Iterative procedure for network inference

Gloria Cecchini[a,*], Björn Schelter[b]

[a]*CSDC, Department of Physics, University of Florence, via G. Sansone 1, 50019, Sesto Fiorentino, Florence, Italy*
[b]*Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Meston Building, Meston Walk, Aberdeen, AB24 3UE, United Kingdom*

Abstract

When a network is reconstructed from data, two types of errors can occur: false positive and false negative errors about the presence or absence of links. In this paper, the vertex degree distribution of the true underlying network is analytically reconstructed using an iterative procedure. Such procedure is based on the inferred network and estimates for the probabilities α and β of *type I* and *type II errors*, respectively. 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 vertex degree distribution tailored to its previously unknown network topology.

Keywords:
network inference, node degree distribution, false positive, false negative, statistical inference

1. Introduction

Complex systems are of key interest in multiple scientific fields, ranging from medicine via physics, mathematics, engineering to economics [1–3]. These systems can be modelled, or represented as networks, where vertices are the elements of the system and links represent the interactions between them.

Networks are ubiquitous in many fields of study [4]; 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 [5–15].

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 [16–19].

To capture particular features of a network, properties or characteristics have been introduced in the literature [20]. For instance, the vertex degree describes the number of links of a vertex [21]. The vertex degree distribution is a key property to study the general structure of the connectivity of a network, and it is predominantly used in this paper 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, such as the functional network of the human brain, 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 [22–24].

---

*Corresponding author
Email addresses:* gloria.cecchini@unifi.it (Gloria Cecchini), b.schelter@abdn.ac.uk (Björn Schelter)

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.

A prototypical class of systems that cannot be directly observed consists of dynamical systems where only the dynamic of the nodes is accessible. In this case, various strategies have been suggested in the literature to reconstruct the network structure [16–19, 22–27]. For the sake of simplicity, the reader can consider this kind of systems to be reconstructed; nevertheless, the method described in this paper is more general, and it can be applied once a measure of connectivity strength for each link is established.

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, i.e. the significance level of the test, is often chosen [28–34]. 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.

In [35] the authors describe 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. In the conclusions, the authors speculate that the simulation study presented could be used as an iterative procedure to achieve a better network reconstruction. This manuscript is dedicated to such a procedure. The results obtained in [36] 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.

This paper is structured as follows. Section 2 is dedicated to present some preliminary results needed for the implementation of the iterative procedure presented in this paper. In Sec. 3 such iterative procedure for network inference is presented. Section 4 shows a simulation study of the method depicted in Section 3.

2. Preliminary results

In this section we give some preliminary results that will be needed to implement the iterative procedure presented later in Sec. 3. In particular, Sec. 2.1 introduces some concepts of test of hypothesis and Sec. 2.2 is dedicated to summarise the results obtained in [36].

2.1. Type I and type II errors

In [35], 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 vertex. Various strategies have been suggested in the literature to recover a measure for the strength of connection for every link [16–19]. For instance, correlation can be used as a measure to this aim, and it will be employed throughout this manuscript. Therefore, assume to assign a correlation value to each couple of nodes, and appoint it to the respective link.

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

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

where Γ is the gamma function and ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the Gaussian hypergeometric function.

When performing a hypothesis test for correlation for every pair of vertices, a $p-$value of the test is found. If the $p-$value is smaller than or equal to the significance level α, then the link between the corresponding vertices is considered to be present. Call r_τ the value of the correlation r such that

$$\alpha = \int_{r_\tau}^1 f(N,0,r)\,dr, \tag{2}$$

meaning that the rejection region is obtained for coefficients $r_\tau \le r \le 1$. Consequently, estimated coefficients $r < r_\tau$ do not lead to a rejection of the null hypothesis. The probability of false negative conclusions

$$\beta = \int_{-1}^{r_\tau} f(N,\rho,r)\,dr \tag{3}$$

is calculated once r_τ is fixed.

Once α is chosen,

$$\beta = \beta(N,\rho,\alpha) \tag{4}$$

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

In the case of network reconstructions, after the value of correlation is estimated for each link, a threshold r_τ is chosen, and α and β are inferred. This is equivalent to infer $f(N,0,r)$ and $f(N,\rho,r)$ from the aggregated distribution of all the estimated correlation values. The detected distribution of all the estimated correlation values is indeed the weighted sum of $f(N,0,r)$ and $f(N,\rho,r)$, for the absent and true links in the underlying network. As an example, Fig. 1 shows the detected distribution in red dotted line, the distributions $f(N,0,r)$ and $f(N,\rho,r)$ in solid blue and dashed black lines, respectively. The dark and light blue areas correspond to α and β, respectively, for the fixed value of r_τ. Given the rate p^* of true links with respect to the total, the detected distribution is given by $(1-p^*)f(N,0,r) + p^*f(N,\rho,r)$.

In a real-world application, an estimate for the distribution of the 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 manuscript assumes that these correlation coefficients are independent and identically distributed (iid) and follow the distribution $f(N,\rho,r)$ in Eq. (1). In this case, it is reasonable to assume that it is possible to estimate ρ, given the distribution of the correlation coefficients.

