Reconstructing topological properties of complex networks using the fitness model

Giulio Cimini[1], Tiziano Squartini[1], Nicolò Musmeci[2], Michelangelo Puliga[3],
Andrea Gabrielli[1,3], Diego Garlaschelli[4], Stefano Battiston[5], Guido Caldarelli[3]

[1] Institute for Complex Systems (ISC-CNR) UoS "Sapienza" University of Rome (Italy)

[2] King's College, London (United Kingdom)

[3] IMT Institute for Advanced Studies, Lucca (Italy)

[4] Lorentz Institute for Theoretical Physics, University of Leiden (Netherlands)

[5] Department of Banking and Finance, University of Zurich (Switzerland)

Abstract

A major problem in the study of complex socioeconomic systems is represented by privacy issues—that can put severe limitations on the amount of accessible information, forcing to build models on the basis of incomplete knowledge. In this paper we investigate a novel method to reconstruct global topological properties of a complex network starting from limited information. This method uses the knowledge of an intrinsic property of the nodes (indicated as *fitness*), and the number of connections of only a limited subset of nodes, in order to generate an ensemble of *exponential random graphs* that are representative of the real systems and that can be used to estimate its topological properties. Here we focus in particular on reconstructing the most basic properties that are commonly used to describe a network: density of links, assortativity, clustering. We test the method on both benchmark synthetic networks and real economic and financial systems, finding a remarkable robustness with respect to the number of nodes used for calibration. The method thus represents a valuable tool for gaining insights on privacy-protected systems.

1 Introduction

The reconstruction of the statistical properties of a network when only limited information is available represents one of the outstanding and unsolved problems in the field of complex networks [1, 2]. A first example is the case of financial networks, for which systemic risk estimation is based on the inter-dependencies among institutions [3, 4]—yet, due to confidentiality issues, the information that regulators are able to collect on mutual exposures is very limited [5]. Other examples include social networks, for which information may be unavailable because of privacy problems or simply for the impossibility to sample the whole system.

Network reconstruction has been typically pursued through Maximum Entropy (ME) algorithms [6, 7, 8], which obtain link weights via a maximum homogeneity principle; however, the strong limitation of these algorithms resides in the assumption that the network is fully connected (for this reason they are known as "dense reconstruction methods"), while real networks show a largely heterogeneous connectivity distribution.

More refined methods like "sparse reconstruction" algorithms [2] allow to obtain a network with arbitrary heterogeneity, but still cannot set an appropriate value for such heterogeneity. Recently, a novel *bootstrapping* (BS) method [9, 10] has been proposed in order to overcome these problems. The BS method uses the limited information on the system to generate an ensemble of networks according to the *exponential random graph* (ERG) model [11]—where, however, the Lagrange multipliers that define it are replaced by *fitnesses*, *i.e.*, known intrinsic node-specific properties related to the network topology [12]. The estimation of the network topological properties is then carried out within the ERG-induced ensemble. The method builds on previous results [13] which showed that, in the particular case of the World Trade Web (see below), the knowledge of a non-topological property (the Gross Domestic Product), if coupled to that of the total number of links, allows to infer the topological properties of the network with great accuracy. This procedure can be restated within a maximum-likelihood framework [14]. The BS method uses these preliminary observations to provide a reconstruction procedure valid in the case when the the degree sequence of the network (*i.e.*, the number of connections for each node) is known only partially.

While past works [2, 6, 7, 8, 9] mainly dealt with using the limited information available on the network to estimate specific high-order properties such as systemic risk, in the present paper we employ the BS method to reconstruct the fundamental properties that are commonly used to describe a network: density of links, assortativity, clustering (see sec. 3). By focusing on these previously untested properties, we are able to enlarge the basket of quantities that are properly estimated by the BS approach. To validate our method we study how its accuracy depends upon the size of the subset of nodes for which the information is available; our case-study includes synthetic networks generated through a *fitness* model [12] as well as real instances of networked systems: 1) the World Trade Web (WTW), [15], *i.e.*, the network whose nodes are the countries and links represent trade volumes among them, and 2) the network of interbank loans of the e-mid (E-mid) interbank money market [16].

