Estimating topological properties of weighted networks from limited information

Giulio Cimini,[1, *] Tiziano Squartini,[1, †] Andrea Gabrielli,[1, 2] and Diego Garlaschelli[3]

[1]*Istituto dei Sistemi Complessi (ISC-CNR) UoS "Sapienza" University of Rome, P.le A. Moro 5, 00185 Rome (Italy)*
[2]*IMT Institute for Advanced Studies, Piazza San Ponziano 6, 55100 Lucca (Italy)*
[3]*Lorentz Institute for Theoretical Physics, University of Leiden, Niels Bohrweg 2, 9506 Leiden (Netherlands)*
(Dated: November 27, 2017)

A fundamental problem in studying and modeling economic and financial systems is represented by privacy issues, which put severe limitations on the amount of accessible information. Here we introduce a novel, highly nontrivial method to reconstruct the structural properties of complex weighted networks of this kind using only partial information: the total number of nodes and links, and the values of the strength for all nodes. The latter are used as *fitness* to estimate the unknown node degrees through a standard *configuration model*. Then, these estimated degrees and the strengths are used to calibrate an *enhanced configuration model* in order to generate ensembles of networks intended to represent the real system. The method, which is tested on real economic and financial networks, while drastically reducing the amount of information needed to infer network properties, turns out to be remarkably effective—thus representing a valuable tool for gaining insights on privacy-protected socioeconomic systems.

Reconstructing the statistical properties of a network when only partial information is available represents a key unsolved problem in the field of statistical physics of complex systems [1, 2]. Yet, addressing this issue can bring to many concrete applications. A paramount example is provided by financial networks, where nodes represent financial institutions and edges stand for the various types of financial ties—such as loans or derivative contracts. These ties result in dependencies among institutions and constitute the ground for the propagation of financial distress across the network. However, due to confidentiality issues, the information that regulators are able to collect on mutual exposures is very limited [3], and this hinders the analysis of the system resilience to the default or distress on one or more institutions—which depends on the structure of the whole network [4, 5]. Typically, the analysis of systemic risk has been pursued by trying to reconstruct the unknown links of the network using Maximum Entropy algorithms [6–8]. These approaches, also known as "dense reconstruction" methods, assume that the network is fully connected and estimate link weights via a maximum homogeneity principle, looking for the weighted adjacency matrix with minimal distance from the uniform matrix that also satisfies the imposed constraints—represented for instance by the budget of individual banks. The strongest limitation of these algorithms lies in the hypothesis that the network is fully connected. In fact, not only empirical networks show a very heterogeneous distribution of the connectivity, but such dense reconstruction was shown to lead to systemic risk underestimation [2, 8]. More refined methods like "sparse reconstruction" algorithms [2] allow to obtain a matrix with an arbitrary level of heterogeneity, but still leave open the question of what value of heterogeneity would be appropriate; moreover, even

when link density is correctly recovered, systemic risk is again underestimated because of the homogeneity principle used to build the network. A more recent approach [9, 10] instead uses the limited topological information available on the network to generate an ensemble of *exponential random graphs* (ERG) through the *configuration model* (CM) [11], where the Lagrange multipliers defining it are replaced by *fitnesses* [12], *i.e.*, node-specific properties assumed to be known—in a way similar to fitness-dependent network models [13]. The estimation of network properties is then carried out within such fitness-induced ensemble. This method overcomes the limitations of its predecessors, but it still suffers from the drawback of being applicable only to binary networks—whereas, the analysis of systemic risk is generally carried out within the weighted representation of the networked system.