A more complicated scenarios consists of correlation coefficients which are still independent, but follow more than one distributions $f(N,\rho,r)$, i.e., there is not a unique value for ρ. In this case, it still should be possible to estimate the distributions $f(N,\rho,r)$, and therefore apply the method developed in this paper, using different values for β, as shown in Eq. 3. A generalization for correlation coefficients which are not independent is likely more complicated, and therefore requires a more in-depth analysis.

2.2. Analytical framework on network inference

The work presented in [36] 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. Here we summarise the main results that will be used later for the iterative procedure.

Consider a network G with n vertices and vertex degree distribution defined by the probability function $\mathcal{P}$. Call G' the network detected when *type I and type II errors* occur; as mentioned above, α is the probability of a *type I error* and β is the probability of a *type II error*. 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 $\mathcal{P}'$.
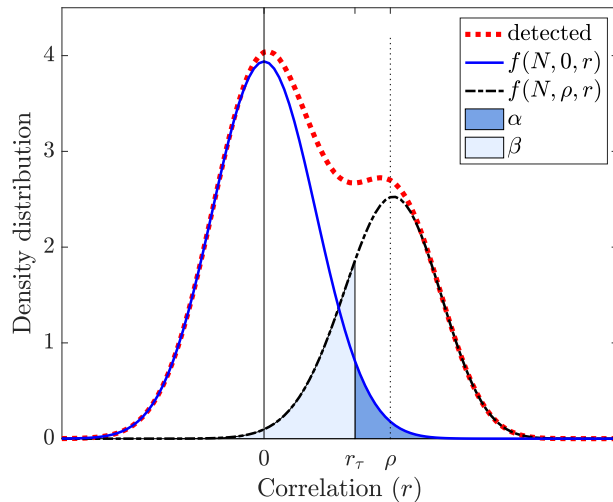
Figure 1: Density distribution of correlation coefficients r for a sample of $N = 100$ data points with true correlation coefficient $\rho = 0$ (solid blue line) and $\rho = 0.25$ (dashed black line), i.e. $f(100, 0, r)$ and $f(100, 0.25, r)$ as described by Eq. (1). The aggregated detected distribution (red dotted line) is given by the weighted sum $(1 - p^*)f(N, 0, r) + p^*f(N, \rho, r)$. 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.

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

$$\mathbb{P}(d' = k' | d = k) = \sum_{i=\max\{0, k-k'\}}^{\min\{k, n-1-k'\}} \binom{k}{i}(1 - \beta)^{k-i}\beta^i \binom{n-1-k}{k'-k+i}\alpha^{k'-k+i}(1 - \alpha)^{n-1-k'-i}, \tag{5}$$

which corresponds to Eq. (A1) in [36]. Hence, calling $\mathcal{P}_i = \mathbb{P}(d = i)$ and $\mathcal{P}'_i = \mathbb{P}(d' = i)$, Eq. (5) can be written in a matrix form as

$$\mathcal{P}' = A\mathcal{P}, \tag{6}$$

where A is the matrix with elements the conditional probabilities $A_{k+1, k'+1} = \mathbb{P}(d' = k' | d = k)$; note that $k, k' \in \{0, \cdots, n-1\}$. Equation (6) corresponds to Eq. (7) in [36]. The matrix $A = A(n, \alpha, \beta)$ depends on n, α and β and has determinant $\det A = (1 - \alpha - \beta)^{\frac{n(n-1)}{2}}$, then it is invertible if and only if $\alpha \neq 1 - \beta$. Note that, since the convergence to zero of the determinant of A scales like $x^{\frac{n(n-1)}{2}}$ for $|x| < 1$, numerical issues arise for relatively small n when inverting the matrix A to find $\mathcal{P}$ through $\mathcal{P} = A^{-1}\mathcal{P}'$.

The vertex degree distribution of the original network is found carefully inverting Eq. (6). Call A_t^+ the pseudoinverse of the truncated matrix A using the singular value decomposition, i.e., only the first t singular values are considered different from zero. A method to assess optimal value for t has been studied in [38, 39], but it still remains an open problem. In this paper, the choice for t lies on the difference between consecutive singular values, i.e., t is the largest index for the singular values σ_i such that $\sigma_t - \sigma_{t+1} \geq \sigma_i - \sigma_{i+1}$ for all i. The vertex degree distribution of the original network

$$\mathcal{P} = A_t^+ \, \mathcal{P}' \tag{7}$$

is thus calculated analytically if the biased one and the probabilities of *type I* and *type II errors* are given. Note that the matrix A depends on n, α, β, where n is the number of vertices in the network.

