SCIENTIFIC REPORTS

OPEN

# Systemic Risk Analysis on Reconstructed Economic and Financial Networks

Received: 26 January 2015
Accepted: 29 September 2015
Published: 28 October 2015

Giulio Cimini[1], Tiziano Squartini[1], Diego Garlaschelli[2] & Andrea Gabrielli[1,3]

**We address a fundamental problem that is systematically encountered when modeling real-world complex systems of societal relevance: the limitedness of the information available. In the case of economic and financial networks, privacy issues severely limit the information that can be accessed and, as a consequence, the possibility of correctly estimating the resilience of these systems to events such as financial shocks, crises and cascade failures. Here we present an innovative method to reconstruct the structure of such partially-accessible systems, based on the knowledge of intrinsic node-specific properties and of the number of connections of only a limited subset of nodes. This information is used to calibrate an inference procedure based on fundamental concepts derived from statistical physics, which allows to generate ensembles of directed weighted networks intended to represent the real system—so that the real network properties can be estimated as their average values within the ensemble. We test the method both on synthetic and empirical networks, focusing on the properties that are commonly used to measure systemic risk. Indeed, the method shows a remarkable robustness with respect to the limitedness of the information available, thus representing a valuable tool for gaining insights on privacy-protected economic and financial systems.**

The estimation of the structural properties of a complex network when the available information on the system is incomplete represents an unsolved challenge[1,2], yet it brings to many important applications. The most typical case is that of financial networks, whose nodes represent financial institutions and links stand for financial ties (*e.g.*, loans or derivative contracts)—the latter indicating dependencies among the institutions themselves, allowing for the propagation of financial distress across the network. The resilience of the system to the default or the distress of one or more institutions considerably depends on the topology of the whole network[3–5]; however, because of confidentiality issues, the information on mutual exposures that regulators are able to collect is very limited[6]. Systemic risk analysis has been typically pursued by reconstructing the unknown links of the network using maximum entropy approaches[7–9]. These methods are also known as "dense reconstruction" techniques because they assume that the network is fully connected—an hypothesis that represents their strongest limitation. In fact, not only real networks show a largely heterogeneous distribution of the connectivity, but such a dense reconstruction was shown to lead to systemic risk underestimation[2,9]. More refined techniques like "sparse reconstruction" algorithms[2] allow to obtain a network with arbitrary density, however they still underestimate systemic risk because of the homogeneity principle used to assign link weights. A more recent approach[10,11], which builds on even earlier results[12], instead uses the limited topological information on the network to generate an ensemble of graphs using the *configuration model* (CM)[13]—where, however, the Lagrange multipliers that define it are replaced by *fitnesses*, *i.e.* known intrinsic node-specific features[14]. The average values of the observables computed on the CM-induced ensemble are then used as estimates for the real network properties. The latter approach overcomes the heterogeneity issue described above, yet it only

[1]Istituto dei Sistemi Complessi (ISC-CNR) UoS "Sapienza" Università di Roma, 00185 Rome, Italy. [2]Lorentz Institute for Theoretical Physics, University of Leiden, 9506 Leiden, Netherlands. [3]IMT Institute for Advanced Studies, 55100 Lucca, Italy. Correspondence and requests for materials should be addressed to G.C. (email: giulio.cimini@roma1.infn.it)

allows to reconstruct systems in which each tie is undirected and unweighted—thus limiting the analysis to unrealistic and oversimplified configurations. Indeed, link directionality has been shown to play an important role in contagion processes and percolation analysis over these and other systems[15,16] by, *e.g.*, speeding up or confining the infection with respect to the undirected case. Since real economic and financial networks are, by their nature, directed, links directionality has to be taken into account when assessing their robustness to shock and crashes. Moreover, the connection weights between the entities of these systems often assume heterogeneous values, which in turn strongly affect the way such entities react to the default or distress of their interacting partners[4]. A recent study[17] has shown that, in order to satisfactorily reconstruct weighted networks, the procedure described above[10,11] cannot be extended naively by enforcing the corresponding weighted information, otherwise the reconstructed network is unrealistically dense[18]. Rather, one should employ a nontrivial combination of weighted and binary properties[17]. However, while this approach is feasible when such properties can be empirically accessed[17,19], it cannot be used when the system is privacy-protected (as in interbank and other financial networks).