Here we aim at overcoming all the limitations of these methods and build an innovative and effective procedure to reconstruct weighted networks, resorting on a minimal amount of available information: the total number of connections and the values of the strength for each node—which will play the role of node fitness. In a nutshell, our method consists in estimating the number of connections for each node via the standard CM calibrated on the fitnesses, and then in using these values as well as node strengths to assess individual link weights through an *enhanced configuration model* (ECM) [14]. To validate our method, we use two real instances of economic and financial systems. The first one is the World Trade Web (WTW) [15], *i.e.*, the network whose N nodes represent countries and whose L links represent trade volumes—so that the weight w_{ij} of the link between nodes i and j is the total monetary flux between these countries resulting from the import/export

among them [16]. The second one is the (E-mid) inter-bank money market [17], where now the nodes represent banks and w_{ij} is the total amount of loans (*i.e.*, of liquidity exchanged) between banks i and j [18]. In both cases, the strength of node i is defined as $s_i^* = \sum_j w_{ij}$, while its degree or number of partners is $k_i^* = \sum_j a_{ij}$ (where $a_{ij} := \lim_{\varepsilon \to 0}[1 + \varepsilon/w_{ij}]^{-1}$). Since we have full information on these networks, we will be able to assess unambiguously the accuracy of our method in estimating their topological properties.

Our network reconstruction procedure builds on two complementary network generation models. The CM [11], a particular class of ERG model [19], consists in generating an ensemble Ω_{CM} of networks which is maximally random—except for the ensemble average of the node degrees $\{\langle k_i \rangle_{\Omega_{CM}}\}_{i=1}^N$ that are constrained to the observed values $\{k_i^*\}_{i=1}^N$. The probability distribution over Ω_{CM} is defined via a set of Lagrange multipliers $\{x_i\}_{i=1}^N$ (one for each node), whose values can be set to satisfy the equivalence $\langle k_i \rangle_{\Omega_{CM}} \equiv k_i^* \; \forall i$ [20]. The ensemble probability that any two nodes i and j are connected is given by:

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

so that x_i quantifies the ability of node i to create links with other nodes. The ECM [14] is instead obtained by specifying both the mean degree and strength sequences $\{k_i^*\}_{i=1}^N$ and $\{s_i^*\}_{i=1}^N$. In this case, two Lagrange multipliers $\{a_i, b_i\}$ are associated to each node i, so that the ensemble probability q_{ij} that any two nodes i and j are connected and the ensemble average $\langle w_{ij} \rangle$ for the weight of such link become [21]:

$$q_{ij} = \frac{a_i a_j b_i b_j}{1 + a_i a_j b_i b_j - b_i b_i}, \qquad \langle w_{ij} \rangle = \frac{q_{ij}}{1 - b_i b_j}. \tag{2}$$

On the other hand, the *fitness* model [12] assumes the network topology to be determined by an intrinsic property (fitness) associated with each node. This approach has been successfully used in the past to model several economic networks, including the network of equity investments in the stock market [22], the E-mid [17] and the WTW [13]. Note that fitnesses are often used within the ERG framework provided an assumed connection between them and the Lagrange multipliers. Our method builds exactly on such assumption.

Given these ingredients, we now formulate the statistical procedure at the basis of our method. We aim at finding the most probable estimate for $X(G_0)$, *i.e.*, the value of a topological property X for the real network G_0 that we want to reconstruct. Such estimate has to rely on and be compatible with some constraints, given by the incomplete information we have on G_0: the total number of nodes N and links L, and the whole strength sequence $\{s_i^*\}_{i=1}^N$. We build on two important assumptions:
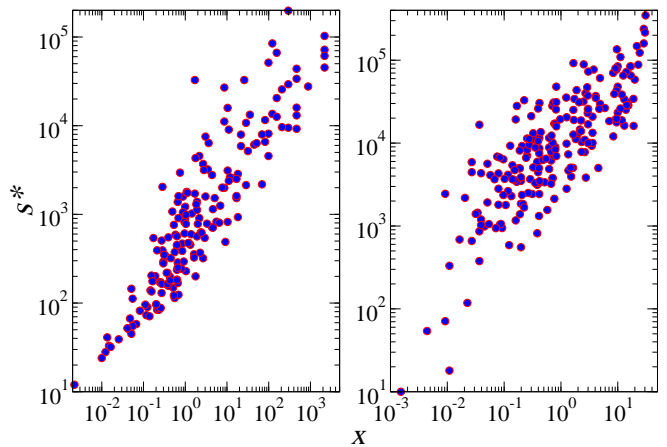

FIG. 1. Relation between node strengths $\{s^*\}$ and their degree-induced Lagrange multipliers $\{x\}$ from CM (obtained by knowing the whole degree sequence). The linearity of such relation is at the basis of assumption 2 of our method that $x_i \propto s_i^* \; \forall i$ accurately describes the binary network topology. Left panel refers to WTW, right panel to E-mid.