As stated above, this reconstruction method assumes the probabilities of *type I* and *type II errors* to be known a priori. To show the impact of violations of this and thereby the robustness of the method, the performance of the reconstruction is analysed when perturbations on α and β are introduced. The perturbations are quantified in percentage using the parameter δ, e.g. the perturbations of β are expressed
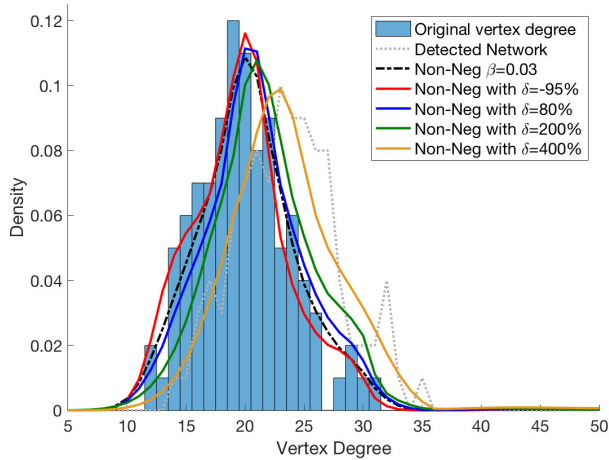
4

Figure 2: 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.

by $\beta^p = \beta + \delta\beta$. 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.

The robustness analysis shows 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. In [36] various example are presented to demonstrate the robustness of the approach. Here we present one of those cases to briefly explain the robustness analysis which will be used later for the iterative procedure in Sec. 3.

Figure 2 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. Note that, since $0 \le \beta \le 1$, the conditions for the perturbations are $-1 \le \delta \le 1/\beta - 1$. Therefore, the perturbation of $\delta = -95\%$ is almost the most extreme value can be chosen in the negative direction. The cut-off for the truncated singular value decomposition method is 0.55 and the probability of *type I error* is $\alpha = 0.05$. The method performs to be robust to large perturbations of β, in both negative and positive directions. Up to $\delta = 80\%$ in the positive direction, the bias of the reconstruction is negligible; only if $\delta = 400\%$ or more, the reconstruction deviates significantly from the true one, although it still performs better than the naïve approach of trusting the identified network structure.

As stated in the concluding section of [36], 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 paper explores the way to adjust the estimates of these two errors, so to improve the reconstruction of the network of interest.

3. Iterative procedure

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 a value of correlation is calculated for each link, the aggregated distribution of all correlation values $(1-p^*)f(N, 0, r) + p^* f(N, \rho, r)$ is found, where p^* is the rate of true links over the total. The aggregated distribution is used to estimate the distributions $f(N, 0, r)$ and $f(N, \rho, r)$. Fix the initial value for α to 0.05 and find r_τ using Eq. (2). All the links with corresponding correlation value larger than r_τ are considered to be present. The resulting network is the
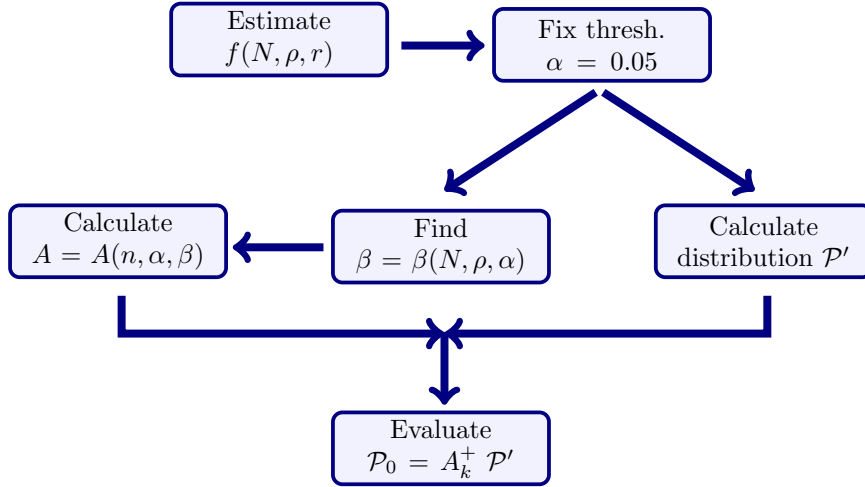
Table 1: Network Inference: procedure scheme. After a value of correlation is calculated for each link, the aggregated distribution is used to estimate the distributions $f(N, 0, r)$ and $f(N, \rho, r)$. Fix the initial value for α to 0.05, and find r_τ using Eq. (2). All the links with corresponding correlation value larger than r_τ are considered to be present. 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. (4), therefore also A is computed. Knowing $\mathcal{P}'$ and A, the probability density function $\mathcal{P}_0$ of the original vertex degree distribution is calculated using Eq. (7).

so-called biased network, and therefore the vertex degree distribution

$$\mathcal{P}'_i = \frac{\text{number of vertices with degree} = i}{\text{total number of vertices}} \tag{8}$$

can be calculated empirically by counting the vertices' degrees. 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. (7) are now available, hence the vertex degree distribution $\mathcal{P}_0$ of the original network is calculated, and this concludes the first iteration step. Table 1 summarises the procedure of the first iteration step.

Section 2.2 shows that wrong estimates of α and β, within certain bounds, do not cause the reconstruction $\mathcal{P} = A_k^+ \mathcal{P}'$ [Eq. (7)] of be rendered invalid. Note that the robustness of 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 1, can be enhanced using this idea. 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. (4)]. 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 2.