2 Method

We start by briefly describing the ERG model and the fitness model, on which the BS method builds.

The ERG model is one of the most common network generation framework [11, 17, 18], which consists in defining an ensemble Ω of networks which is maximally random, except for the ensemble average of a set of network properties $\{\langle C_a \rangle_\Omega\}$—constrained to some specific values $\{C_a^*\}$. The probability distribution over Ω can then be defined via a set of control parameters $\{\theta_a\}$, namely the set of Lagrange multipliers associated with the constraints $\{C_a^*\}$. A particular yet widely used case of the ERG model is known as the Configuration Model (CM) [11], which is obtained by specifying the mean degree sequence $\{k_i^*\}_{i=1}^N$ of the network. In this case, each node i is identified by the Lagrange multiplier θ_i associated to its degree k_i. By defining $x_i \equiv e^{-\theta_i} \ \forall i$, the ensemble probability that any two nodes i and j are connected reads [11]:

$$p_{ij} = \frac{x_i x_j}{1 + x_i x_j} \tag{1}$$

so that x_i quantifies the ability of node i to create links with other nodes (induced by its degree k_i) [19].

On the other hand, the *fitness* model [12] assumes the network topology to be determined by an intrinsic non-topological property (known as *fitness*) associated with each node of the network, and has successfully been used in the past to model several empirical economical networks [20, 16, 13].

The BS method [9] combines these two network generation models, working as follows. We start from incomplete information about the topology of a given network G_0 (consisting of N nodes): we assume to

know the degree sequence $\{k_i^*\}_{i \in I}$ of only a subset I of the nodes (with $|I| = n < N$) and an intrinsic, non-topological property $\{y_i\}_{i \in V}$ for all the nodes—that will be our fitness (see below). Using this information, we want to find the most probable estimate of the value $X(G_0)$ of a topological property X computed on the network G_0, compatible with the aforementioned constraints. The method builds on two important assumptions:

1. The network G_0 is interpreted as drawn from an ERG-induced ensemble Ω. We then expect the quantity $X(G_0)$ to mostly vary within the range $\langle X \rangle_\Omega \pm \sigma_X^\Omega$, where $\langle X \rangle_\Omega$ and σ_X^Ω are respectively average and standard deviation of property X estimated over the ensemble Ω.

2. The non-topological fitnesses $\{y_i\}$ are assumed to be proportional the degree-induced exponential Lagrange multipliers $\{x_i\}$ through a universal (unknown) parameter z:[1] $x_i \equiv \sqrt{z}y_i \ \forall i$. Therefore eq. (1) becomes:

$$p_{ij} = \frac{z y_i y_j}{1 + z y_i y_j}. \tag{2}$$

Thanks to these two assumptions we can turn the problem of evaluating $X(G_0)$ into the one of choosing the optimal ERG ensemble Ω compatible with the constraints on G_0, which is the most appropriate to extract the real network G_0 from—given that we know only partial information. Once Ω is determined (by the set $\{x_i\}$ and thus by the set $\{y_i\}$), we can use the average $\langle X \rangle_\Omega$ as a good estimation for $X(G_0)$ and σ_X^Ω as the typical statistical error. Now, since we know the rescaled fitness values $\{y_i\}$, the problem becomes equivalent to that of finding the most likely value of z that defines the ensemble Ω according to eq. (2).