1. G_0 can be seen as drawn from an appropriate ECM ensemble Ω_{ECM}, so that $X(G_0)$ can be estimated as $\langle X \rangle_{\Omega_{ECM}}$;

2. The strengths $\{s_i^*\}_{i=1}^N$ represent degree-induced node fitnesses, and are thus assumed to be proportional to the Lagrange multipliers $\{x_i\}_{i=1}^N$ of the CM via a universal parameter z: $x_i \equiv \sqrt{z} s_i^* \; \forall i$.

The first assumption allows us to map the problem of evaluating $X(G_0)$ into that of choosing the optimal ECM ensemble Ω_{ECM} compatible with the known constraints on G_0. In other words, the question to address becomes: what ECM ensemble is the most appropriate to extract the real network G_0 from, given that we know only partial information? Then, once Ω_{ECM} is determined, we can use the average $\langle X \rangle_{\Omega_{ECM}}$ as a good estimation for $X(G_0)$. However, in order to build an appropriate Ω_{ECM}, we need to know not only the strengths but also the degrees for all the nodes, and this is where assumption 2 comes in handy: the unknown degrees can be estimated within a CM ensemble Ω_{CM} built using the strengths as degree-induced fitnesses (Figure 1).

Technically, our method consists in the following operative steps. I) We first find the unknown parameter z that defines Ω_{CM} (see assumption 2) by comparing the average number of links of a network belonging to Ω_{CM} with the (known) total number $L(G_0)$ of links in G_0:

$$\langle L \rangle_{\Omega_{CM}} = \frac{1}{2} \sum_i \sum_{j(\neq i)} \frac{z s_i^* s_j^*}{1 + z s_i^* s_j^*} \equiv L(G_0) \tag{3}$$

(since $\{s_i^*\}_{i=1}^N$ are known, eq. (3) is an algebraic equation in z). We then use this z to estimate the unknown degrees
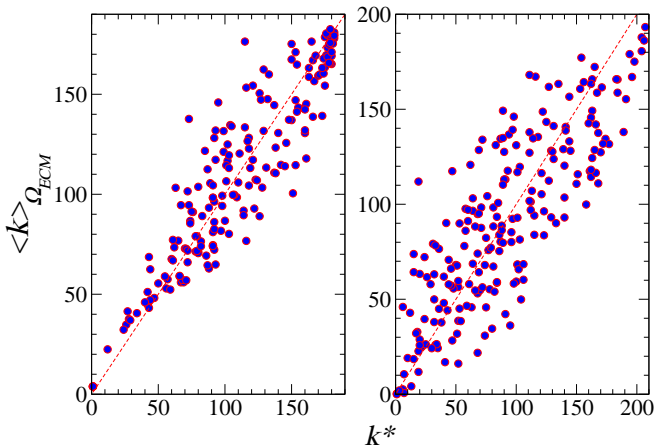
FIG. 2. Relation between k^* and $\langle k \rangle_{\Omega_{ECM}}$ for WTW (left panel) and E-mid (right panel).

through eq. (1):

$$\langle k_i \rangle_{\Omega_{CM}} = \sum_{j(\neq i)} p_{ij} = \sum_{j(\neq i)} \frac{z s_i^* s_j^*}{1 + z s_i^* s_j^*} \qquad \forall i. \qquad (4)$$

II) We use the degrees estimated in this way in the system of $2N$ nonlinear equations that define the ECM:

$$\begin{cases} \langle k_i \rangle_{\Omega_{CM}} = \displaystyle\sum_{j(\neq i)} \frac{a_i a_j b_i b_j}{1 + a_i a_j b_i b_j - b_i b_j} \\ s_i^*(G_0) = \displaystyle\sum_{j(\neq i)} \frac{a_i a_j b_i b_j}{(1 + a_i a_j b_i b_j - b_i b_j)(1 - b_i b_j)} \end{cases} \forall i.$$

$$(5)$$

The solution is the set of Lagrange multipliers $\{a_i, b_i\}_{i=1}^N$ that define the ECM ensemble—through the linking probabilities $\{q_{ij}\}_{i,j=1}^N$ and the average weights $\{\langle w_{ij} \rangle\}_{i,j=1}^N$ as of eq. (2)—and allow to compute $\langle X \rangle_{\Omega_{ECM}}$, either analytically or numerically.