In order to achieve a realistic and faithful reconstruction of economic and financial networks, here we develop an improved procedure that allows to reconstruct links directionality, and at the same time we implement an effective and self-consistent prescription to assign link weights. Our method can thus be employed specifically for systemic risk estimation, by assessing those network properties that have been shown to play a crucial role in contagion processes and in the propagation of distress over a networked system: the k-core structure[20], the percolation threshold[21], the mean shortest path length[22] and the DebtRank[4]. In particular, we perform an extensive analysis in order to quantify the accuracy of our method with respect to the size of the subset of nodes for which the topological information is available. Validation of the method is carried out on benchmark synthetic networks generated through a fitness-induced CM, as well as on two representative empirical systems, namely the International Trade Network or World Trade Web (WTW)[23] and the Electronic Market for Interbank Deposits (E-mid)[24]. In both cases, we have full information on these systems and we can thus unambiguously assess the accuracy of the method in reconstructing them.

## Previous approaches

Before explaining our method in detail, let us introduce some notation and recall previous results that we build upon. We will deal with weighted directed networks, *i.e.*, graphs composed by a set $V$ of nodes (with $|V| = N$) and described by a weighted directed adjacency matrix—whose generic element $w_{i\to j}$ represents the weight of the connection that runs from node $i$ to node $j$. The incoming total weight or *in-strength* for a generic node $i$ is then defined by $s_i^{in} = \sum_{j\in V} w_{j\to i}$, whereas, its outgoing total weight or *out-strength* reads $s_i^{out} = \sum_{j\in V} w_{i\to j}$. It is also convenient to introduce the binary directed adjacency matrix that describes the binary topology: $a_{i\to j} = \Theta[w_{i\to j}]$ ($\Theta$ is the Heaviside step function: $\Theta[a] = 1$ for $a > 0$ and $\Theta[a] = 0$ otherwise). This allows to define node $i$'s number of incoming connections or *in-degree* $k_i^{in} = \sum_{j\in V} a_{j\to i}$ and number of outgoing connections or *out-degree* $k_i^{out} = \sum_{j\in V} a_{i\to j}$. Finally, the binary undirected adjacency matrix—whose elements are obtained as $a_{ij} \equiv a_{ji} = \Theta\left[w_{i\to j} + w_{j\to i}\right]$—is used to define the number of incident connections or undirected *degree* of node $i$: $k_i = \sum_{j\in V} a_{ij} \equiv \sum_{j\in V} a_{ji}$.

In what follows, we are going to suppose that we only have partial information about the network: rather than knowing all the entries of the weighted adjacency matrix, we assume to know only local, node-specific information. In general, this information can be either topological (*e.g.*, the degrees of nodes[18]) or non-topological (*e.g.*, the economic size of nodes[12]). Before describing our specific implementation, we recall some important results that have been found previously using both schemes. At a binary network level, it has been shown that the topology of economic networks (including the ones we consider in this paper) can be accurately reconstructed from the knowledge of node degrees only[18,25,26]. Alternatively, since node degrees often turn out to be in an approximately monotonic (but highly nonlinear) relationship with some intrinsic economic property of nodes (like the GDP of countries in the WTW[12] or the portfolio volume in case of shareholding networks[27]), a good binary network reconstruction can be also achieved starting from the knowledge of such intrinsic node properties, rather than from node degrees themselves. The earliest and most clearcut illustration of this nontrivial result has been provided for the WTW[12,28], where it was shown that the observed topology can be reproduced from the knowledge of the GDP of all countries, plus the total number of links. This result, which supports the hypotheses of the *fitness model*[14], was later shown to remain valid even if one assumes to know the degrees of only a small subset of the nodes[10,29] (a framework known as *bootstrap* that we use also below), and if the analysis is extended to other financial systems such as interbank networks. The robustness of the reconstruction under bootstrap for interbank networks is very important for concrete applications, since knowing even only the total number of interbank connections is practically impossible, while knowing the degree of a few banks is in many cases easier[30]. Using the above technique, the level of systemic risk associated with the *binary* structure of a financial network can be estimated fairly well[10,29].

