Data reliability in complex directed networks

Joaquín Sanz[1,2], Emanuele Cozzo[1,2] and Yamir Moreno[1,2,3]

[1] Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, Zaragoza 50018, Spain.

[2] Department of Theoretical Physics, Faculty of Sciences, University of Zaragoza, Zaragoza 50009, Spain.

[3] Complex Networks and Systems Lagrange Lab, Institute for Scientific Interchange, Turin, Italy

E-mail: jsanz@bifi.es

Abstract.

The availability of data from many different sources and fields of science has made it possible to map out an increasing number of networks of contacts and interactions. However, quantifying how reliable these data are remains an open problem. From Biology to Sociology and Economy, the identification of false and missing positives has become a problem that calls for a solution. In this work we extend one of newest, best performing models -due to Guimerá and Sales-Pardo in 2009- to directed networks. The new methodology is able to identify missing and spurious directed interactions with more precision than previous approaches, which renders it particularly useful to analyze data reliability in systems like trophic webs, gene regulatory networks, communication patterns and several social systems. We also show, using real-world networks, how the method can be employed to help searching for new interactions in an efficient way.

Keywords: Regulatory networks (Theory), network reconstruction, statistical inference.

1. Introduction

The last several years have witnessed many advances in what is today known as network science. Although the study of networks is not new, the availability of data in many different fields, ranging from techno-social systems to biological networks, has paved the way to solve relevant questions that were not accessible just a few years ago due to the lack of relevant data. It is however evident that we have advanced less in some fundamental questions, already put forward back in 2001 [1]. One of such challenges is to understand how models, methods and results of networks theory change when one consider different kinds of links: directed or undirected, weighted or unweighted. Though there are many good examples of real networks than can be easily treated as undirected [2], probably there is an even larger number of systems in which links' directionality and/or weights make a difference. These systems include gene regulatory

networks [3, 4], food webs [5] or some interaction networks extracted from social media communication patterns [6, 7].

On the other hand, the lack of data quality and complete information about interactions is an ubiquitous problem in most research areas where the framework of network modeling is applied. For example, classical social survey methods must deal with problems like sampling biases [8], or data loss [9, 10], which can compromise network-level analyses. The problem is even more acute when moving from social to biological systems like transcriptional regulatory maps, in which the promise of high-throughput biochemical techniques of revealing the system backbone (i.e., transcriptomes) has to deal with the inaccuracy that these methods often show. Microarray essays −the main tool to quantitatively measure the activity of large amounts of genes in a highly parallel fashion− constitute a paradigmatic example of a powerful, but sometimes inaccurate or hardly reproducible technique [11–13].

Focusing on the subfield of gene regulatory networks, one additional limitation to the network approach is the diversity -even conceptual- of the high number of different techniques used to infer regulatory interactions [14–16]. Lastly, the most important issue is probably the fact that the environmental conditions under which regulatory interactions take place are, in general, different for each interaction, and for a high proportion of cases only roughly known. This leads to the paradox that in many cases, reported regulations [14, 15, 17] identified through very diverse experimental techniques, and under specific experimental conditions, are rarely similar when links identified through different experiments are compared.

It is then of utmost importance to develop new ways to assess data reliability in complex directed networks, specially because most of the efforts up to now have been directed towards solving the problem in undirected, unweighted graphs [18–21], and there are only few works devoted to address the problem in directed systems [22, 23]. In this paper, we capitalize on a previous method proposed to study the very same problem but for undirected systems [24]. Specifically, we generalize the method proposed by Guimera & Sales-Pardo [24] to the case in which links are directed, like in a regulatory network. By doing so, we are able to successfully identify missing and spurious interactions in several real-world networks.

By comparing the performance of our method to that of previous approaches [22,23] dealing with directed networks, we obtain, with some exceptions, better results at predicting missing and spurious interactions, paying the prize of greater computational requirements. This exhaustive comparison allows us to plot a general outlook of the problem of data reliability determination on directed networks, identifying the strengths and weakness of each method. Finally, we test whether the methods can be used to predict new links in a genome-wide transcriptional regulatory network [14], providing a robust methodology that could help and guide the experimental search for unnoticed regulations. Our results indicate that the approach proposed here is also the best performing method when facing this kind of situation.

2. Results

2.1. Stochastic block models for reliability determination in directed networks

Following [24], let us suppose that we are working on a certain graph whose adjacency matrix is A^o, which is just an imperfect realization of a certainly ideal, "true" network A to which we have no access. Being X a certain measurable property of the network, we will call $p(X = x|A^o)$ the probability that, once observed the graph A^o, X is equal to x in the ideal system A. Then we have:

$$p(X = x|A^O) = \int_M p(X = x|m)p(m|A^O)dm \tag{1}$$

where $p(m|A^O)$ is the probability that m is the model in a class M that gave the observation A^O, and $p(X = x|m)$ stands for the probability of model m to generate networks in which $X = x$. It is worth noticing that, depending on the way the family of models M is defined, eq. 1 can adopt the form of a sum instead of an integral. Since the term $p(m|A^O)$ is certainly difficult to estimate, we must reformulate the problem by using the Bayes theorem, to get:

$$p(X = x|A^O) = \frac{\int_M p(X = x|m)p(A^O|m)p(m)dm}{\int_{M'} p(A^O|m')p(m')dm'} \tag{2}$$

where $p(A^0|m)$ is the probability that m gave A^O among all possible adjacency matrices and $p(m)$ is the a priori probability of model m.

At this point, we need to select a class of models to integrate the former expression. The main hypothesis that lies beneath this method consists of assuming that the required family is that of stochastic-blocks-models (SBM). In the case of undirected networks, any of these SBM can be characterized by a partition P of the set of nodes into blocks, and a probability matrix $\mathbf{Q}$ such that the element $Q_{\alpha,\beta}$ defines the probability that any of the nodes belonging to the block α be connected to any of the nodes within block β. So, the probability of two nodes being connected depends only on the blocks these nodes belong to within the partition P. Note that under these assumptions, $\mathbf{Q}$ is symmetric.

In order to deal with directed networks several possibilities are conceptually feasible. Here, we propose the following variation of the model. Instead of considering one single partition P of the nodes' space, we will consider two partitions, a senders partition P_s and a receivers partition P_r. Every node i must then belong, independently, to a block in each partition: $i \in \sigma_i$ with $\sigma_i \in P_s$ and $i \in \rho_i$ with $\rho_i \in P_r$, as it is sketched in figure 1. The partitions just take into account the fact that in directed networks, out-going and incoming links are treated separately. Thus out-going links of node i will be determined by block σ_i to which it belongs in the partition P_s. On its turn, and in an independent way, the in-degree will be given by the block ρ_i in the other partition P_r in which the node i is located. Within this scheme, the probability of node i sending a link to node j is Q_{σ_i,ρ_j}. Remarkably, the probability of observing the opposite link is different, and equal to Q_{σ_j,ρ_i}.

This scheme, yet having the virtue of its computational tractability, conceptually captures the behavior of systems like transcriptional regulatory networks in which the statistics associated to in-degrees are very different to those regarding out-degrees [14,15], being both relatively uncorrelated. This can be easily understood if one considers that the biochemical properties that define the susceptibility of a protein to be regulated by others are different to those that make the protein a regulator. While the information that will ultimately define the identity and the strength of the transcriptional regulations affecting a protein reside in its promoter region, its eventual ability to bind to the promoters of other target proteins depends on the presence and identity of a regulator domain within its protein sequence. Consequently, these two eventual roles of the protein are determined by DNA sequences that are independent and that, at least in principle, can evolve separately, both in prokaryotic [25] and eukaryotic cells [26].

