Journal of Neuroscience Methods 307 (2018) 31–36

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

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

Gloria Cecchini[a,b,*], Marco Thiel[a], Björn Schelter[a], Linda Sommerlade[a]

[a] Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Meston Building, Meston Walk, Aberdeen AB24 3UE, United Kingdom
[b] Institute of Physics and Astronomy, University of Potsdam, Campus Golm, Karl-Liebknecht-Straße 24/25, 14476 Potsdam-Golm, Germany

ARTICLE INFO

Keywords:
Network inference
Node degree distribution
False positive
False negative
Statistical inference

ABSTRACT

Background: A reliable inference of networks from data is of key interest in the Neurosciences. Several methods have been suggested in the literature to reliably determine links in a network. To decide about the presence of links, these techniques rely on statistical inference, typically controlling the number of false positives, paying little attention to false negatives.
New method: In this paper, by means of a comprehensive simulation study, we analyse the influence of false positive and false negative conclusions about the presence or absence of links in a network on the network topology. We show that different values to balance false positive and false negative conclusions about links should be used in order to reliably estimate network characteristics. We propose to run careful simulation studies prior to making potentially erroneous conclusion about the network topology.
Results: Our analysis shows that optimal values to balance false positive and false negative conclusions about links depend on the network topology and characteristic of interest.
Comparison with existing methods: Existing methods rely on a choice of the rate for false positive conclusions. They aim to be sure about individual links rather than the entire network. The rate of false negative conclusions is typically not investigated.
Conclusions: Our investigation shows that the balance of false positive and false negative conclusions about links in a network has to be tuned for any network topology that is to be estimated. Moreover, within the same network topology, the results are qualitatively the same for each network characteristic, but the actual values leading to reliable estimates of the characteristics are different.

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 (Bullmore and Sporns, 2009; Pessoa, 2014; Petersen and Sporns, 2015; Sporns et al., 2004). 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 (Newman, 2010, 2002), e.g., Erdös and Rényi (1959, 1960), Watts (1999), Watts and Strogatz (1998), Barabási and Albert (1999), Barabási and Pósfai (2016) networks. In our work we consider binary undirected networks of these three topologies.

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 (Newman, 2010, 2002). Some of the key characteristics are the node degree distribution, the number of links, the global clustering coefficient, and the efficiency. We considered these characteristics in our study since they are meaningful in random networks and give a global description in large networks (Newman, 2010).

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 (Mader et al., 2015; Zerenner et al., 2014) and (ii) the successful usage of the characteristics above to uniquely determine the topology of network (Bialonski et al., 2010, 2011).

The correct reconstruction of networks is hampered not only by false conclusions about links due to statistical uncertainties, but also by

* Corresponding author at: Room 343, Meston Building, Meston Walk, Aberdeen AB24 3UE, United Kingdom.
E-mail addresses: gloria.cecchini@abdn.ac.uk (G. Cecchini), m.thiel@abdn.ac.uk (M. Thiel), b.schelter@abdn.ac.uk (B. Schelter), l.sommerlade@abdn.ac.uk (L. Sommerlade).

https://doi.org/10.1016/j.jneumeth.2018.06.011
Received 19 February 2018; Received in revised form 19 June 2018; Accepted 19 June 2018
Available online 26 June 2018

unobserved processes (Elsegai et al., 2015; Guo et al., 2008; Ramb et al., 2013) and noise contamination (Nalatore et al., 2007; Newbold, 1978; Sommerlade et al., 2015) to name just a few challenges of network reconstruction. Classical statistical methods to estimate links in a network aim to identify present links with high certainty, e.g. Jalili and Knyazeva (2011), Quinn and Keough (2002), Devore (2011), Schinkel et al. (2011), De Vico Fallani et al. (2014), Chavez et al. (2010), Honey et al. (2007). Therefore, typically the rate of false positive conclusions about links is chosen and consequences for the rate of false negative conclusions about links are accepted. We investigate if these common rules of false positive conclusions and false negative conclusions should be modified to achieve a more reliable inference of the correct topology of network. To this aim, we analyse their influence on the network topology and characteristic.

The manuscript is structured as follows. An introduction to network topologies and their characteristics is given in Section 2.1. Section 2.2 explains statistical errors and their influence on the network topology. A simulation study in the case of Erdös–Rényi, Watts–Strogatz and Barabási–Albert networks is presented in Section 3.