On the other hand, at a weighted network level the situation is much more complicated, and still unsolved at present. If one attempts to reconstruct the network starting from the strengths of nodes (the most direct proxy for nodes size) and without adjusting manually the network density[31–33], the result

is—depending on the methodology adopted—either a very dense network[18], or a completely connected one[6]. Indeed, in the latter case the link weights are assigned according to the so-called "gravity model" as:

$$\widetilde{w}_{i\to j} = \frac{s_i^{out} s_j^{in}}{W}, \tag{1}$$

where $W = \sum_{i,j\in V} w_{i\to j} = \sum_{i\in V} s_i^{out} \equiv \sum_{j\in V} s_j^{in}$ is the total observed weight[6]. The above formula shows that the reconstructed in-strength and out-strength of each node $i$, which are given by $\sum_{j\in V} \widetilde{w}_{j\to i}$ and $\sum_{j\in V} \widetilde{w}_{i\to j}$ respectively, coincide with the observed quantities $s_i^{in}$ and $s_i^{out}$ as desired. However, it also highlights that the reconstructed network is fully connected, a limitation that can be understood as the result of the fact that, in absence of purely topological information, the known total weight is redistributed over many more (all, in fact) pairs of nodes than those actually connected in the real network[17]. As we have mentioned in the Introduction, in the case of interbank networks this results in a very poor estimation of systemic risk.

Recently it has been shown that, in order to satisfactorily reconstruct a weighted network, one should simultaneously specify both node strengths and node degrees[17]. This results in an accurate reconstruction, however requires the knowledge of a lot of information. How to relax this requirement in an effective manner is an open question at the moment. For the WTW, a recent study[19] has shown that, as in the purely binary case[12], it is possible to reproduce both the topology and the weighted structure of the network by replacing the knowledge of the degree and strength sequence with that of the total number of links and total link weight respectively, plus the knowledge of the GDP of all countries. While powerful, this simplification is generally not feasible for financial networks[34]. In particular, for real interbank networks the full strength sequences (*i.e.*, total loans and liabilities) are typically publicly available—thus there is no need to assume that it is unknown, whereas, the total number of links is not (since, as we have already mentioned, it is feasible to collect information on the connectivity for only a subset of nodes). The aim of this paper is to introduce a reconstruction method that is appropriate for directed and weighted financial networks, and that allows to estimate systemic risk to a high level of accuracy.

## Method

In accordance with the above discussion, in this paper we are going to adopt a bootstrap-like scenario and assume incomplete information about the topology of a given network $G_0$. In particular, we suppose to know the in-degree and out-degree sequences $\{k_i^{in}\}_{i\in I}$ and $\{k_i^{out}\}_{i\in I}$ only for a subset $I \subset V$ of all nodes (where $|I| = n < N$). Moreover, we suppose to know a pair of properties $\{\chi_i\}_{i\in V}$ and $\{\psi_i\}_{i\in V}$ for *all* the nodes—that will be our *fitnesses*. These fitnesses should be thought of as intrinsic economic properties that are responsible for the inward (in-degree) and outward (out-degree) connectivity of nodes (see points I and II below); in this respect, it is quite straightforward (and actually very common[2,6–9,17,18,31–33]) to associate them with the nodes in- and out- strengths, respectively—but in general other proxies can be used. Given these ingredients, our network reconstruction method invokes a two-step statistical procedure (in which connection probabilities are estimated first, and link weights later) in order to find the most probable estimate for the value $X(G_0)$ of a given property $X$ computed on the network $G_0$, compatible with the constraints given by the aforementioned information we have on $G_0$.

First, we aim at reconstructing the binary topology of the network. To this end, we build on two important hypotheses.