An estimation for the value of z can be found using the incomplete degree sequence through the following relation [9]:

$$\sum_{i \in I} \langle k_i \rangle_\Omega \equiv \sum_{i \in I} \sum_{\substack{j=1 \\ j \neq i}}^{N} p_{ij} = \sum_{i \in I} k_i^* \tag{3}$$

in which the first equality comes from the definition of the ERG model, and the second one is the application of the maximum-likelihood argument [11]—the whole equation being restricted only to the nodes belonging to the subset I. Since $\langle k_i \rangle_\Omega = \sum_{j(\neq i)} p_{ij}$ contains the unknown z through eq. (2), and $\{y_i\}$ and $\sum_{i \in I} k_i^*$ are known, eq. (3) defines an algebraic equation in z, whose solution can be used to estimate the degree-induced Lagrange multipliers $x_i = \sqrt{z}y_i \ \forall i$ of the ERG ensemble $\Omega(z)$, and at the end to obtain an estimation of $X(G_0)$.

Summing up, the BS algorithm consists in the following steps. Given a network G_0, the knowledge of some non-topological property $\{y_i\}$ for all the nodes and the knowledge of the degrees of a subset I of nodes:

- compute the sum of the degrees of the nodes in I ($\sum_{i \in I} k_i^*$) and use it together with $x_i = \sqrt{z}y_i \ \forall i$ to solve eq. (3) and to obtain the corresponding value of z;

- using the estimated z and the knowledge of $\{y_i\}$, generate the ensemble $\Omega(z)$ by placing a link between each pair of nodes i and j with probability p_{ij} given by eq. (2);

- compute the estimate of $X(G_0)$ as $\langle X \rangle_\Omega \pm \sigma_X^\Omega$, either analytically or numerically.

---

[1]Fitnesses are often used within the ERG framework provided an assumed connection between them and the Lagrange multipliers. For instance, countries Gross Domestic Products (GDPs) work well as fitnesses when modeling the WTW, and eq. (1) accurately describes the WTW topology when $-\theta_i \propto \log(\mathrm{GDP}_i)$ [13]. In any case, this second assumption (or any other relation $y_i = f(x_i)$) can be appropriately tested on the subset I of nodes for which the degree is known.

3 Topological Properties

As stated in the introduction, in testing the BS method we will focus on the topological properties (each playing the role of X in the previous discussion) which are commonly regarded as the most significant for describing a network. To define these properties, we use the formalism of the adjacency matrix—where $a_{ij} = 1$ if nodes i and j are connected, and $a_{ij} = 0$ otherwise. We consider:

- link density (or connectance):

$$D := \frac{\sum_{i<j} a_{ij}}{N(N-1)/2} = \frac{2L}{N(N-1)} \tag{4}$$

  which is the ratio between the actual number of links in the network and the maximal one compatible with the number of nodes N;

- mean average nearest-neighbors degree:

$$k_{nn} := \frac{\sum_i k_{nn,\,i}}{N} \quad \text{where} \quad k_{nn,\,i} := \frac{\sum_{j(\neq i)} a_{ij} k_j}{k_i} = \frac{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{jk}}{\sum_{j(\neq i)} a_{ij}} \tag{5}$$

  i.e., the arithmetic mean of the degrees of the neighbors of each node, averaged over all nodes;

- the mean clustering coefficient:

$$c := \frac{\sum_i c_i}{N} \quad \text{where} \quad c_i := \frac{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{ik} a_{jk}}{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{ik}} \tag{6}$$

  i.e., the ratio between the number of observed links in each node neighborhood and the maximum possible number of such links, averaged over all nodes [21];

- average rich-club coefficient:

$$\phi = \sum_k P(k)\,\phi(k) \quad \text{where} \quad \phi(k) = \frac{\psi - D}{1 - D} \quad \text{and} \quad \psi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)} \tag{7}$$

  with $P(k)$ representing the fraction of nodes having degree equal to k and $\phi(k)$ the ratio between the $E_{>k} = \sum_{i:k_i>k} \sum_{j:k_j>k} a_{ij}$ edges actually connecting the $N_{>k} = \sum_i \Theta\left[\sum_j a_{ij} - k\right]$ nodes with degree higher than k and the maximum possible number $N_{>k}(N_{>k}-1)/2$ of such edges (or, in other words, the density of links of the subgraph consisting of only the nodes with degree higher than k) [22].