The robustness analysis consists of calculating several different biased networks varying α, starting from the vertex degree distribution $\mathcal{P}_0$, i.e., $\mathcal{P}'(\alpha) = A(n, \alpha, \beta) \mathcal{P}_0$. For each $\mathcal{P}'(\alpha)$, a robustness analysis is performed, as explained in Sec. 2.2. Namely, call $\mathcal{P}_0^*(\alpha, \delta)$ the density of the reconstructed vertex degree distribution, using as biased vertex degree distribution $\mathcal{P}'(\alpha)$, and a perturbation δ on α.

The value of α, that gives the most robust result, is chosen analysing the difference between the distributions $\mathcal{P}_0$ and $\mathcal{P}_0^*(\alpha, \delta)$. To quantify the bias between these distributions, 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 to perform a test of hypothesis. The Kolmogorov-Smirnov distance is the largest difference between the two cumulative distributions [33], i.e.,

$$\mathcal{D}_{\alpha,\delta} = \mathcal{D}(\mathcal{F}_0^*(\alpha, \delta), \mathcal{F}_0) = \max |\mathcal{F}_0^*(\alpha, \delta) - \mathcal{F}_0|, \tag{9}$$
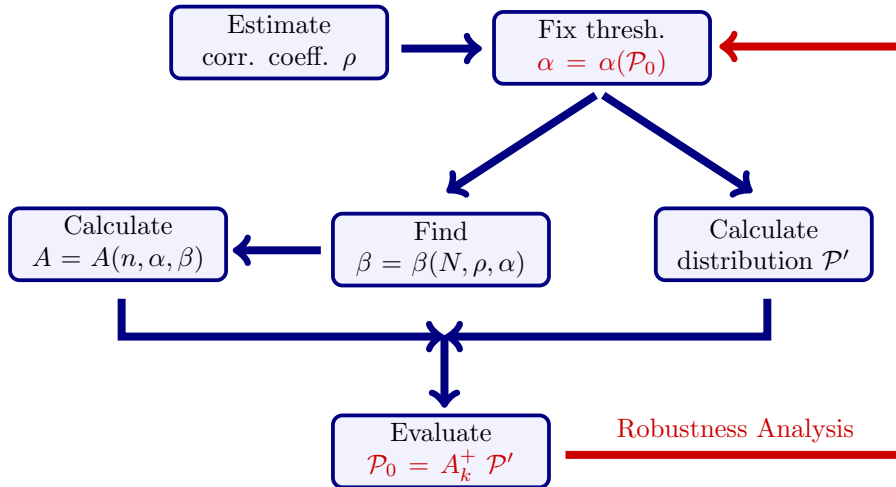
Table 2: Network Inference: procedure scheme with iteration. After a value of correlation is calculated for each link, the aggregated distribution is used to estimate the distributions $f(N, 0, r)$ and $f(N, \rho, r)$. Fix the value for α according to the result of the previous iteration step, and find r_τ using Eq. (2). All the links with corresponding correlation value larger than r_τ are considered to be present. 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. (4), therefore also A is computed. Knowing $\mathcal{P}'$ and A, the probability density function $\mathcal{P}_0$ of the original vertex degree distribution is calculated using Eq. (7). Perform a robustness analysis to iterate the procedure.

is the distance between $\mathcal{P}_0$ and $\mathcal{P}_0^*(\alpha, \delta)$, where $\mathcal{F}_0$ and $\mathcal{F}_0^*(\alpha, \delta)$ are their respective cumulative distributions. In mathematics, defined on a vector space, this distance is called Tchebychev distance, or maximum metric.

Once the distance $\mathcal{D}_{\alpha,\delta}$ between the two distributions is calculated, the value for α which gives the shortest distance varying δ is chosen. This approach follows the minimum distance estimation method developed in [40–44].

4. 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 $n = 100$ vertices and probability of connection $p = 0.2$. For the sake of simplicity, assume this is a dynamical system, and the dynamics of the nodes are accessible. Performing a correlation analysis for every pair of vertices, values for the correlation are found. As explained in Sec. 2.1, it is reasonable to assume that correlation values corresponding to connected nodes follow the density distribution $f(N, \rho, r)$ [Eq. (1)]. Therefore, instead of simulating the dynamics of the nodes, a value for the correlation is assigned at random with density distribution $f(N, \rho, r)$ to each link, with $\rho = 0.25$ and the number of data points $N = 100$. For each absent link the value for the correlation is chosen at random with distribution $f(N, 0, r)$. The aggregated distribution of all correlation values is used to estimate the true correlation coefficient ρ.

