] E. Olbrich, T. Kahle, N. Bertschinger, N. Ay, and J. Jost. Quantifying structure in networks. *Eur. Phys. J. B*, 77, 2010.

[51] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc. Networks*, 32, 2010.

[52] L. Pessoa. Understanding brain networks and brain organization. *Phys. Life Rev.*, 11, 2014.

[53] S. E. Petersen and O. Sporns. Brain Networks and Cognitive Architectures. *Neuron*, 88, 2015.

[54] A. Pikovsky. Reconstruction of a neural network from a time series of firing rates. *Phys. Rev. E*, 93, 2016.

[55] A. Pikovsky. Reconstruction of a random phase dynamics network from observations. *Phys. Lett. A*, 382, 2018.

[56] A. Pikovsky, M. G. Rosenblum, and J. Kurths. *Synchronization, A Universal Concept in Nonlinear Sciences*. Cambridge University Press, Cambridge, 2001.

[57] G. P. Quinn and M. J. Keough. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, 2002.

[58] P. Sahoo. *Probability and Mathematical Statistics*. 2015.

[59] S. Schinkel, G. Zamora-López, O. Dimigen, W. Sommer, and J. Kurths. Functional network analysis reveals differences in the semantic priming task. *J. Neurosci. Meth.*, 197, 2011.

[60] S. L. Simpson, S. Hayasaka, and P. J. Laurienti. Exponential random graph modeling for complex brain networks. *PLoS ONE*, 6, 2011.

[61] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag. Organization, development and function of complex brain networks. *Trends Cogn. Sci.*, 8, 2004.

[62] M. Valencia, J. Martinerie, S. Dupont, and M. Chavez. Dynamic small-world behavior in functional brain networks unveiled by an event-related networks approach. *Phys. Rev. E*, 77, 2008.

[63] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, New York, 2004.

[64] D. J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton studies in complexity. Princeton University Press, 1999.

[65] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[66] Y. Xue, J. Wang, L. Li, D. He, and B. Hu. Optimizing transport efficiency on scale-free networks through assortative or dissortative topology. *Phys. Rev. E*, 81, 2010.

[67] H. Yanai, K. Takeuchi, and Y. Takane. *Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition*. Springer, 2011.

[68] C. H. Yeung and K. Y. M. Wong. Phase transitions in transportation networks with nonlinearities. *Phys. Rev. E*, 80, 2009.

[69] T. Zerenner, P. Friederichs, K. Lehnertz, and A. Hense. A gaussian graphical model approach to climate networks. *Chaos*, 24, 2014.

[70] M. Zhang, Y. Hong, and N. Balakrishnan. The generalized poisson-binomial distribution and the computation of its distribution function. *J. Stat. Comput. Sim.*, 88, 2018.