4 Dataset

In order to validate our BS method we use two instances of real economic systems. The first one is the World Trade Web [15], *i.e.*, the network whose nodes represent the world-countries and links represent trade volumes between them:[2] thus, w_{ij} is the total amount of the export of country i to country j. The second one is the interbank loan network of the so-called E-mid interbank money market [16]. In this case, the nodes represent the banks and a link w_{ij} between banks i and j represents the amount of the loan from bank

---

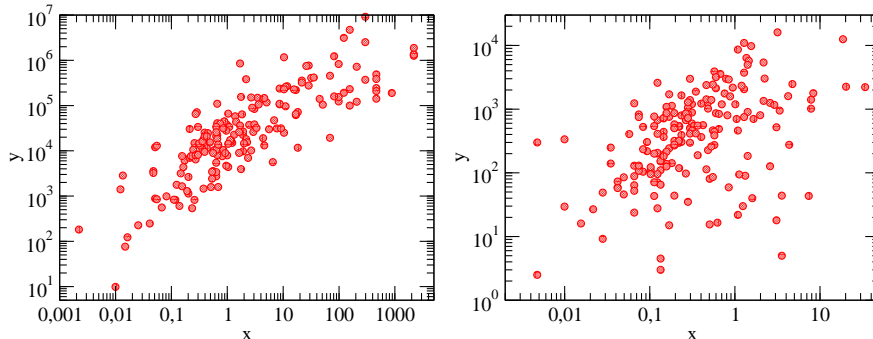[2]For WTW, we use trade volume data for the year 2000.

4

Figure 1: Scatter plot of node fitness $\{y_i\}$ vs exponential Lagrange multiplier $\{x_i\}$ for WTW (left panel) and E-mid (right panel).

i to bank j.[3] These datasets are particularly suited for our study, as the node fitnesses $\{y_i\}$ can be naturally identified with country GDP for the WTW and with the banks total exposures (*i.e.*, the total lending) for E-mid. Thus in both cases each node fitness coincides with its total strength: $y_i \equiv s_i = \sum_j w_{ij}$. The binary undirected version of these networks (that we want to reconstruct) is then built as $a_{ij} = \Theta[w_{ij} + w_{ji}]$.

5 Test of the BS method

Before proceeding to results, we remark that the BS method is subject to two different types of errors. The first one is due to the limited information available for calibrating the ERG model: since we know only the degrees of a subset I of nodes, we can just obtain an estimate of the best z of the ERG ensemble through eq. (3). The second error comes from the assumption that the node fitnesses $\{y_i\}$ are proportional to the degree-induced Lagrange multipliers $\{x_i\}$. Fig. 1 shows the relation between $\{y_i\}$ and $\{x_i\}$ for the two empirical networks. Indeed, there are deviations from linearity—which would correspond to a perfect realization of the fitness model. Note that a better correlation is observed for the WTW: thus, we can expect the BS method to work better in this case.

The quantitative estimation of the BS method effectiveness in reconstructing the topological properties of the two case-studies networks thus proceeds in two steps, as follows.

Test on synthetic networks. To assess the errors which are only due to the limited information available about the degree sequence, we first perform a benchmark test on *synthetic* networks generated through the fitness model. This means that we use the fitnesses $\{y_i\}$ (GDPs for the WTW and total loans for E-mid) to evaluate the "real" z of the ERG ensemble by solving eq. (3) with all nodes included in I (*i.e.*, assuming to know the whole degree sequence), and then draw a network G_0 from $\Omega(z)$ by numerically generating it through eq. (2). G_0 is now the network to reconstruct through the BS method (*i.e.*, with partial information), and

---

[3]For E-mid, we consider snapshots of loans aggregated on a monthly scale (as also done in other works [16]) because of the high volatility of the links at shorter time scales. In the following, we report the results about the snapshot for February 1999. We performed the same analysis also for other monthly snapshots and we found comparable results.