In real world applications, the expectation maximization algorithm for Gaussian mixture models is often used to estimate two distributions given the overall observed data. This algorithm is an iterative procedure to find maximum likelihood estimates of the parameters in the model. The expectation maximization algorithm for a Gaussian mixture model is used in this paper to estimate the distributions for true and absent links of the correlation coefficients. The choice of fitting normal distributions, instead of the exact distribution equation [Eq. (1)], is motivated by the fact that we want to keep our approach as general as possible. Figure 3a shows the histogram of all correlation values, and the estimated distributions, using the expectation maximization algorithm for Gaussian mixture models.
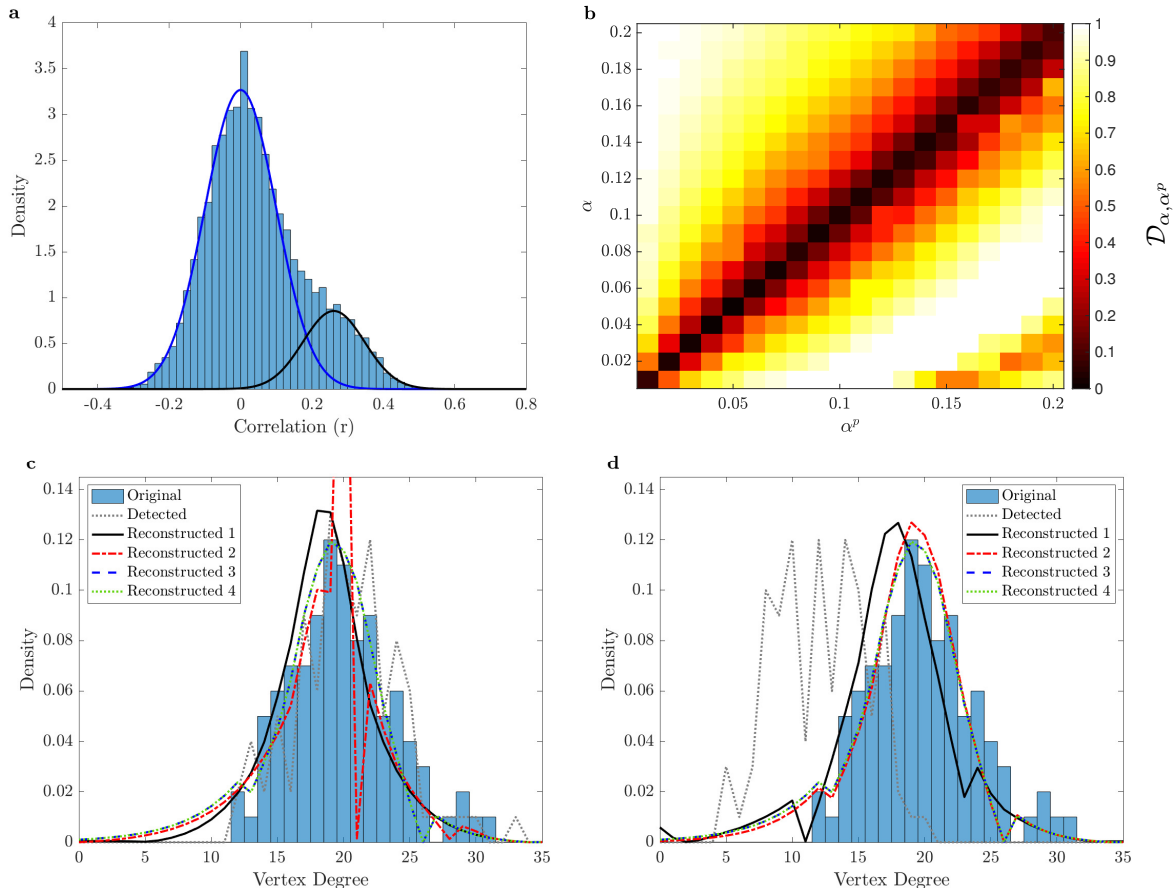
Figure 3: Reconstruction of an Erdős-Rényi network with $n = 100$ vertices and probability of connection 0.2. a) Histogram of all correlation values, and the estimated distributions, using the expectation maximization algorithm for Gaussian mixture models. b) Kolmogorov-Smirnov distance $\mathcal{D}_{\alpha,\alpha^p}$ obtained from $\mathcal{D}_{\alpha,\delta}$, where $\alpha^p = \alpha + \delta\alpha$ (x-axis). c) Results of four iteration steps, starting with $\alpha = 0.05$. d) Results of four iteration steps, starting with $\alpha = 0.01$.

Once the distributions for true and absent links are estimated, and α is set to 0.05, all the correlation values above the associated threshold r_τ are selected; the corresponding links are considered to be present, and the biased, or detected, network is obtained. The procedure explained in Sec. 3 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 $\mathcal{P}_0^*(\alpha,\delta)$ of the vertex degree distributions are calculated. Evaluate the Kolmogorov-Smirnov distance $\mathcal{D}_{\alpha,\delta}$ using Eq. (9). For the sake of simplicity in the representation of the results, the value of the perturbed α, i.e., $\alpha^p = \alpha + \delta\alpha$ is used instead of δ, and the distance $\mathcal{D}_{\alpha,\alpha^p}$ is plotted, as in Fig. 3b.

Once $\mathcal{D}_{\alpha,\delta}$ is evaluated, the value for α which gives the shortest distance varying δ is chosen. Namely, the aim is to minimise the bias of the reconstruction in a neighbourhood of α. For the same values of the perturbation δ, the distance $\mathcal{D}_{\alpha,\delta}$ gives different results, depending on α. This means that there exist some values for α which give smaller values for the overall distance $\mathcal{D}_{\alpha,\delta}$ in a neighbourhood of α. Once the radius of the neighbourhood is chosen, e.g. $-0.2 < \delta < 0.2$, the overall bias of the neighbourhood is calculated as $\sum_{-0.2 < \delta < 0.2}(1 - \delta)\mathcal{D}_{\alpha,\delta}$, the value for α which gives the smallest result is chosen, and the procedure is iterated.