I) The binary topology of $G_0$ is drawn from an ensemble $\Omega$ induced by a directed CM[25]—meaning that $\Omega$ is a set of binary directed networks that are maximally random, except for the ensemble averages (*i.e.*, expected values) of the in- and out- degrees $\left\{\langle k_i^{in}\rangle_\Omega\right\}_{i\in V}$ and $\left\{\langle k_i^{out}\rangle_\Omega\right\}_{i\in V}$ that are constrained to the observed values $\{k_i^{in}\}_{i\in V}$ and $\{k_i^{out}\}_{i\in V}$, respectively[13]. The directed CM prescribes that the probability distribution over $\Omega$ is defined via a set of Lagrange multipliers $\{x_i, y_i\}_{i\in V}$ (two for each node), whose values can be adjusted in order to satisfy the equivalence $\langle k_i^{in}\rangle_\Omega \equiv k_i^{in}$ and $\langle k_i^{out}\rangle_\Omega \equiv k_i^{out}, \forall i \in V$[25]. The values of $x_i$ and $y_i$ are thus induced by the in- and out- degree of node $i$, respectively, and the ensemble probability for a directed connection between any two nodes $i$ and $j$ reads[13]:

$$p_{i\to j} \equiv \left\langle \widetilde{a}_{i\to j}\right\rangle_\Omega = \frac{x_j y_i}{1 + x_j y_i}, \tag{2}$$

where $\widetilde{a}_{i\to j}$ is a particular realization of the reconstructed link, having expected value over the ensemble equal to $p_{i\to j}$: $\widetilde{a}_{i\to j} = 1$ with probability $p_{i\to j}$, and $\widetilde{a}_{i\to j} = 0$ otherwise. Eq. (2) thus shows that $x_i$ ($y_i$) quantifies the ability of node $i$ to receive incoming (form outgoing) connections.

II) The fitnesses $\{\chi_i\}_{i\in V}$ and $\{\psi_i\}_{i\in V}$ are assumed to be linearly correlated, respectively, to the in-degree-induced and out-degree-induced Lagrange multipliers $\{x_i\}_{i\in V}$ and $\{y_i\}_{i\in V}$ through universal (unknown) parameters $\alpha$ and $\beta$: $x_i \equiv \sqrt{\alpha}\,\chi_i$ and $y_i \equiv \sqrt{\beta}\,\psi_i$, $\forall\, i \in V$. Therefore eq. (2) becomes:

$$p_{i\to j} = \frac{\sqrt{\alpha}\,\chi_j\sqrt{\beta}\,\psi_i}{1 + \sqrt{\alpha}\,\chi_j\sqrt{\beta}\,\psi_i} = \frac{z\chi_j\psi_i}{1 + z\chi_j\psi_i}, \tag{3}$$

where we have defined $z \equiv \sqrt{\alpha\beta}$. Such an hypothesis is inspired by the *fitness model*[14], which assumes the network topology to be determined by intrinsic properties associated to each node of the network. We recall that this approach has been already used in the past to model several economic and financial networks[12,24], possibly within the CM framework assuming a connection between fitnesses and Lagrange multipliers[27].

These two hypotheses allow us to build the optimal CM ensemble $\Omega$ induced by the fitnesses $\{\chi_i\}_{i\in V}$ and $\{\psi_i\}_{i\in V}$, that is compatible with the binary constraints on $G_0$—given by the knowledge of $\{k_i^{in}\}_{i\in I}$ and $\{k_i^{out}\}_{i\in I}$. Indeed, because of the limited available information, finding the CM of the real system[13] is impossible, and we thus have to impose it by assigning *ad hoc* values to the Lagrange multipliers—whence the name "fitness-induced" CM (FiCM). In practice, since we know the fitness values $\{\chi_i, \psi_i\}_{i\in V}$, in order to determine unambiguously $\Omega$ we have to find the most likely value of the proportionality constant $z$ that defines $\Omega$ according to eq. (3). This can be done using the partial knowledge of the degree sequences to estimate the appropriate value of $z$ through a maximum-likelihood argument[12], *i.e.*, by comparing, for the nodes in the set $I$, the average number of incoming and outgoing connections in the ensemble $\Omega$ with their in-degrees and out-degrees observed in $G_0$:

$$\sum_{i\in I}\left[\langle k_i^{in}\rangle_\Omega + \langle k_i^{out}\rangle_\Omega\right] = \sum_{i\in I}[k_i^{in} + k_i^{out}]. \tag{4}$$

In the above expression, $\langle k_i^{in}\rangle_\Omega = \sum_{j(\neq i)} p_{j\to i}$ and $\langle k_i^{out}\rangle_\Omega = \sum_{j(\neq i)} p_{i\to j}$ contain the unknown parameter $z$ through eq. (3), and since $\{\chi_i, \psi_i\}_{i\in V}$ and $\{k_i^{in}, k_i^{out}\}_{i\in I}$ are known, eq. (4) defines an algebraic equation in $z$, whose solution allows to build the FiCM ensemble—even with the knowledge of the in- and out- degree of just a single node.