2.2. Links reliabilities

Each of the SBM is fully defined by determining the two partitions above and the probability matrix, hence $m = (P_s, P_r, \mathbf{Q})$. Additionally, we define the reliability of a certain link $i \to j$ as the probability:

$$R_{i\to j} = P(A_{i,j} = 1|A^o). \tag{3}$$

On the other hand, let us consider a couple of nodes (i, j) so that $i \in \sigma_i$ in the senders partition P_s and $j \in \rho_j$ in the receivers partition P_r. The probability of observing a link from node i to node j in a network generated by our model is:

$$P(A_{i,j} = 1|P_s, P_r, \mathbf{Q}) = Q_{\sigma_i,\rho_j}. \tag{4}$$

Consequently, the probability of observing the graph A^o as a realization of the same model is given by the binomial product:

$$P(A^O|P_s, P_r, \mathbf{Q}) = \prod_{\sigma\in P_s,\rho\in P_r} Q_{\sigma\rho}^{l^O_{\sigma\rho}}(1 - Q_{\sigma\rho})^{r_{\sigma\rho}-l^O_{\sigma\rho}}, \tag{5}$$

where $l^O_{\sigma,\rho}$ is the number of links observed between nodes placed in σ in P_s, and nodes placed in ρ in P_r. Regarding $r_{\sigma,\rho}$ it is the maximum possible value for $l^O_{\sigma,\rho}$, that is, the product of the sizes of blocks $\sigma \in P_s$ and $\rho \in P_r$. Substituting the three last expressions into Eq. 2, we get, after integration over all possible probability matrices for each case, that the reliabilities of links are:

$$R_{i\to j} = \frac{1}{Z} \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r) \frac{l^O_{\sigma_i,\rho_j} + 1}{r_{\sigma_i,\rho_j} + 2} e^{-H(P_s,P_r)}, \tag{6}$$

with P_S and P_R standing, respectively, for the spaces of all possible partitions of nodes as link senders (S) and link receivers (R). Node i belongs to block σ_i in P_s; while node

j is located in ρ_j at P_r. Finally, $P(P_s, P_r)$ is here the a priori probability of observing a subset of models defined by P_s and P_r, under the assumption that once partitions are fixed, all possible models that one can get by changing the probability matrices are equally probable. In addition, the partition function Z in the last equation takes the form:

$$Z = \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r) e^{-H(P_s, P_r)} \tag{7}$$

and the hamiltonian function is:

$$H(P_s, P_r) = \sum_{\substack{\sigma \in P_s \\ \rho \in P_r}} \left[ln\left(r_{\sigma\rho} + 1\right) + ln \binom{r_{\sigma\rho}}{l^O_{\sigma\rho}} \right] \tag{8}$$

Up to this point, the scheme of the method is totally analogous to the baseline method for undirected systems presented in [24]. However, the generalization of the method to directed networks requires further refinements. More precisely, as it is detailed in the Appendix, we must adopt here the following hypothesis. Let $\vec{\chi}_{P_x}$ be the vector whose components are the (ordered) number of nodes present in each of the blocks within partition P_x. We have that

$$\begin{aligned} &P(P_s, P_r) = constant \;\; \forall (P_s, P_r) \;\; with \;\; \vec{\chi}_{P_s} = \vec{\chi}_{P_r}, \\ &P(P_s, P_r) = 0 \;\; \forall (P_s, P_r) \;\; with \;\; \vec{\chi}_{P_s} \neq \vec{\chi}_{P_r}. \end{aligned} \tag{9}$$

Then, the a priori probabilities cancel out in Eqs. (6) and (7), and thus, the mathematical forms of these expressions are identical to those given in [24], except for the fact that here, sums and products are taken over the combination of two partition spaces: P_s and P_r, with the additional constraint that the only couple of partitions (P_s, P_r) that computes are those for which $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$ (See Appendix).

Nevertheless, the reliabilities sums have always the form of a canonical ensemble average, which allows us to use again a Metropolis algorithm to sample among all the possible pairs of partitions compatible with the condition $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$, those yielding to smaller hamiltonians and thus contributing the most to the sum (see Appendix). When the sampling finishes, we recover the reliabilities of all possible directed links in the network despite of their directionality –obviously, in general $R_{i,j} \neq R_{j,i}$–. Moreover, by ranking the links one can test which are the more reliable ones, no matter whether a given link was observed in our graph A^o or not. This is what we do in the following sections.

2.3. Method accuracy

In order to check the performance of our approach, we perform a series of tests on top of different networks as in [24]. To this end, we use six well-known directed networks (see Appendix): a social network of radio calls among a closed set of operators [27], a network of hyperlinks in an on-line glossary [28, 29], the trophic webs of Narragansett bay [30]

and the Everglades [31], the cell-fate determination gene network of flower development of *Arabidopsis thaliana* [32] and the regulatory network among transcription and sigma factors of *Mycobacterium tuberculosis* [14].

Assuming that these networks are error-free, we randomly remove a certain proportion of links. Then, we run our algorithm and rank the links reliabilities as coming out of the algorithm. We define the accuracy of the method when it comes to identify missing interactions as the probability that removed links are assigned a higher reliability ranking -i.e., they are false negatives- as compared to those that are true negatives. On the contrary, to test whether the method is able to identify spurious interactions accurately, we randomly add a proportion of links between nodes which are already senders and receivers in the original network. As before, link reliabilities are computed and the ordered ranking is used to check the accuracy of the method, which in this case is given by the (mean) probability that spurious interactions -now they are like false positives- are ranked lower than true links.

In order to evaluate the performance of our method, we compare its accuracy with two of the latest (and to the best of our knowledge the only two dealing with directed networks) alternative approaches to the problem, due to Zhang et al. [23] and Kim and Leskovec [22], respectively.

Results of the accuracy tests are shown in Fig. 2. In the left panels, we have represented the accuracy of the methods regarding the identification of missing links, and in the right panels, we show the accuracies related to spurious links. Black series correspond to the SBM-based algorithm presented here, red data series correspond to the method in [23] and green series to KronEM algorithm developed in [22].

As we can see, in the one hand, the SBM-based method systematically outperforms that of Zhang et al., except for the case of the social network of radio calls, for which the performance of the latter is slightly better, mainly regarding the prediction of spurious links. In fact, a deeper analysis of the two methods, as we discuss next, show that they give highly correlated outcomes. In the other hand, the SBM-based approach outperforms KronEM algorithm in eight panels, and underperforms it in three. At the last case -spurious links in the glossary network- both methods perform very similarly.

2.4. Alternative methods: Zhang's approach

According to the first of these alternative methods, due to Zhang et al. [23], the reliability of a link is thought to be proportional to the number of bi-fans (graph formed by two senders and two receivers each one of which receives a link from each sender [33]) in which each link participates. Similarly, the reliability of a non existing link can be evaluated as the number of bi-fans that would be generated by adding the absent interaction. The so evaluated scores are integers and obviously have no absolute probabilistic interpretation; nevertheless, pairs of nodes can be ranked by their scores and, in this sense, it is a useful tool for missing and spurious links identification also.