In order to check whether Ω_{ECM} defined above is a proper ensemble to draw the real network G_0 from, we first compare for each node i the degree $k_i^*(G_0)$ of the real network and $\langle k_i \rangle_{\Omega_{ECM}} = \sum_{j(\neq i)} q_{ij}$ estimated through our method [23]. As Figure 2 shows, this results in a scattered cloud around the identity, whose behavior reflects the noisy yet very high correlation between strengths and degrees—as we are not using the real $k^*(G_0)$ in eq. (5) but $\langle k_i \rangle_{\Omega_{CM}}$ obtained from the CM induced by node strengths. We move further and focus on the topological properties that are commonly regarded as the most significant for describing a weighted network structure: the *average nearest neighbors strength*

$$s_i^{nn} := \frac{\sum_{j(\neq i)} a_{ij} s_j}{k_i} = \frac{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} w_{jk}}{\sum_{j(\neq i)} a_{ij}} \qquad (6)$$

and the *weighted clustering coefficient*

$$c_i^w := \frac{\sum_{j(\neq i)} \sum_{k(\neq i,j)} w_{ij} w_{ik} w_{jk}}{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{ik}}, \qquad (7)$$

together with the binary version of these quantities: the *average nearest neighbors degree*

$$k_i^{nn} := \frac{\sum_{j(\neq i)} a_{ij} k_j}{k_i} = \frac{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{jk}}{\sum_{j(\neq i)} a_{ij}} \qquad (8)$$

and the *binary clustering coefficient*

$$c_i^k := \frac{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{ik} a_{jk}}{\sum_{j(\neq i)} \sum_{k(\neq i,j)} a_{ij} a_{ik}}. \qquad (9)$$

The ECM ensemble averages for these quantities are obtained from eqn. (6-9) by replacing the binary adjacency matrix elements a_{ij} with the linking probabilities q_{ij}, and the real link weights w_{ij} with their ensemble averages $\langle w_{ij} \rangle$. Figure 3 shows a remarkable agreement between the values of these quantities computed on G_0 and their ECM ensemble averages—which can therefore be used as good estimates for the real quantities $X(G_0)$. Such test reveals the effectiveness of our method in reconstructing the topological properties of the real network.

It is important to remark that the applicability of our method strongly depends on the accuracy of assumption 2, *i.e.* on whether the CM induced by node strengths is able to provide good estimates for the unknown degrees. This is indeed the case of the WTW [13] and the E-mid [17], but also of other economic and financial networks of different nature [22]. Another important remark is that our method is based on a combination of CM and ECM rather than directly on the Weighted Configuration Model (WCM) [20], because the latter not only fails to reproduce the network topological properties (as shown by Figure 3), but also predicts a far denser network than observed. This happens not because strengths carry a "lower level" information than that of degrees—rather, they can be used to infer the degrees themselves, and this is what our method points out: the information on strength values should not be used to directly reconstruct the network, but to estimate the degree first, and only then to compute the quantities of interest. In this respect, note that using directly the knowledge of the strength sequence and number of links as fixed constraints to build a maximum-entropy ensemble would result in a different mathematical expressions. In particular, we would arrive at a variant of eq. (2) where $a_i = a$ $\forall i$. We have checked that, just like the WCM, this model gives a bad prediction of the network, leading to the conclusion that inferring the expected degrees first through eq. (4) is a crucial step of the approach we are using here: the information on links presence is indispensable to achieve a faithful network reconstruction.