We now turn to reconstructing the weighted topology of $G_0$. A key ingredient of our approach will be the following consideration. As already mentioned, eq. (1) ensures that the reconstructed in- and out-strengths of all nodes are equal to the observed ones *only when the reconstructed network is fully connected*. However, if the topology is more complex (hence determined by a nontrivial probability $p_{i\to j}$ that node $i$ connects to node $j$), then in order to reproduce the observed strengths eq. (1) has to be modified as follows:

$$\widetilde{w}_{i\to j} = \frac{s_i^{out}s_j^{in}}{W p_{i\to j}}\widetilde{a}_{i\to j}, \tag{5}$$

This prescription ensures that the expected value of the reconstructed in-strength and out-strength of node $i$ are

$$\langle s_i^{in}\rangle_\Omega = \left\langle \sum_{j\in V}\widetilde{w}_{j\to i}\right\rangle_\Omega = \frac{s_i^{in}}{W}\sum_j \frac{s_j^{out}}{p_{j\to i}}\left\langle \widetilde{a}_{j\to i}\right\rangle \equiv s_i^{in}$$

and

$$\langle s_i^{out}\rangle_\Omega = \left\langle \sum_{j\in V}\widetilde{w}_{i\to j}\right\rangle_\Omega = \frac{s_i^{out}}{W}\sum_j \frac{s_j^{in}}{p_{i\to j}}\left\langle \widetilde{a}_{i\to j}\right\rangle \equiv s_i^{out} \tag{6}$$

as desired: eq. (5) ensures that the observed in- and out-strength sequences are correctly replicated by the method, irrespectively of whether the topology (as predicted by the set $\{p_{i\to j}\}_{i,j\in V}$) is reproduced. For instance, if $p_{i\to j} = 1\ \forall i, j \in V$ we recover the standard eq. (1)[6–9], whereas, if $p_{i\to j} = 1 - \lambda\ \forall i, j \in V$ we recover a variant of the sparse reconstruction method[2]. Our purpose here is ensuring that $p_{i\to j}$ correctly reconstructs the degree sequence, and hence both the binary and weighted topology of the network.

We formalize the above discussion as follows. In the most general case (*i.e.*, for generic node fitnesses), in order to obtain a weighted topology we place $\forall i, j$ a weight $\widetilde{w}_{i\to j}$ on the directed link from $i$ to $j$ according to the following prescription:

$$\widetilde{w}_{i\to j} = \frac{\chi_j\psi_i}{W p_{i\to j}}\widetilde{a}_{i\to j} \equiv \frac{1}{W}\left(z^{-1} + \chi_j\psi_i\right)\widetilde{a}_{i\to j}, \tag{7}$$