In order to understand in depth the relationship between SBM and Zhang's method,

let us recall some technical details of our approach. As it is thoroughly explained in the Appendix, in the Metropolis algorithm used by the SBM-based method, the partitions that give lower hamiltonians and thus mostly contribute to the reliability sums are those that find a better compromise between two conflicting constraints. The first of these constraints is that blocks in the partitions have to be as large as possible. The second constraint forces the amounts of links $l^o_{\sigma,\rho}$ existing between any pair of blocks $\sigma \in P_s$ and $\rho \in P_r$ to be either close to the maximum (i.e. equal to $r_{\sigma\rho}$, the maximum possible value given the size of the blocks) or to the minimum. Once said that, given a certain partition, the bigger the quotient $r = (l^o_{\sigma_i,\rho_j} + 1)/(r_{\sigma_i\rho_j} + 2)$, the bigger the reliability of the link $i \to j$ will be. In the partition depicted in figure 1, for example, for the absent link $3 \to 1$, we have $r = (8 + 1)/(9 + 2) = 9/11$, while, for the link $3 \to 6$, $r = (1 + 1)/(12 + 2) = 2/14$. The example is relevant because it evidences that a pair of nodes with a high link-reliability between them also tends to form a high number of bi-fans, as it is the case of pairs of nodes between blocks $1 \in P_s$ and $1 \in P_r$. The reason is that, to have a high reliability, the number of links between the blocks involved have to be saturated, or nearly saturated, (i.e. $l^o_{\sigma_i,\rho_j}$ near to $r_{\sigma_i\rho_j}$) and that, additionally, as it has been said before, blocks tend to be as large as possible.

To show that this relationship between both methods exist, we have calculated the correlation coefficients between Zhang scores and SBM-based reliabilities. The results, given in table 1 for the six networks analyzed show a high correlation between the outcome of both methods: links with high Zhang-scores tend to have high SBM-based reliabilities and vice-versa. In order to perform an additional test, we can calculate the probability of any pair of nodes (i, j) to co-occur in a common block either at the senders partition:

$$P_{\sigma_i=\sigma_j} = \frac{1}{Z} \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r)\delta(\sigma_i - \sigma_j)e^{-H(P_s,P_r)} \tag{10}$$

or at the receivers partition:

$$P_{\rho_i=\rho_j} = \frac{1}{Z} \sum_{\substack{P_s \in P_S \\ P_r \in P_R}} P(P_s, P_r)\delta(\rho_i - \rho_j)e^{-H(P_s,P_r)} \tag{11}$$

where δ stands for the Kronecker delta function. For the same pair of nodes (i, j) we calculate the number of bi-fans that are generated with nodes i and j playing the receivers roles $N^{rec}_{bf}(i, j)$, and we can compare it with the expected number of bi-fans that they would generate at random in a network of N nodes given their in-degrees $k_{in}(i)$ and $k_{in}(j)$. This expected value is $N^{rec}_{bf}|_{exp}(i, j) = k_{in}(i)k_{in}(j)/N$, and the deviation of the observed number of bi-fans coming from nodes (i, j) and the expected value is $\Delta^{rec}_{bf}(i, j) = N^{rec}_{bf}(i, j) - N^{rec}_{bf}|_{exp}(i, j)$. To test whether bi-fans tend to be formed by couples of receivers that share a common block in the senders partition, we calculate the average of $\Delta^{rec}_{bf}(i, j)$ for those receiver couples having a co-occurence probability

Network	Pearson coefficient
Radio calls	0.912
Glossary	0.917
Narragansett bay	0.904
Everglades	0.850
A.thaliana	0.920
M.tuberculosis	0.943

Table 1. Correlations between SBM-based reliabilities and Zhang scores. To obtain these correlation coefficients we calculate the average SBM-based reliabilities of all pairs of nodes sharing a common Zhang score. The Pearson coefficients represented show the existence of correlations between the scores and the averaged reliabilities. For the calculations, we have considered the generalized Zhang scores, which also take into account the degenerated bi-fans in figure 3.

Network	$\langle\Delta_{bf}^{rec}(i,j)\rangle$ for $P_{\rho_i=\rho_j} \leq 0.5$	$\langle\Delta_{bf}^{rec}(i,j)\rangle$ for $P_{\rho_i=\rho_j} > 0.5$	p-value
Radio calls	5.42 ± 0.37	10.30 ± 2.37	$1.20 \cdot 10^{-4}$
Glossary	$(1.80 \pm 0.85) \cdot 10^{-1}$	$(3.12 \pm 1.61) \cdot 10^{-1}$	$2.25 \cdot 10^{-1}$
Narragansett bay	$(4.47 \pm 1.10) \cdot 10^{-1}$	1.63 ± 0.26	$< \cdot 10^{-5}$
Everglades	6.85 ± 0.58	31.09 ± 2.94	$< \cdot 10^{-5}$
A.thaliana	$(-5.11 \pm 3.75) \cdot 10^{-1}$	2.28 ± 1.60	$8.48 \cdot 10^{-2}$
M.tuberculosis	$(3.31 \pm 1.1) \cdot 10^{-2}$	$(1.40 \pm 0.64) \cdot 10^{-1}$	$7.63 \cdot 10^{-3}$

Table 2. Bi-fans tend to aggregate around pairs of receivers sharing common blocks. $\Delta_{bf}^{rec}(i,j)$ is the difference between the number of bi-fans in which nodes (i,j) participate as receivers $N_{bf}^{rec}(i,j)$ and the expected value of the same quantity $N_{bf}^{rec}|_{exp}(i,j) = k_{in}(i)k_{in}(j)/N$ in the null case. Positive values of $\Delta_{bf}^{rec}(i,j)$ means that nodes (i,j) have more common in-neighbors than expected at random. In this table we show that receiver pairs with greater co-existence probabilities $P_{\rho_i=\rho_j}$ have greater $\Delta_{bf}^{rec}(i,j)$ on average. The p-values stand for the probability of the mean value of $\Delta_{bf}^{rec}(i,j)$ for the first population (pairs (i,j) with $P_{\rho_i=\rho_j} \leq$) being equal or greater than the mean of the second population. We can repeat the exercise to test the correlation between $P_{\sigma_i=\sigma_j}$ and the deviation of the number of bi-fans generated by receivers couples $\Delta_{bf}^{send}(i,j) = N_{bf}^{send}(i,j) - N_{bf}^{send}|_{exp}(i,j)$; the results are very similar; with all p-values under 20% and 4 out of 6 under 5%. Degenerated bi-fans have also been taken into account.

$P_{\rho_i=\rho_j} \leq 0.5$, and we compare it to the same quantity averaged on couples with $P_{\rho_i=\rho_j} > 0.5$. The results of this test, given in table 2, show that in most networks a larger probability of co-existence at receivers' partitions comes together with a higher number of bi-fans formed by the couple of receivers.

After these observations, the reasons behind Zhang's method performance could be reinterpreted as a simple consequence of the existence of an underlying block structure. Under this assumption, the mapping between both methods allows us to overcome one of most clear limitations in [23], i.e. its inability to evaluate self-loops reliabilities (a self-loop never joins a bi-fan). Once we have seen that the appearance of bi-fans around

highly reliable links can be rooted in the underlying block structure, we notice that the structures sketched in figure 3, are, from the perspective of stochastic block models, absolutely identical. Recalling the example partition in figure 1, the "pure" bi-fan formed by nodes $(1, 2, 4, 5)$, appears as a consequence of links saturation between blocks $1 \in P_s$ and $1 \in P_r$, just in the same way that the degenerated structure formed by nodes $(1, 2, 4)$ does. Thus, what we propose here is a variation of Zhang's method in which degenerated bi-fans are treated in the same way that "pure" are, and therefore counted when it comes to evaluate the scores, even when they violate one of the main requirements of Zhang et al.'s original approach [23]. The last four rows of figure 2 show results obtained for networks that contain self-loops. As it can be seen, the generalization of the method has little impact on the food webs because of their low number of self-loops, while in the regulatory networks analyzed, in which self-loops are more frequent, the generalized method noticeably outperforms the original Zhang et al.'s proposal, thus supporting our hypothesis.