Figure 3c shows the results of four iteration steps. As shown in 2.1, 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 last two iteration
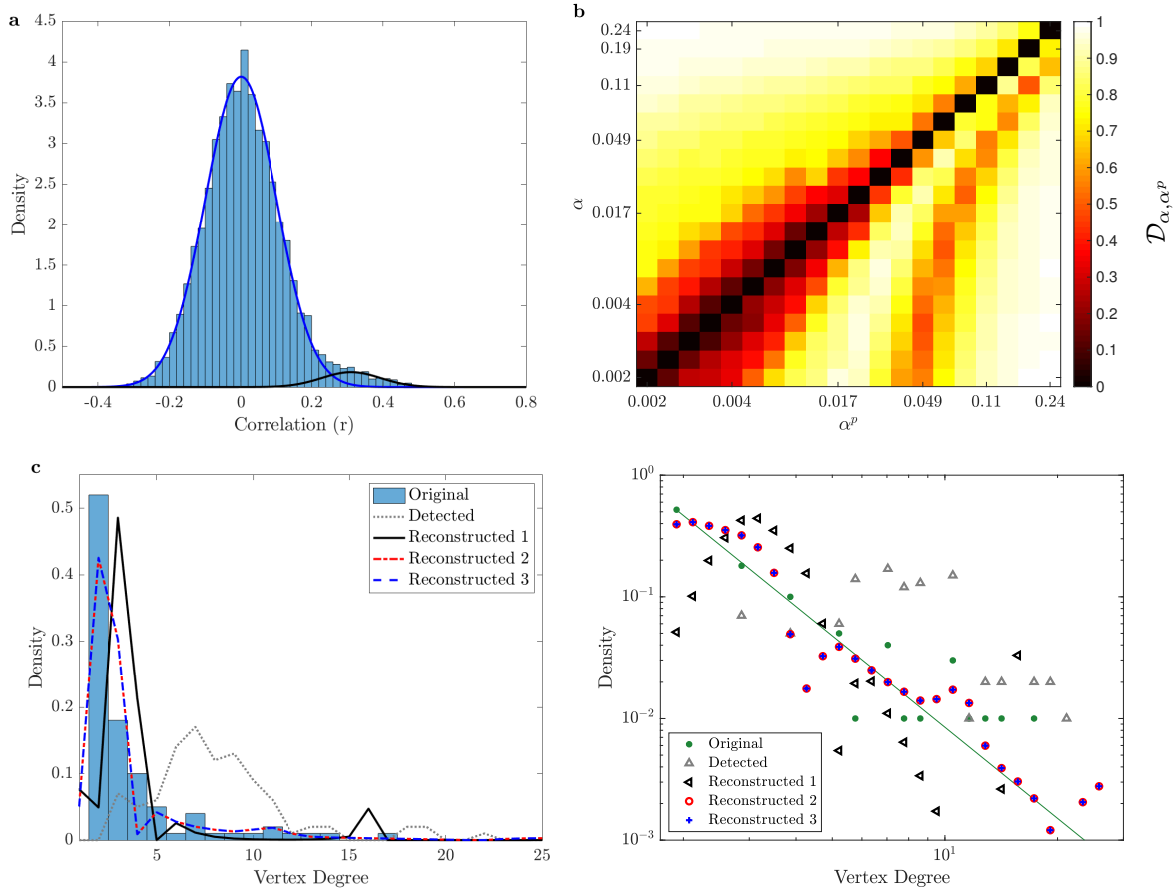
Figure 4: Reconstruction of a scale-free network with $n = 100$ vertices. a) Histogram of all correlation values, and the estimated distributions, using the expectation maximization algorithm for Gaussian mixture models. b) Kolmogorov-Smirnov distance $\mathcal{D}_{\alpha,\alpha^p}$ obtained from $\mathcal{D}_{\alpha,\delta}$, where $\alpha^p = \alpha + \delta\alpha$ (x-axis). c) Results of three iteration steps with initial $\alpha = 0.05$. d) Results of three iteration steps presented in a log-log plot.

steps demonstrate that the process converges, and once the correct reconstruction is achieved, the following results do not deviate from it.

Figure 3d shows the results of four iteration steps, when the initial value of α is set to 0.01 (instead of 0.05). In this case, the detected vertex degree distribution deviates considerably from the original one, implying that the naïve approach would lead to a wrong estimate of the vertex degree distribution. The result of the first iteration improves the result, reducing the bias and getting the correct shape of the distribution. At the other iteration steps, the bias is completely corrected. Note that the last two iteration steps give the same reconstructed distribution, this means that the procedure converged to a solution and the following results do not deviate from it.