the value of X is computed both on G_0 itself as $X(G_0)$, as well as on the whole ensemble $\Omega(z)$ as $\langle X \rangle_{\Omega(z)}$.[4] Inferring G_0 then consists of the following operative steps:

- choose a value of $n < N$ (the number of nodes for which the degree is known);

- build a set of $M = 100$ subsets $\{I_\alpha\}_{\alpha=1}^M$ of n nodes picked at random;

- for each subset I_α, use the degree sequence in I_α to evaluate z_α from eq. (3), and use it to build the ensemble $\Omega(z_\alpha)$;

- use the linking probabilities from eq. (2) to compute the value of property X over the ensemble $\Omega(z_\alpha)$ as $X_\alpha = \langle X \rangle_{\Omega(z_\alpha)}$;

- compute the relative root mean square error (rRMSE) of property X over the subsets $\{I_\alpha\}$:

$$r_X \equiv \sqrt{\frac{1}{M} \sum_{\alpha=1}^M \left[\frac{X_\alpha}{X_0} - 1 \right]^2} \tag{8}$$

In the rRMSE expression, X_0 denotes a reference value of property X, which can be either $X(G_0)$ (the value of X measured on G_0), as well as $\langle X \rangle_{\Omega(z)}$ (the value of X on the whole ensemble $\Omega(z)$). The consequent two alternative rRMSEs (which we denote as r_X^0 and $r_X^{\Omega_0}$, respectively) provide different estimates of the BS method accuracy. In fact, r_X^0 tests the ability to reproduce a single outcome of the ensemble sampling, whereas, $r_X^{\Omega_0}$ refers to the theoretical values of X expected on the ensemble. Note that since the ensemble $\Omega(z)$ is generated through the fitness model, by construction $r_X^{\Omega_0} \to 0$ for $n \to N$; however, r_X^0 does not necessarily tend to zero, because the generated configuration G_0 is a single realization of $\Omega(z)$ and thus in general $X(G_0) \neq \langle X \rangle_{\Omega(z)}$.

Test on real networks. The testing procedure for real networks is equivalent to the one described above, with the only difference that G_0 and X_0 are now the empirical network and the value of X computed on such network respectively—the rRMSE in this case is denoted as r_X^R. We recall again that since the fitness model is only an approximation for real networks, in this case we expect larger errors from the BS reconstruction than those computed for the synthetic networks.

Results. In order to study the accuracy of the BS reconstruction, we study how the rRMSE for the various topological properties we consider varies as a function of the size n of the subset of nodes used to calibrate the ERG model (*i.e.*, for which information about the degrees is available). Results are shown in Fig.2 for WTW and in Fig. 3 for E-mid. We observe that in all cases there is a rapid decrease of the relative error as the number of nodes n, used to reconstruct the topology, increases. This is an indication of the goodness of the estimation provided by the BS method. As expected, the rRMSE is higher for real networks than for synthetic networks, and the difference between the two respective curves gives a quantitative estimation of the error made in modeling real networks with the fitness model. The fact that such difference is higher for E-mid than for WTW is directly related to the better correlation between node fitnesses and node degrees observed in the latter case. Note also that the errors for E-mid are higher than the ones for WTW also for the corresponding synthetic networks. This feature is easily explained by the higher link density of WTW