2.5. Alternative methods: KronEM algorithm

Instead Zhang's approach, –essentially based on local topological information–, the so-called Kronecker expectation-maximization (KronEM) algorithm [22], is based on a family of stochastic network models. These models have two main ingredients: a Kronecker matrix built after the expansion of a low-dimensional matrix θ verifying $\theta_{i,j} \in (0, 1) \forall (i, j)$ via Kronecker multiplication by itself, and a bijection of the node set $\Sigma : i \to \sigma(i)$. In order to describe a network of N nodes, if we are working with a matrix θ of dimensions nxn (typically $n = 2$), we will need to iterate the Kronecker product k times, being k the lowest integer higher than $log_n(N)$. Once done so, the matrix elements $\theta^k_{i,j}$ –always verifying $\theta^k_{i,j} \in (0, 1) \forall (i, j)$– can be interpreted as links reliabilities; more precisely, the matrix entry $\theta^k_{\sigma(i),\sigma(j)}$ is the probability assigned to the link $i \to j$.

It is worth noticing that, as exposed before, the so-constructed kronecker probability matrix has, –unless N is an integer power of n–, more rows than nodes exist in the network. This situation is used in [22] to make inferences on what they tell "the hidden part of the graph", i.e. the part of the graph artificially described by the surplus of rows of the matrix θ^k. However, the adaptation of the algorithm to the problem that we face here is trivial, and allows us to focus only in the "real part" of the graph (as it is already done in [22], section 4.3).

The algorithm proposed in [22] to determine link reliabilities is based in finding, among all possible Kronecker models able to describe a certain network, one that maximizes the overall likelyhood:

$$P(G|\Sigma, \theta) = \Pi_{A_{i,j}=1}(\theta^k_{\sigma(i),\sigma(j)})\Pi_{A_{i,j}=0}(1 - \theta^k_{\sigma(i),\sigma(j)}) \tag{12}$$

and, as it can be shown from [22] such a maximization is feasible on computational time of the order of the number of nodes. The computational requirements, as it can be

Network	Nodes	Time (s)
Radio calls	44	155
Glossary	67	113
Narragansett bay	32	50
Everglades	69	218
A.thaliana	15	2
M.tuberculosis	65	302

Table 3. Computational time required for generating one single reliabilities rank using SBM approach. In each case, 1000 partitions are sampled.

seen in table 3, are heavier for the SBM approach, as it was already determined in [22] for the undirected version of the algorithm presented here.

The reason for this situation does not come from the behavior of the elemental step of our Metropolis algorithm itself; which indeed scales with the number of occupied blocks (i.e., at most, with the number of nodes as well).

Most the time requirements of our algorithm come from the amount of iterations needed for uncorrelating two subsequent partitions in the sampling procedure. This decorrelation intervals depend in not trivial ways not only on the number of nodes, but also on the number of links and, in general, on the topology of the graph. This behavior explains the inexact correlations between computational times and system sizes in table 3.

KronEM algorithm does not present, in principle, these problems. The reason is that although both approaches are model-based bayesian methods, while the SBM approach bases its predictions on recovering a whole ensemble of stochastic models, kronEM algorithm aims at simply pick one optimal model to optimize the likelyhood. This situation, yet having the virtue of reducing the computational requirements of a single run, makes the prediction of the algorithm more volatile, and sensitive to initial conditions as it was already admited in [22].

2.6. Guiding experiments

Once we have tested the general performance of the SBM-based method when compared to KronEM, Zhang's, and generalized Zhang's approaches, we discuss their application in an important and specific domain, that of transcriptional regulatory networks. In this field of research, computational data reliability tools could help mitigate either the relatively poor quality and reduced size of some networks available [34,35] or to integrate vast amounts of information coming from high-throughput experimental techniques.

On the other hand, there are several organisms, –even relevant pathogens– for which the whole transcriptional map is not at hand, despite the fact that having the network would help in the search of new drug targets or vaccines. This is the case of the transcriptional regulatory network of *Mycobacterium tuberculosis*. The bacillus of tuberculosis, responsible of one of the most threatening diseases worldwide,

is probably one of the bacteria whose transcriptome has been best studied during the last years [14, 36, 37]. In 2008 the transcriptional regulatory network of the pathogen consisted of 782 genes and 937 interactions [36], but the last updated version, published in 2011, contains as many as 1624 genes and 3212 interactions [14]. Moreover, the updated version, also added 357 new links between some of the 782 genes that were reported in 2008.

All the aforementioned facts, together with the running costs of experiments are calling for methods that could optimize the search of new interactions. To test whether and to what extent our algorithm could contribute to cure new datasets and guide the experimental search of new transcriptional relations and regulators, we perform a simple exercise with the *M. Tb* datasets of 2008 and 2011. Specifically, we check whether the appearance of the 357 links in the 2011 compilation that connects pairs of genes already integrated in the 2008 network could have been inferred from the analysis of the 2008 network itself.

To simulate the way in which our method could help to identify these new interactions, let us suppose that we are interested on a certain gene of the 2008 network and we look for undiscovered regulations it might receive from any of the regulators already present in the network in 2008 –obviously excluding those that had been already found to regulate its activity at the moment–. If no biological clue is available about what regulators are the more likely candidates to act on our gene, we are forced to experimentally try, one after another, all the possibilities. If the result of some of these experiments is positive, and so the interactions exist, we will identify them at a linear rate, as it is represented in grey in figure 4, panel a. In the same figure, the black curve represents the rate at which all these novel interactions are detected when the possible targets are checked according to their reliabilities calculated using the SBM approach. As it can be seen, the SBM-based method greatly enhances the rate at which new links are discovered, with respect to the random case but also, to a lower extent, with respect to KronEM and Zhang's algorithms .

If the situation is the opposite, and we are interested in unveiling new links coming from any of the regulators of the network in 2008, the rate at which targets of the new links could be experimentally found is represented in figure 4, panel b, when choosing the candidates according to their SBM-based reliabilities, according to the alternative methods and when the order is random. In this case, the performance of SBM and Zhang's methods slightly outperforms KronEm, which is not better than the random procedure. These differences, though moderate, in practice could represent saving time and resources. In fact, starting from the regulators (Fig. 4, panel b) and aiming at finding 50% of the new targets, one has to seek the 39% of the targets with the highest SBM-based reliabilities in each case. This implies that the SBM-based method uses 78% of the time and resources needed if the identification is made randomly. Going back to the results shown in Fig. 4, panel a, that case produces even better results: to find the 50% of the regulations received by a target gene, one must only seek a 20% of the total of regulators. Therefore, the SBM-based method remarkably outperforms the random

search by using as less as 40% of the resources spent in the null case, but also the search orderings based on the alternative methods.

3. Conclusions

We have proposed an extension of the method in [24] to determine link reliabilities in directed networks. This opens the path to the potential application of our technique for the detection of missing and spurious interactions in systems as important as food-webs, transcriptional regulatory networks or certain social networks, all of which are directed networks. A related and interesting problem that however remains to be explored is whether reliability rankings are correlated with significance measures [38–40] of the links identified. For instance, a genuine question is whether finding a highly ranked but lowly significant link is worth the computational cost involved in the calculation. We let this kind of questions for future works.