The method described in the previous section is now applied to a Scale-Free network of 100 vertices. Values for the correlation are assigned at random with density distribution $f(N, \rho, r)$ to each link, with $\rho = 0.3$ and the number of data points $N = 100$. Figure 4a shows the histogram of all correlation values, and the estimated distributions, using the expectation maximization algorithm for Gaussian mixture models. Figure 4b depicts 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. A non-linear x- and y-scale for α is here used to emphasize the role of small values for α on the distance $\mathcal{D}_{\alpha,\alpha^p}$. Note that this is in agreement with the results shown in Fig. 5 in [35]; in

9

that case, 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 4c shows the results of three iteration steps, when the initial value of α is set to 0.05. Also in this case, the last two iteration steps give the same reconstructed distribution, meaning that the procedure converged to a solution and the following results do not deviate from it. Figure 4d presents the plot in Fig. 4c in log-log scale to demonstrate that the scale-freeness property of the network has been correctly recovered.

Additionally, simulations with both larger and smaller number of nodes have been performed. Results are qualitatively the same as the ones presented in this section. Likewise, for different values of ρ the method performs similarly. However, for very small values of ρ, the aggregated distribution of all correlation values does not carry any relevant information about the distributions of true and absent links. Namely, the distribution of the true links is shifted towards zero, and therefore it is superposed to the other one. Consequently, the expectation maximization algorithm for Gaussian mixture models is not accurate in distinguishing the two distributions for the correlation coefficients. Hence, the correct reconstruction of the network is not guaranteed. In this case, i.e., when the aggregated distribution of all correlation values does not carry any relevant information, different measures for the strength of connection should be used instead of correlation.

5. Conclusion

The analysis presented in [36] is advanced in this manuscript, and an iterative procedure to reconstruct the vertex degree distribution is suggested. This procedure should be used when reconstructing a network from data to guarantee that the vertex degree distribution is correctly recovered. In particular, when the estimates of α and β might not be accurate and large errors on these estimates might occur, the method described in this manuscript ensures a robust and better reconstruction than other methods described previously.

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.

The method presented in this manuscript is a general procedure that can be applied to any kind of network. Moreover, this iterative procedure does not depend on the dynamic of the nodes, nor on the dimension of the network. Throughout the paper correlation has been used as a measure of connectivity strength for each link; nevertheless, the method works regardless the kind of measure chosen to establish the strength of connection. This implies that different types of measures can be used instead of correlation, broadening the applicability of the procedure described in this work.

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. Furthermore, with the aim to make this procedure even more general, the case in which the inference of the distributions for the correlation coefficients is not accurate should be carefully studied.

Note that the method presented in this manuscript aims to reconstruct the vertex degree distribution. A generalization for other characteristics is likely more complicated, and therefore requires a more in-depth analysis. Nonetheless, this further investigation could lead to reconstructing the network itself, i.e., link by link estimation. Namely, once the functional relationship between the reconstructed and true underlying network is found in terms of vertex degree distribution and eventually also, e.g., the shortest path length and the clustering coefficient, the true underlying network should be inferred combining all these results. A step further consists of creating a whole new procedure for network inference. Firstly, the vertex degree distribution and eventually other characteristics of the original network are estimated. Secondly, the links in the network which have zero probability of being false positive or false negative are identified. Such links are fixed, and the method aims to reconstruct the remaining links matching the characteristics found at the previous step.

Acknowledgements

References

[1] B. A., B. M., V. A., Dynamical Processes on Complex Networks, Cambridge University Press, Cambridge, 2008.

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424. doi:https://doi.org/10.1016/j.physrep.2005.10.009.

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

[4] G. Fagiolo, Clustering in complex directed networks, Phys. Rev. E 76. doi:10.1103/PhysRevE.76.026107.

[5] R. Albert, A. L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. 74. doi:10.1103/RevModPhys.74.47.

[6] J. R. Banavar, F. Colaiori, A. Flammini, A. Maritan, A. Rinaldo, Topology of the fittest transportation network, Phys. Rev. Lett. 84. doi:10.1103/PhysRevLett.84.4745.

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

[8] S. N. Dorogovtsev, J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW, Oxford University Press, Oxford, 2003. doi:10.1093/acprof:oso/9780198515906.001.0001.

[9] V. M. Eguíluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, A. V. Apkarian, Scale-free brain functional networks, Phys. Rev. Lett. 94. doi:10.1103/PhysRevLett.94.018102.

[10] M. Kurant, P. Thiran, Extraction and analysis of traffic and topologies of transportation networks, Phys. Rev. E 74. doi:10.1103/PhysRevE.74.036114.

[11] M. E. J. Newman, The structure and function of complex networks, SIAM Rev. 45. arXiv:http://dx.doi.org/10.1137/S003614450342480, doi:10.1137/S003614450342480.

[12] G. Ódor, B. Hartmann, Heterogeneity effects in power grid network models, Phys. Rev. E 98. doi:10.1103/PhysRevE.98.022305.

[13] M. Valencia, J. Martinerie, S. Dupont, M. Chavez, Dynamic small-world behavior in functional brain networks unveiled by an event-related networks approach, Phys. Rev. E 77. doi:10.1103/PhysRevE.77.050905.