2. Materials and methods

In this section, network topologies and their characteristics are described (Section 2.1). We summarise statistical errors and suggest a distance measure to quantify their influence on the estimation of network characteristics (Section 2.2).

2.1. Network characteristics

A network G is defined as a set of nodes with links between them. To quantify the topology of networks, different network characteristics have been described (Olbrich et al., 2010). Here, we consider four network characteristics: node degree, number of links, global clustering coefficient and efficiency. The node degree describes the number of links of a node. For example, if the node v has k links attached, its node degree is $d_v = k$. Typically the node degree distribution is used to characterise the entire network. The number of links refers to half of the sum over the node degrees.

The global clustering coefficient describes how well the neighbours of a node are connected. More precisely it measures the conditional probability that given one node connected to other two nodes, these are also connected to each other (Olbrich et al., 2010).

For two randomly selected nodes i, j in a network of n nodes, the shortest path length ℓ_{ij} measures the number of steps separating them if the shortest path is taken. The average path length $\gamma = \frac{1}{n(n-1)} \sum_{i \neq j} \ell_{ij}$ gives a measure of the sparsity of the network. The efficiency $\epsilon = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\ell_{ij}}$ is defined as the sum of the inverse of the shortest paths lengths. Since the shortest path is infinitely long for unconnected nodes, taking the average of the shortest path length in a network with unconnected nodes is not meaningful. Efficiency for unconnected nodes will be zero, therefore a meaningful network average of efficiency can be obtained.

Different network topologies have been described (Newman, 2003). Here, we investigate Erdös and Rényi (1959, 1960, Watts (1999), Watts and Strogatz (1998) and Barabási and Albert (1999) networks, as key examples of networks.

Erdös–Rényi networks are random networks in which each pair of nodes is connected with independent probability p_c. The probability mass function of the node degree distribution of a Erdös–Rényi network

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

is a binomial distribution (Newman, 2010).

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 with node degree $2c$. Nodes are arranged on a circle; therefore, each node has c nearest clockwise as well as c nearest counterclockwise neighbours. With probability p_r each link connecting a node to one of its nearest neighbours is reconnected to another node randomly chosen. The node degree distribution has probability mass function

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

in the assumption that $n \gg c$, (Menezes et al., 2017). Here, we study Watts–Strogatz networks with $c = 2$.

Barabási–Albert networks are so-called scale free networks. They are constructed by adding nodes to an existing network. The degree of the existing nodes influences the probability for a new link. Each new node is connected to the network with a certain number b of links. The probability for one of these b links to be formed with any existing node is proportional the degree of that node. The node degree distribution has probability mass function (Barabási and Pósfai, 2016)

$$\mathbb{P}(d_v = k) = \frac{2b(b+1)}{k(k+1)(k+2)}. \tag{3}$$

Note that Eq. (3) is of type $\mathbb{P}(d_v = k) = c_1 k^{-c_2}$, where c_1 and c_2 are constants, i.e., it follows a power law.

2.2. Inference reliability

In the Neurosciences different methods to identify nodes of a brain network exist. For our purposes the method of identifying nodes is not relevant. Therefore, we assume a fixed set of nodes. Once nodes have been fixed, 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 (Jalili and Knyazeva, 2011; Quinn and Keough, 2002; Devore, 2011; Schinkel et al., 2011; De Vico Fallani et al., 2014; Chavez et al., 2010; Honey et al., 2007).

Two types of errors exist when making these statistical inferences: (i) an absent link ($\overline{C}$) may be erroneously assumed to be present by the method (C^D), this is a false positive conclusion and referred to as a *type I error*; (ii) a present link (C) may remain undetected ($\overline{C^D}$) by the method, this is a false negative conclusion and referred to as a *type II error*. We call $\alpha = \mathbb{P}(C^D|\overline{C})$ the probability of a false positive conclusion and $\beta = \mathbb{P}(\overline{C^D}|C)$ the probability of a false negative conclusion. These two probabilities (α and β) are related and cannot be fixed independently. A standard choice is to set $\alpha = 0.05$ and neglect investigation of β, focussing on reliably detecting individual links of the network.

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 network we detect 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. We quantify the bias for each characteristic using a distance between distributions. Several distance measures are conceivable and have been investigated; for sake of simplicity and to make the arguments clearer, we only consider the distance

$$\delta = |\mu_1 - \mu_2| \tag{4}$$

between two distributions, 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 numerically we simulate