The accuracy and robustness of the method has been tested exhaustively on networks of different sizes and topological properties. Results of intensive numerical simulations have shown that missing and spurious interactions can be detected successfully, with higher precision than previous approaches in most cases. Additionally, we have numerically shown that the method can be used to guide the experimental search for missing links, as the reliability ranking resulting from the application of the algorithm to an incomplete network provides a very good guideline for experimental tests that eventually lead to the discovery of new interactions in a highly efficient way. This potentiality has important implications for our current efforts to map out transcriptional regulations, particularly, in cases such as that of *Mycobacterium tuberculosis*, where experimental lab protocols are very slow and expensive. At a conceptual level, this exercise makes explicit the ability of our method to predict real, arbitrarily correlated errors in complex directed networks, rather than randomly generated missing or spurious interactions.

In the other hand, after an exhaustive comparison of our method with the method proposed in [23], we have been able to provide a rationale for the latter approach: when the topology of a real system can be successfully described by a block model, bi-fans systematically appear around highly reliable links. Additionally, the mapping between both methods makes it immediate the generalization of Zhang's approach to deal with self-loops. This generalization enhances the performance of the original algorithm when self-loops are statistically relevant, as it happens in some gene regulatory networks (see figure 2 lower row).

Our SBMs-based model has however an important limitation. It is prohibitively costly in terms of computational time for large systems, and for sure much more expensive than Zhang's approach, or even than single runs of kronEM algorithm. Therefore, the method proposed here is mainly aimed at relatively small systems. For larger networks, Zhang's generalized method, whose outcomes are highly correlated to our approach in small systems, can be used as a low-cost resource. KronEM algorithm,

in turn, represents an intermediate solution, both in terms of computational expenses and method accuracy and consistence. This kind of situation in which the prize to pay for reaching more accurate tools also appears when facing other problems such as community detection in networks [41].

Alternatively, if more accuracy is required, we believe that the SBM-based method presented in this paper could also be applied to subgraphs, overcoming in this way the size limitations. For instance, one can try to partition the whole system first by using one of the many algorithms available for community detection and then apply the reliability technique only to the detected communities. This kind of solutions will be explored in future work.

Appendix. Some relevant aspects regarding the methodology.

3.1. Phase space

In [24], the mathematical form of the hamiltonian, in the undirected model is, as said before, equivalent to 8, except for the fact that there is only a partition family to sum over. Let us write it as:

$$H_u(P) = \sum_{\alpha<\beta} \left[ln\left(r_{\alpha\beta}+1\right) + ln\binom{r_{\alpha\beta}}{l^O_{\alpha\beta}} \right] \tag{13}$$

The restriction $\alpha < \beta$ (both blocks belonging to the partition P) appears only in order not to sum each term of the sum twice. Let's inspect the two different terms:

$$H_{1u}(P) = \sum_{\alpha<\beta} ln\left(r_{\alpha\beta}+1\right) \tag{14}$$

$$H_{2u}(P) = \sum_{\alpha<\beta} ln\binom{r_{\alpha\beta}}{l^O_{\alpha\beta}} \tag{15}$$

The first term depends, essentially, on how "concentrated" the partition is. Briefly, it is minimal when the nodes tend to concentrate in a few number of blocks. In the case of having all the nodes on the same block, Eq. 14 gives $ln(1+N(N-1)/2)$, where N is the number of nodes, which is approximately equal to $2ln(N)$ when N is large enough. Instead, if we have the opposite situation in which each node is assigned to a different block, then $H_1 = N(N-1)ln(2) >> 2ln(N)$. So, the term H_{1u} minimizes when the partitions are compact, and maximizes in the opposite case. As for the second term, the picture is the opposite. The presence of the combinatory number implies that, to minimize H_{2u}, the partitions of nodes should be a kind of "straight fit" for the links connecting blocks: given any two random blocks α and β, there should be a number of links between the blocks near to the maximum -the product of the block sizes, i.e. $r_{\alpha,\beta}$- or to the minimum (i.e. no link between the blocks). So, if we aim at getting the minimum of this term alone, one must go to the segregated partition in which each node belongs to a different block, for which the term directly vanishes.

Therefore, minimizing the hamiltonian implies finding a compromise between aggregation and segregation of nodes into blocks, as the two terms have clear opposite effects, and no one of the extreme situations are globally convenient. How this picture change when we move to the bipartite scheme? The addition of new degrees of freedom to the system generates an undesirable situation in which, if we perform a Metropolis algorithm letting freely evolve the two partitions, we will reach a situation in which in the P_s space, all nodes gather together into a single block, while in the P_r space we will get an split into as many blocks as nodes are. The reason is that, for the system, such configuration is globally stable, because the two hamiltonian terms, under this configuration, reach values that are far away of the possible maximum. However, in this case, the final configuration is absolutely uninformative.

The above problem comes from the fact that the system is not constrained enough and it is allowed to adopt partitions in each one of the subspaces with very different degrees of aggregation. So, we should impose a further constraint so that the system can only adopt couples of partitions with the same aggregation state (i.e. $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$), the stable. This will allow to get partitions that give rise to minimum hamiltonians being at the same time fully informative and having a compromise at intermediate levels of aggregation between links assignments and block sizes. In this case, the algorithm will be qualitatively analogous to that of the undirected case.

3.2. Metropolis algorithm

In order to perform our Metropolis algorithm, we start by assigning, at random, each node to one block, for example in the space P_s. Then we copy the partition generated to P_r. To ensure independence between the partitions but always verifying the constraint $\vec{\chi}_{P_s} = \vec{\chi}_{P_r}$, we proceed to randomize the partition P_r by iteratively changing the block of couples of nodes (also chosen randomly) a high enough number of times. In this way, the blocks numbered equally in both partitions contains the same number of nodes. Thus, at each Metropolis step, we choose a couple of nodes belonging to the same block in both the partitions P_s and P_r and we try to change both at the same time to the same destination block (each one on its own partition). To ensure that any couples of nodes has the same probabilities of being chosen, we proceed as follows: we start by choosing randomly one node n_1 in one partition. Then we move it to the twin block containing the very same node n_1 in the complementary partition. Inside this twin block, we randomly choose the second node to move, n_2. After the nodes n_1, n_2 are selected and tentatively moved, we recalculate H and accept the move if $H(t+1) < H(t)$. As usual, if the hamiltonian raises up, we accept the move with probability $P = e^{(H(t)-H(t+1))}$ in the standard case. Such an algorithmic scheme guarantees an ergodic exploration of the phase space, and ensures without problems detailed balance. In this way, after a certain transient, the hamiltonian reaches its equilibrium value and at that point, we start the sampling procedure, taking care that two consecutive samples are uncorrelated enough.

3.3. Technical aspects

While the method does not raise any problem when analyzing systems of small size (let us say $N < 200$ nodes and $E < 1000$ links approx.), as those studied in section 2, for larger networks, there sometimes appear some conceptual problems that can make the sampling procedure more difficult. First, it has been observed that the amplitude of oscillations of stationary hamiltonians, in general, increases with the size of the network analyzed. This range can be near 1000 hamiltonian units for systems of less than 2000 nodes, such as those of E.coli [15] or M.tuberculosis [14] transcriptional regulatory networks. Since the distribution of the hamiltonians is qualitatively normal around the average value (results not shown), the higher the amplitude of the oscillation is, the lower the proportion of samples that will contribute significantly to the sum is (let us say, those with H, at most, 10 units greater than the minimum). This problem, when it comes to analyze big networks, will force us to get a too high number of samples to get a minimum amount of relevant ones. The latter can be prohibitive in terms of computational time (recall, in addition, that the computational time of a single Metropolis step also increases with the size of the system).