[14] Y. Xue, J. Wang, L. Li, D. He, B. Hu, Optimizing transport efficiency on scale-free networks through assortative or dissortative topology, Phys. Rev. E 81. doi:10.1103/PhysRevE.81.037101.

[15] C. H. Yeung, K. Y. M. Wong, Phase transitions in transportation networks with nonlinearities, Phys. Rev. E 80. doi:10.1103/PhysRevE.80.021102.

[16] R. Cestnik, M. Rosenblum, Reconstructing networks of pulse-coupled oscillators from spike trains, Phys. Rev. E 96. doi:10.1103/PhysRevE.96.012209.

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

[18] S. Li, F. Li, W. Liu, M. Zhan, Network reconstruction by linear dynamics, Physica A 404.

[19] A. Pikovsky, Reconstruction of a neural network from a time series of firing rates, Phys. Rev. E 93. doi:10.1103/PhysRevE.93.062313.

[20] E. Olbrich, T. Kahle, N. Bertschinger, N. Ay, J. Jost, Quantifying structure in networks, Eur. Phys. J. B 77. doi:10.1140/epjb/e2010-00209-0.

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

[22] M. Asllani, T. Carletti, F. Di Patti, D. Fanelli, F. Piazza, Hopping in the crowd to unveil network topology, Phys. Rev. Lett. 120 (2018) 158301. doi:10.1103/PhysRevLett.120.158301.

[23] R. Burioni, M. Casartelli, M. di Volo, R. Livi, A. Vezzani, Average synaptic activity and neural networks topology: a global inverse problem, Sci. Rep. 4 (1) (2014) 4336. doi:10.1038/srep04336.

[24] S. G. Shandilya, M. Timme, Inferring network topology from complex dynamics, New J. Phys. 13 (1) (2011) 013004. doi:10.1088/1367-2630/13/1/013004.

[25] M. G. Leguia, Z. Levnaji, L. Todorovski, B. enko, Reconstructing dynamical networks via feature ranking, Chaos 29 (9) (2019) 093107. doi:10.1063/1.5092170.

[26] A. Banerjee, J. Pathak, R. Roy, J. G. Restrepo, E. Ott, Using machine learning to assess short term causal dependence and infer network links, Chaos 29 (12) (2019) 121104. doi:10.1063/1.5134845.

[27] M. J. Panaggio, M.-V. Ciocanel, L. Lazarus, C. M. Topaz, B. Xu, Model reconstruction from temporal data for coupled oscillator networks, Chaos 29 (10) (2019) 103116. doi:10.1063/1.5120784.

[28] M. Chavez, M. Valencia, V. Latora, J. Martinerie, Complex networks: new trends for the analysis of brain connectivity, Int. J. Bifurcat. Chaos 20. doi:10.1142/S0218127410026757.

[29] F. De Vico Fallani, J. Richiardi, M. Chavez, S. Achard, Graph analysis of functional brain networks: practical issues in translational neuroscience, Philos. T. R. Soc. B 369. doi:10.1098/rstb.2013.0521.

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

[31] C. J. Honey, R. Kötter, M. Breakspear, O. Sporns, Network structure of cerebral cortex shapes functional connectivity on multiple time scales, Proc. Natl. Acad. Sci. 104. doi:10.1073/pnas.0701519104.

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

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

[34] S. Schinkel, G. Zamora-López, O. Dimigen, W. Sommer, J. Kurths, Functional network analysis reveals differences in the semantic priming task, J. Neurosci. Meth. 197. doi:https://doi.org/10.1016/j.jneumeth.2011.02.018.

[35] G. Cecchini, M. Thiel, B. Schelter, 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. doi:https://doi.org/10.1016/j.jneumeth.2018.06.011.

[36] G. Cecchini, B. Schelter, Analytical approach to network inference: Investigating degree distribution, Phys. Rev. E 98. doi:10.1103/PhysRevE.98.022311.

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

[38] M. Frank, J. M. Buhmann, Selecting the rank of truncated svd by maximum approximation capacityarXiv:arXiv:1102.3176v3.

[39] M. Gavish, D. L. Donoho, The optimal hard threshold for singular values is $4/\sqrt{3}$, IEEE T. Inform. Theory 60. doi:10.1109/TIT.2014.2323359.

[40] C. A. Drossos, A. N. Philippou, A note on minimum distance estimates, Ann. I. Stat. Math. 32 (1) (1980) 121–123. doi:10.1007/BF02480318.

[41] W. C. Parr, W. R. Schucany, Minimum distance and robust estimation, J. Am. Stat. Assoc. 75 (371) (1980) 616–624. doi:10.1080/01621459.1980.10477522.

[42] J. Wolfowitz, Consistent estimators of the parameters of a linear structural relation, Scand. Actuar. J. 1952 (3-4) (1952) 132–151. doi:10.1080/03461238.1955.10430689.

[43] K. Matusita, On the estimation by the minimum distance method, Ann. I. Stat. Math. 5 (2) (1953) 59–65. doi:10.1007/BF02949801.

[44] J. Wolfowitz, The minimum distance method, Ann. Math. Statist. 28 (1) (1957) 75–88. doi:10.1214/aoms/1177707038.