---

[4]In the latter case, we use $p_{ij} = \langle a_{ij} \rangle_{\Omega(z)}$ as the expected values of the adjacency matrix elements a_{ij} in the definitions (4,5,6,7).
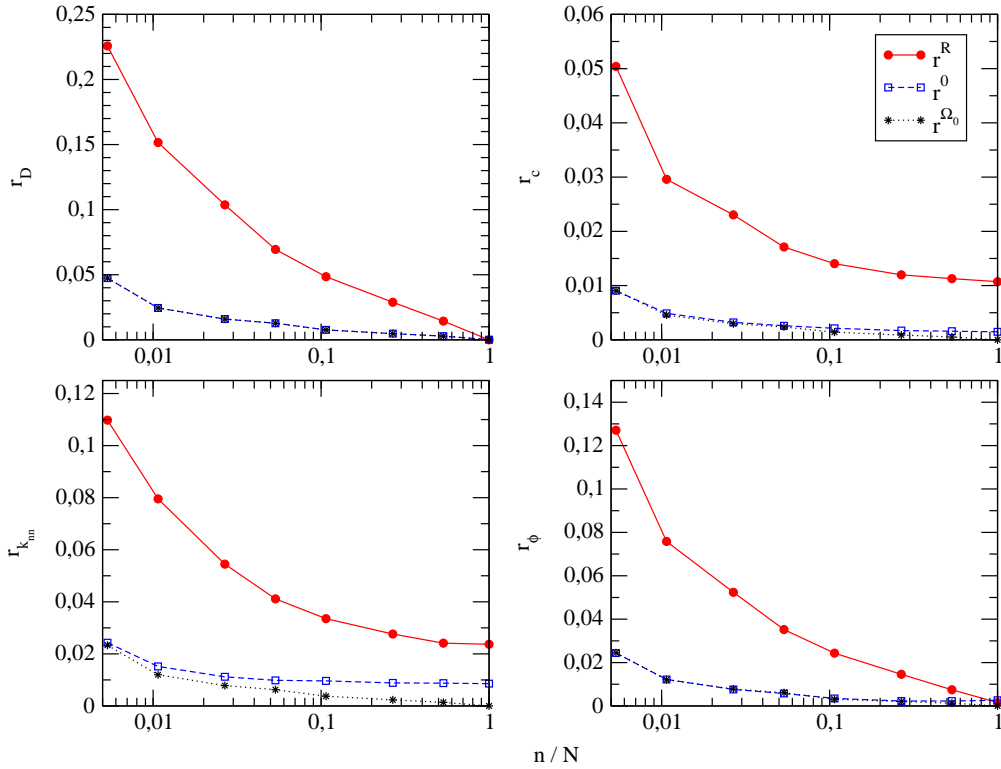
Figure 2: rRMSE of various topological properties for different values of n, obtained by the BS method on the real WTW network and on its synthetic version obtained through the fitness model. Top left: density D; top right: clustering coefficient c; bottom left: average nearest-neighbor degree k_{nn}; bottom right: rich-club coefficient ϕ.

($D \simeq 0.59$) with respect to that of E-mid ($D \simeq 0.20$): denser networks are easier to reconstruct because nodes have more links, and thus carry more information for the BS method to exploit (as fluctuations are tamed out).