Here we propose an alternative procedure that can be implemented when the networks under study are too large and computational resources do not allow a full exploration of the phase space. The alternative is as simple as discarding all the samples with $H < \langle H_{stat} \rangle - \gamma \cdot \sigma_{H_{stat}}$, where γ is a coefficient that can be chosen depending on the computational time we require and the number of samples we are looking for. This resource, although in principle could limit the performance of the method, does not affect it significantly, as showed in Fig. 5, panel a.

The black bars in figure 5, panel a, show the consistency of the standard method of sampling without any threshold. We define this consistency as the proportion of reliabilities pairs $R_{i,j}$ $R_{k,l}$ whose relative ordering is preserved in successive reliability ranks obtained with the same method. Moreover, in red bars, the comparison is made between a rank obtained with the standard procedure and another rank for which only the samples that lie over a threshold $\langle H_{stat} \rangle - \gamma \sigma_{H_{stat}}$ have been preserved and considered (here, $\gamma = 2$). Finally, the bars in blue show the internal consistency of the threshold method, that is, the mean proportion of reliability pairs whose order is conserved when we compare pairs resulting from two independent rankings generated using the threshold criterium. As it can be seen, the three measures, for the six systems shown, are consistently high and quantitatively similar between them, thus providing evidence that the threshold method could help in situations where the required computational time is prohibitively large if we aim at getting enough samplings. This kind of procedure has been used for the analysis of the transcriptional regulatory network update represented in figure 4, considering 10.000 partitions with hamiltonian over $\langle H_{stat} \rangle - \gamma \sigma_{H_{stat}}$ with $\gamma = 2$.

There is an additional problem that generally appears when the networks have high mean connectivities, or, strictly speaking, when the mean connectivity is of the

order of half the number of total possible links in the network, that is, in a directed network, $N^2/2$. In these cases, the information stored in the adjacency matrix is high, and so, being high the number of constraints, the dependency of the hamiltonian on the partitions defines a rough energy landscape that sometimes can become difficult to deal with. This situation can lead the system to fall into a local minimum after the thermalization process, and get trapped there. So, once arrived to the stationary state, if the basin of that local minimum is small, we will observe that the system is not able to uncorrelate sufficiently, and thus, even if its energy is small enough to consider it an acceptable minimum, our sampling will be very poor. One solution to this issue would be that of parallelizing the algorithm starting each parallel process from a random initial configuration. In this way, the process will ideally reach independent minima and thus the sampling would be N times reacher, being N the number of parallel processes.

In the above solution is is not possible, the strategy would be to introduce a pseudo temperature $T > 1$ in the Metropolis algorithm, just to ensure the system is able to abandon local minima and explore the whole configurational space looking for other ones. The adoption of this strategy has the problem that, the higher is the temperature, the higher is also the oscillation of amplitudes of the stationary hamiltonian, and therefore the application of a threshold might also be needed.

In Fig. 5, panel b, we show the consistency of our method when the above strategy is implemented (using $T = 2$) in combination with a threshold criterium to select the samples, accepting only those with $H > \langle H_{stat} \rangle - 2\sigma_{H_{stat}}$. Though these operations, again, could compromise the quality of our sampling, we found that the consistency of the ranks generated with the method (Fig. 5, panel b, red bars) compared to those generated by the standard procedure is higher than 85%. On its turn, when we check the internal consistency of the ranks generated with the method is even better and could be greater than that reached with a standard sampling.

3.4. Network models

- **Killworth-Bernard radio calls network**. In their work [27], the authors asked to 44 radio operators (nodes) to rank from 0 to 9 the frequency they had used to call the rest of operators during last month. We have reconstructed our network by assigning a link when the rank associated to it was greater than 1, which produces 400 connections. Dataset available at [29].
- **Graph theory glossary network**. This network is constructed based upon an on-line glossary of definitions of technical terms about graph theory [28]. In the network, each node represents one concept; and a link points from one concept to another if the latter is linked in the definition of the former. The network has 67 nodes (5 of the 72 terms in the glossary are not connected to any other) and 122 links (114 unidirectional links and 4 reciprocal interactions). No self-link is allowed since a defined concept cannot be used in its own definition. Dataset available in [29].

- **Narragansett bay trophic web**. The system [30] originally contained 220 interactions between 35 nodes. We have removed the links involving the nodes associated to input, output and respiration fluxes, in order to take into consideration only the trophic relationships between organisms. The effective size of our system is, thus, 32 nodes and 158 links. Dataset available in [29].
- **Everglades trophic web**. The network described in [31] contains 69 nodes and 916 interactions. It describes the trophic interactions of the Everglades ecosystem in the wet season. Dataset available in [29].
- ***Arabidopsis thaliana* flower development cell fate determination network**. The network contains 15 networks and 37 interaction among the genes that control cell-fate during the process of flower development of the model plant *Arabidopsis thaliana* [32].
- ***Mycobacterium tuberculosis* transcriptional regulatory backbone**. From the whole genome wide network compiled in [14], we have extracted the subnetwork that connects the transcription and sigma factors. The dataset analyzed is the giant component of that network, and contains 65 genes and 130 interactions.

Acknowledgments

J.S. was supported by MINECO through an FPU fellowship. E.C. is funded by the FPI program of the Government of Aragón, Spain. This work has been partially supported by MINECO through Grant FIS2011-25167, Comunidad de Aragón (Spain) through a grant to the group FENOL and by the EC FET-Proactive Project MULTIPLEX (grant 317532).

References

[1] Strogatz S, 2001 *Nature* **410** 268-276.
[2] Barabasi A L and Albert R, 1999 *Science* **286** 509-512
[3] Bolouri H and Davidson E H, 2002 *BioEssays* **24** 12, 1118-1129.
[4] Babu M M, Luscombe N M, Aravind L, Gerstein M. and Teichmann S A, 2004 *Curr. Op. Struct. Biol.* **14** 283-291.
[5] Dunne J A, Williams R J and Martinez N D, 2002 *Proc. Nat. Acad. Sci.* **99** 20, 12917-12922
[6] Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Flammini A and Menczer F, 2011 *Proc. 5th Intl. Conference on Weblogs and Social Media*, **89**
[7] Borge-Holthoefer J et al., 2011 *PLoS One*, **6** 8, e23883
[8] Kossinets G, 2006 *Soc. Networks* **28** 247-268
[9] Schafer J L and Graham J W, 2002 *Psychol. Methods* **7**(2), 147-177
[10] Butts C T, 2003 *Soc. Networks* **25** 103-140
[11] Draghici S, Khatri P, Eklund A C and Szallasi Z, 2006 *Trends in Genet.* **22**(2), 101-109
[12] Brettschneider J, Collin F, Bolstad B M and Speed T P, 2008 *Technometrics* **50**(3), 241-264
[13] Ioannidis J P et al., 2009 *Nat. Genet.* **41**(2), 149-155
[14] Sanz J, Navarro J, Arbués J, Martín C, Marijuán P C and Moreno Y, 2011 *PLoS One* **6**(7), e22178
[15] Gama-Castro S et al., 2011 *Nuc. Acid Res.* **39** Database Issue D98-D105
[16] Sanz J, Cozzo E, Borge-Holthoefer J and Moreno Y, 2012 *BMC Sys. Biol.* **6**:110