Further work is needed to address several issues that remain open, including testing the accuracy of our method in estimating higher-order topological properties. Possibly, for these cases the method could require a larger initial information to obtain the same effectiveness. Nevertheless, in its present version our
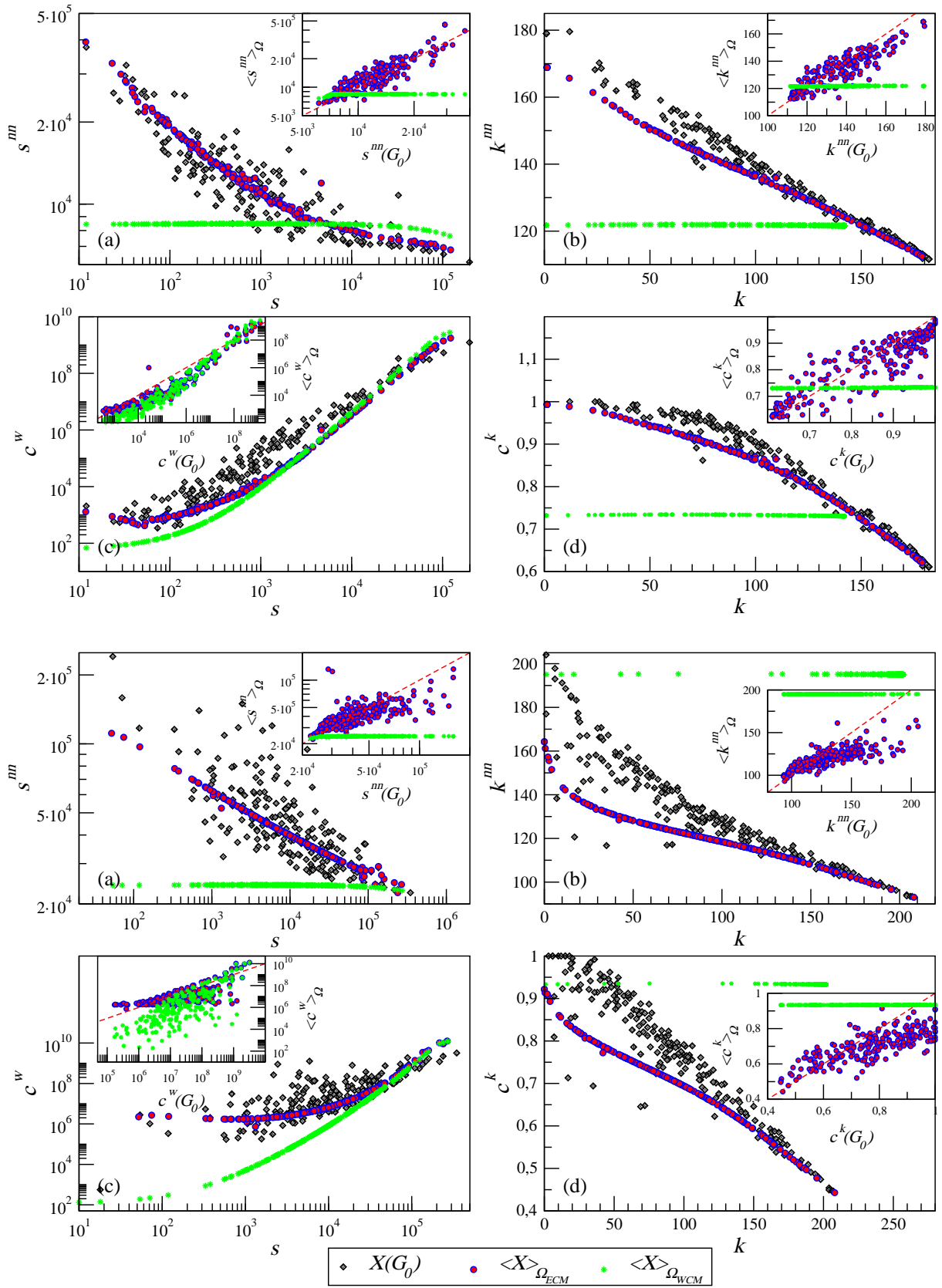
FIG. 3. Scatter plots of s vs s^{nn} (a), k vs k^{nn} (b), s vs c^w (c) and k vs c^k (d) for the real quantities ($X(G_0)$), those estimated by our method ($\langle X \rangle_{\Omega_{ECM}}$), and those computed by a WCM-based reconstruction ($\langle X \rangle_{\Omega_{WCM}}$). Insets: relations $X(G_0)$ vs $\langle X \rangle_{\Omega_{ECM}}$ and $X(G_0)$ vs $\langle X \rangle_{\Omega_{WCM}}$ for the same quantities. Upper plots refer to WTW, lower plots to E-mid.