6   Conclusions

In this paper we tested a novel network reconstruction (BS) method that allows to estimates the topological properties of a network by using only partial information about its connectivity, as well as a non-topological quantity (interpreted as fitness) associated to each node. This method is particularly useful to overcome the lack of topological information which often hinders the study of complex networks. We tested the method on empirical networks, as well as on synthetic networks generated through the fitness model, and studied how well it can estimate the fundamental topological properties which are widely employed to describe network patterns: connectivity, assortativity, clustering coefficient and rich-club coefficient. We found that these properties are reconstructed accurately, for instance with a tolerance usually varying from 2% to
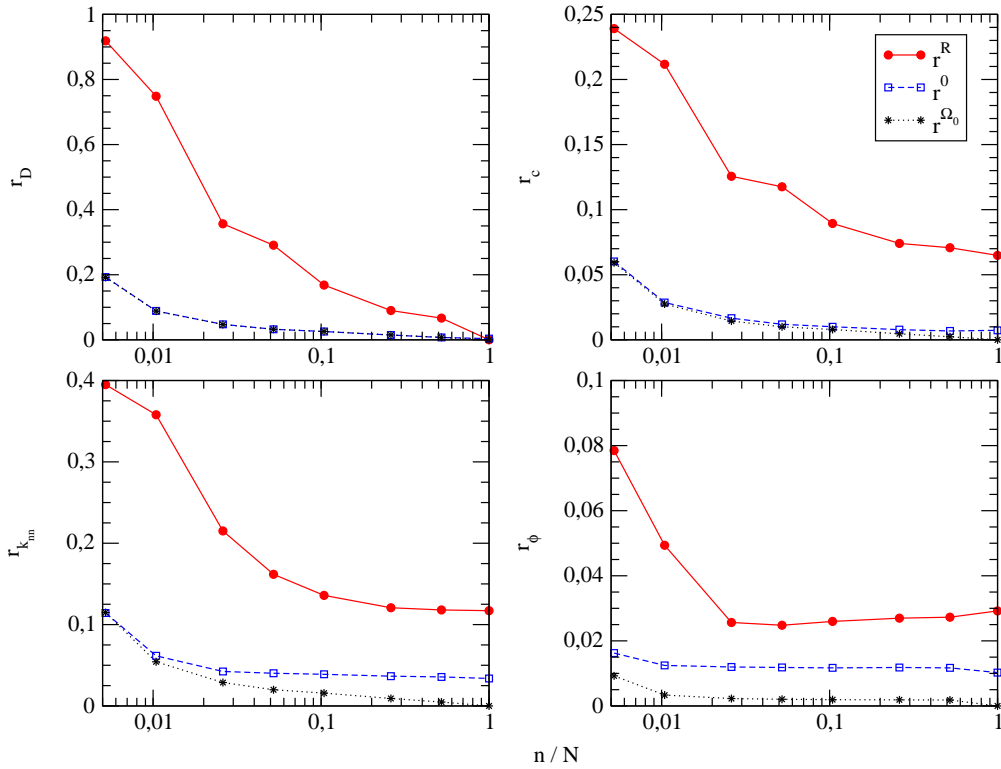
Figure 3: rRMSE of various topological properties for different values of n, obtained by the BS method on the real E-mid network and on its synthetic version obtained through the fitness model. Top left: density D; top right: clustering coefficient c; bottom left: average nearest-neighbor degree k_{nn}; bottom right: rich-club coefficient ϕ.

15% (depending on the property examined) using only 5% of the nodes. We also found that the BS method brings to better estimates in denser networks (where more information is available); additionally, the method effectiveness strongly depends on the accuracy of the fitness model used to describe the empirical dataset. In the case of the WTW, the fitness model is fairly accurate in describing how links are formed across countries depending on their GDP [13] and the BS is thus efficient in reconstructing the network topological properties. In the case of E-mid, the fitness model is less accurate and so is the BS method, but the latter can still lead to useful outcomes.

While at first thought it can be surprising that a small fraction of nodes enables to estimate with high accuracy global emerging properties of the network, it is important to remark that the BS method assumes the knowledge of the fitness of all nodes and the validity of the fitness model in describing the data. Therefore, a limitation of this method could arise when considering higher-order topological properties as the community structure. Possibly, in these situations the method could require a larger initial information to obtain the same results. Investigation of these cases is left for future research. Finally note that, contrarily to past works [2, 6, 7, 8, 9], in this paper we focused on estimating previously untested network quantities, namely

the most fundamental properties of a network. This allows us to enlarge the number of properties whose estimations are in remarkable agreement with observations. Indeed, any network reconstruction method must be tested against these basic quantities before being employed for more demanding tasks that are specific for a class of networks (such as systemic risk estimation in financial networks). Thus, the validation of the BS method we presented here also allows us to extend its applicability to any set of dependencies among components in a complex system.

Acknowledgments

References

[1] Clauset, A., Moore, C., Newman, M.: *Hierarchical structure and the prediction of missing links in networks.* Nature 453(7191), 98-101 (2008).