[17] Makita Y, Nakao M, Ogasawara N and Nakai K, 2004 *Nuc. Acid Res.* **32** Database Issue D75-D77.
[18] Jansen R et al., 2003 *Science* **302**,449-453.
[19] Clauset A, Moore C and Newman M.E.J., 2008 *Nature* **453** 98-101.
[20] Yan B and Gregory S., 2012 *Phys. Rev. E* **85** 056112.
[21] Lü L and Zhou T, 2010 *Phys. A* **390**1150-1170.
[22] Kim M and Leskovec J., 2011 *SIAM* International conference on data mining.
[23] Zhang QM, Lü L, Wang WQ, Zu YX and Zhou T, 2013 *PLoS One* **8**(2), e55437
[24] Guimerá R and Sales-Pardo M, 2009 *Proc. Nat. Acad. Sci.* **106**(52), 22073-22078
[25] Babu M, Teichmann S and Aravind L, 2006 *J. Mol. Biol.* **358**(2), 614-633
[26] Prud'homme B, Gompel N and Carroll S B, 2007 *Proc. Nat. Acad. Sci.* **104**(1), 8605-8612.
[27] Killworth B and Bernard H, 1976 *Human Org.* **35** 269-286
[28] Bill Cherowitzo's graph theory glossary: http://www-math.ucdenver.edu/ wcherowi/courses/m4408/m4408f.html
[29] pajek datasets web page: http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm
[30] Monaco M E and Ulanowicz R E, 1997 *Mar. Ecol. Prog. Ser.* **161** 239-254
[31] Ulanowicz R E, Heymans J J, and Egnotovich M S, 2000. Network analysis of trophic dynamics in South Florida ecosystems, FY 99: the graminoid ecosystem. Annual Report to the United States Geological Service Biological Resources Division Ref. No.[UMCES] CBL 00-0176.
[32] Espinosa-Soto C, Padilla-Longoria P and Alvarez-Buylla ER, 2004 *The Plant Cell* **16** 2923-2939
[33] Milo R, Shen-Orr S, Itzkovitch S, Kashtan N, Chklovskii D, et al., 2002 *Science* **298**(5594): 824-827.
[34] Stathopoulos A and Levine M, 2005 *Developmental Cell* **9** 449-462
[35] Suel G M, García-Ojalvo J, Liberman L M and Elowitz M B, 2006 *Nat. Lett.* **440** 545-550
[36] Balazsi G, Heath A P, Shi L and Gennaro M L, 2008 *Mol. Sys. Biol.* **4** 225
[37] Jacques P E, Gervais A L, Cantin M, Lucier J F and Dallaire G, 2005 *Bioinformatics* **21**(10), 2563-2565
[38] Tumminello M, Aste T, Di Matteo T, Mantegna R N, 2005 *Proc. Nat. Acad. Sci.* **102**(30) 10421-10426
[39] Serrano M A, Boguñá M and Vespignani A, 2009 *Proc. Nat. Acad. Sci.* **106**(16) 6483-6488
[40] Radicchi F, Ramasco J J, Fortunato S., 2011 *Phys. Rev: E* **83** 046101
[41] Fortunato S., 2010 *Phys. Rep.* **486** 75-174

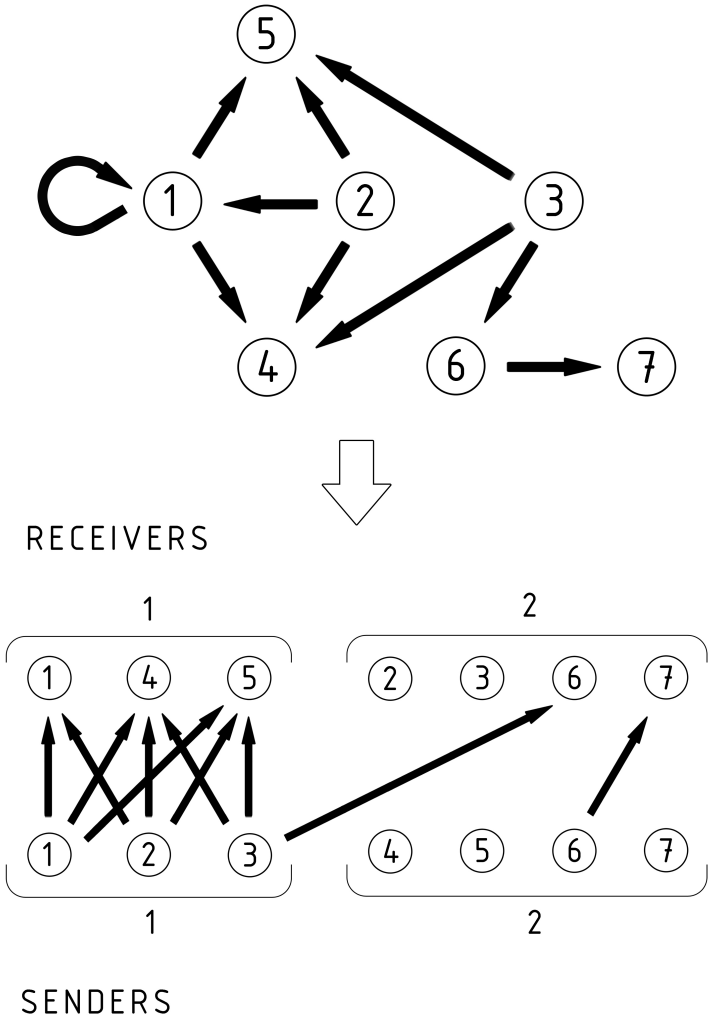

Figure 1. Stochastic block models for directed networks. In this example, partitions count with two blocks of three and four nodes, respectively. Notice that the number of occupied blocks and the number of nodes within them have to be the same both in senders and receivers partitions. The partition represented has a low hamiltonian, because blocks are large enough and the number of links between each pair of blocks $l^0_{\sigma,\rho}$ is either close to 0 or to the maximum possible value in each case $r_{\sigma,\rho}$. We have: $(l^o_{1,1}, r_{1,1}) = (8, 9)$, $(l^o_{1,2}, r_{1,2}) = (1, 12)$, $(l^o_{2,1}, r_{2,1}) = (0, 12)$, $(l^o_{2,2}, r_{2,2}) = (1, 16)$. According to this picture, links from nodes of block $1 \in P_s$ to nodes in block $1 \in P_r$ are much more reliable than any others: thus, the link $3 \to 1$ is a prototypical missing link, while the links $3 \to 6$ and $6 \to 7$ are prototypical false positives.