method exploits a very limited information, which is indeed minimal but also often available for economic and

financial systems: besides global statistics (N and L), the strengths (that can be the operating revenue of firms, or the tier-1 capital of banks) are or should be accessible public data. In conclusion, our method is particularly useful to overcome the lack of topological information that often hampers systemic risk estimation in financial networks. More generally, our method can be applied to any network representing a set of dependencies among components in a complex system for which the available information is limited, and it is thus of general interest in the field of statistical physics of networks.

* giulio.cimini@roma1.infn.it

† tiziano.squartini@roma1.infn.it

[1] A. Clauset, C. Moore, M. E. J. Newman. Nature **453**(7191), 98-101 (2008).

[2] I. Mastromatteo, E. Zarinelli, M. Marsili. J. Stat. Mech. **2012**(03), P03011 (2012).

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

[4] S. Battiston, D. Gatti, M. Gallegati, B. Greenwald, J. Stiglitz. J. Econ. Dyn. Contr. **36**(8), 1121-1141 (2012).

[5] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, G. Cal-

darelli. Sci. Rep. **2**, 541 (2012).

[6] I. van Lelyveld, F. Liedorp. Int. J. Cent. Bank. **2**, 99-134 (2006).

[7] H. Degryse, G. Nguyen. Int. J. Cent. Bank. **3**, 123-171 (2007).

[8] P. Mistrulli. J. Bank. Fin. **35**(5), 1114-1127 (2011).

[9] N. Musmeci, S. Battiston, G. Caldarelli, M. Puliga, A. Gabrielli. J. Stat. Phys. **151**(3-4), 720-734 (2013).

[10] G. Caldarelli, A. Chessa, A. Gabrielli, F. Pammolli, M. Puliga. Nat. Phys. **9**, 125 (2013).

[11] J. Park, M. E. J. Newman. Phys. Rev. E **70**(6), 066117 (2004).

[12] G. Caldarelli, A. Capocci, P. De Los Rios, M. A. Muñoz. Phys. Rev. Lett. **89**(25), 258702 (2002).

[13] D. Garlaschelli, M. I. Loffredo. Phys. Rev. Lett. **93**(18), 188701 (2004).

[14] R. Mastrandrea, T. Squartini, G. Fagiolo, D. Garlaschelli. New J. Phys. **16** 043022 (2014).

[15] K. S. Gleditsch. J. Confl. Res. **46**(5), 712-724 (2002).

[16] We use trade volume data for year 2000, expressed in units of 10^8\$. Original volumes were divided by 10 in order to keep the $\{b_i\}_{i=1}^N$ in eq. (5) away from 1 and thus avoid numerical instability.

[17] G. De Masi, G. Iori, G. Caldarelli. Phys. Rev. E **74**(6), 066112 (2006).

[18] We consider snapshots of loans aggregated on annual scale (as also done in other works [17]) because of the high volatility of the links at shorter time scales. Here we report results for snapshots taken in 1999—yet, analysis for other annual snapshots brings to comparable results.

[19] S. Dorogovtsev. Phys. J. **9**(11), 51 (2010).

[20] T. Squartini, D. Garlaschelli. New J. Phys. **13**, 083001 (2011).

[21] D. Garlaschelli, M. I. Loffredo. Phys. Rev. Lett. **102**(3), 038701 (2009).

[22] D. Garlaschelli, S. Battiston, M. Castri, V. Servedio, G. Caldarelli. Phys. A **350**(2), 491-499 (2005).

[23] Comparing $\forall i$ node strength $s_i^*(G_0)$ and $\langle s_i \rangle_{\Omega_{ECM}} = \sum_{j(\neq i)} \langle w_{ij} \rangle$ is instead only a consistency check which returns, as it should, an identity (apart from small numerical errors).