[2] Mastromatteo, I., Zarinelli, E., Marsili, M.: *Reconstruction of financial networks for robust estimation of systemic risk.* J. Stat. Mech. Theory Exp. 2012(03), P03011 (2012).

[3] Battiston, S., Gatti, D., Gallegati, M., Greenwald, B., Stiglitz, J.: *Liaisons dangereuses: increasing connectivity, risk sharing, and systemic risk.* J. Econ. Dyn. Control 36(8), 1121-1141 (2012).

[4] Battiston, S., Puliga, M., Kaushik, R., Tasca, P., Caldarelli, G.: *DebtRank: too central to fail? Financial networks, the fed and systemic risk.* Sci. Rep. 2, 541 (2012).

[5] Wells, S.: *Financial interlinkages in the United Kingdom's interbank market and the risk of contagion.* (Bank of England's Working paper 230, 2004).

[6] van Lelyveld, I., Liedorp, F.: *Interbank contagion in the dutch banking sector.* Int. J. Cent. Bank. 2, 99-134 (2006).

[7] Degryse, H., Nguyen, G.: *Interbank exposures: an empirical examination of contagion risk in the Belgian banking system.* Int. J. Cent. Bank. 3(2), 123-171 (2007).

[8] Mistrulli, P.: *Assessing financial contagion in the interbank market: maximum entropy versus observed interbank lending patterns.* J. Bank. Finance 35(5), 1114-1127 (2011).

[9] Musmeci, N., Battiston, S., Caldarelli, G., Puliga, M., Gabrielli, A.: *Bootstrapping Topological Properties and Systemic Risk of Complex Networks Using the Fitness Model.* J. Stat. Phys. 151(3-4), 720-734 (2013).

[10] Caldarelli, G., Chessa, A., Gabrielli, A., Pammolli, F., Puliga, M.: *Reconstructing a credit network.* Nature Physics 9, 125 (2013).

[11] Park, J., Newman, M.: *Statistical mechanics of networks.* Phys. Rev. E 70(6), 066117 (2004).

[12] Caldarelli, G., Capocci, A., De Los Rios, P., Muñoz, M.: *Scale-free networks from varying vertex intrinsic fitness.* Phys. Rev. Lett. 89(25), 258702 (2002).

[13] Garlaschelli, D., Loffredo, M.: *Fitness-dependent topological properties of the World Trade Web.* Phys. Rev. Lett. 93(18), 188,701 (2004).

[14] Garlaschelli D., Loffredo, M.: *Maximum likelihood: Extracting unbiased information from complex networks.* Phys. Rev. E 78, 015101 (2008).

[15] Gleditsch, K. S.: *Expanded Trade and GDP Data.* J. Confl. Res. 46(5), 712-724 (2002).

[16] De Masi, G., Iori, G., Caldarelli, G.: *A fitness model for the Italian Interbank Money Market.* Phys. Rev. E 74(6), 066112 (2006).

[17] Dorogovtsev, S.: *Lectures on complex networks.* Phys. J. 9(11), 51 (2010).

[18] Garlaschelli, D., Loffredo, M. I.: *Generalized Bose-Fermi Statistics and Structural Correlations in Weighted Networks.* Phys. Rev. Lett. 102, 038701 (2009).

[19] Squartini, T., Garlaschelli, D.: *Analytical maximum-likelihood method to detect patterns in real networks.* New Journ. Phys. 13, 083001 (2011).

[20] Garlaschelli, D., Battiston, S., Castri, M., Servedio, V., Caldarelli, G.: *The scale-free topology of market investments.* Physica A 350(2), 491-499 (2005).

[21] Watts, D. J., Strogatz, S.: *Collective dynamics of 'small-world' networks.* Nature 393 (6684) 440-442 (1998).

[22] Colizza, V., Flammini, A., Serrano, M. A., Vespignani, A.: *Detecting rich-club ordering in complex networks.* Nature Physics 2, 110-115 (2006).