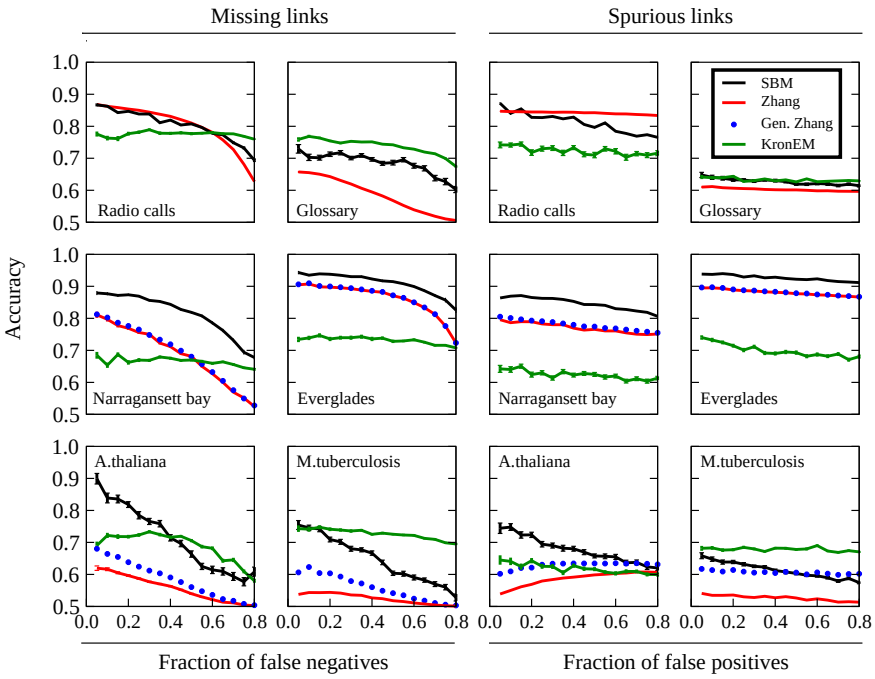

Figure 2. Method accuracy in the detection of missing (left) and spurious interactions (right) in six directed networks, according to the four methods explored in the text. The first two networks (radio calls and glossary) lack self-loops by construction, and hence, it makes no sense to generalize Zhang's approach there. For the same reason, self-links are not allowed as spurious interactions for these two networks. Not shown error bars are smaller than symbol size or line thickness.

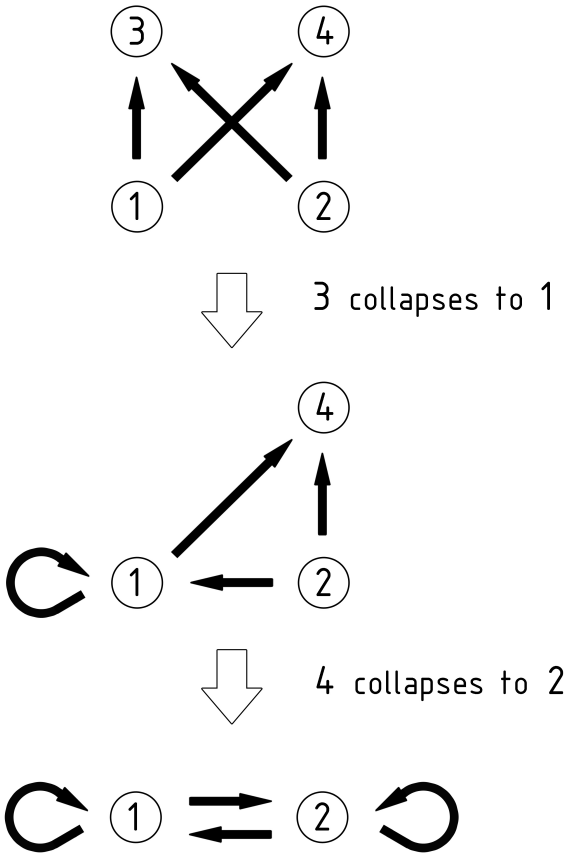

Figure 3. Bi-fan degeneration. From a "pure" bi-fan (upper graph), if the identity of one –or both– of the couples sender-receiver coincides, we obtain these motifs, that we call degenerated bi-fans.

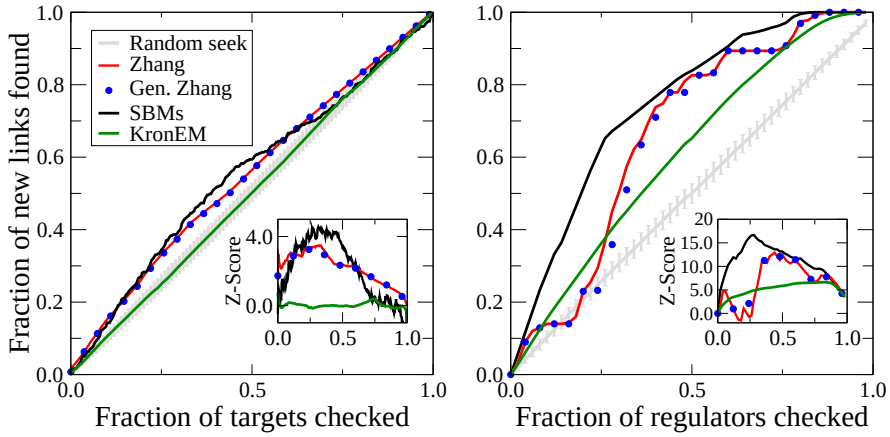

Figure 4. *Mycobacterium tuberculosis* transcriptional regulatory network update analysis. Black: SBM-based seek. Red: Zhang's scores. Blue: generalized Zhang's scores. Green: KronEM algorithm. Grey: random seek. (a): Proportion of regulators checked versus proportion of new links found, when focusing on targets receiving the new links.(b): Regulators based search: Proportion of targets checked versus proportion of new links found, when focusing on regulators sending the new links. In the insets, the Z-Score of the methods' performance is computed, when compared to the random procedures, whose error bars ($\sigma = 1$) are represented in grey. As it can be seen, all three methods outperform the random procedure, mostly at first stages, and more remarkably in the case of target based search (panel a).

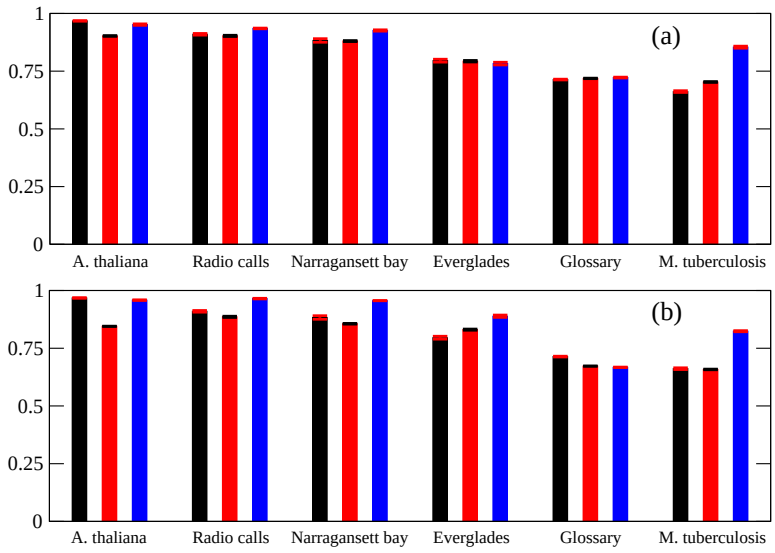

Figure 5. Coherence of ranks defined as the proportion of reliabilities that preserve ordering in successive realizations obtained with diverse sampling strategies: Panel (a): black bars: standard sampling procedure. Blue bars: Threshold sampling. We have set $\gamma = 2$. See the text for further details. Red bars: relative coherence of standard sampling ranks vs. threshold sampling ranks ($\gamma = 2$). Panel (b): black bars: standard sampling procedure. Red bars: relative coherence of standard sampling ranks vs. Hot-threshold sampling ranks ($T = 2$, $\gamma = 2$). Blue bars: Hot-threshold sampling.($T = 2$, $\gamma = 2$). See the text and Appendix for details.