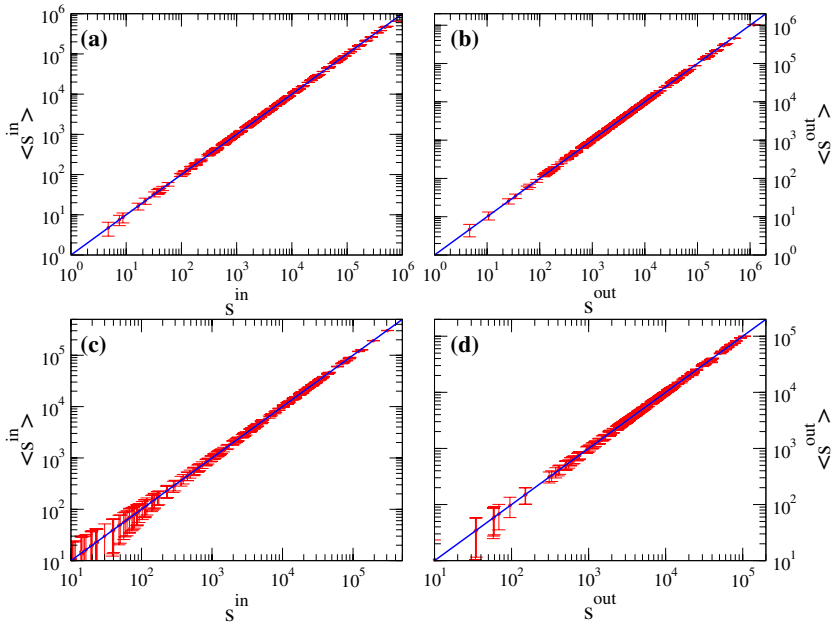


**Figure 1. Reconstruction of the strength sequences.** Scatter plots of node in-strengths $s^{in}$ and out-strengths $s^{out}$ observed for the real network $G_0$ and their ensemble averages obtained from eq. (8). Upper panels (**a**,**b**) refer to WTW, lower panels (**c**,**d**) to E-mid.

where the last equality comes from eq. (3). In this expression, the normalization $W$ represents the expected *induced total weight of the network*, defined as the geometric mean of the sum of the fitnesses: $W = \sqrt{(\sum_i \chi_i)(\sum_i \psi_i)}$. Indeed, this definition is consistent with $W \equiv \sum_{ij} \langle \widetilde{w}_{i\to j} \rangle_\Omega$, where $\langle \widetilde{w}_{i\to j} \rangle_\Omega = (\chi_j \psi_i) \langle \widetilde{a}_{i\to j} \rangle_\Omega / (W p_{i\to j}) = (\chi_j \psi_i)/W$. This procedure assures that the expected values of a node $i$'s total in- and out-strengths are directly proportional to $\chi_i$ and $\psi_i$, respectively and $\forall i \in V$. Now, using the natural interpretation of fitnesses as the empirical nodes strengths observed in $G_0$ ($\chi_i = s_i^{in}$ and $\psi_i = s_i^{out}, \forall i \in V$), brings to the situation described in the previous paragraph: $W = \sum_{i\in V} \chi_i \equiv \sum_{i\in V} \psi_i$, $\langle s_i^{in} \rangle_\Omega \equiv s_i^{in}$ and $\langle s_i^{out} \rangle_\Omega \equiv s_i^{out}$. We stress again that in this way we successfully preserve, on average, the strength sequences of the real network $G_0$ (and thus its total weight), as shown in Fig. 1. In other words, our network reconstruction method is based on a null model constraining the in-degree and out-degree sequence of a subset of nodes, together with the in-strength and out-strength sequence of the whole set of nodes. The final result is that the appropriate modification of the standard gravity model of eq. (1) is, as for eq. (5), the "degree-corrected gravity model":

$$\widetilde{w}_{i\to j} = \frac{z^{-1} + s_i^{out} s_j^{in}}{W} \widetilde{a}_{i\to j}. \tag{8}$$

With respect to eq. (1), eq. (8) has two important differences. On one hand, only the links that are actually created are assigned a non-zero weight; on the other hand, with respect to eq. (1) there is an extra offset $z^{-1}$ which depends (through $\{p_{i\to j}\}_{i,j\in V}$) on the observed density, and whose role is precisely that of redistributing the "missing" weight (required to reconstruct the desired in- and out-strengths) from the disconnected pairs of nodes to the connected ones. Remarkably, these modifications also allow to obtain much better estimates of higher order weighted network properties, as compared to the standard gravity approach (Fig. 2).

Finally, once the FiCM ensemble $\Omega$ is determined and link weights are placed, statistical mechanics of networks prescribes that the value $X(G_0)$ of property $X$ computed on $G_0$ typically varies in the range $\langle X \rangle_\Omega \pm \sigma_X^\Omega$, where $\langle X \rangle_\Omega$ and $\sigma_X^\Omega$ are respectively average and standard deviation of $X$ estimated over $\Omega$[13]. We can thus use $\langle X \rangle_\Omega$ as a good estimation for $X(G_0)$.

Summing up, the algorithm works as follows. Given a network $G_0$, two fitness values $\chi$ and $\psi$ for each of the $N$ nodes, and the in-degrees and out-degrees only for a subset $I$ of $|I| = n < N$ nodes:

- We compute the sum of the in-degrees and out-degrees of the nodes in $I$, and use it to obtain the value of $z$ by solving eq. (4);

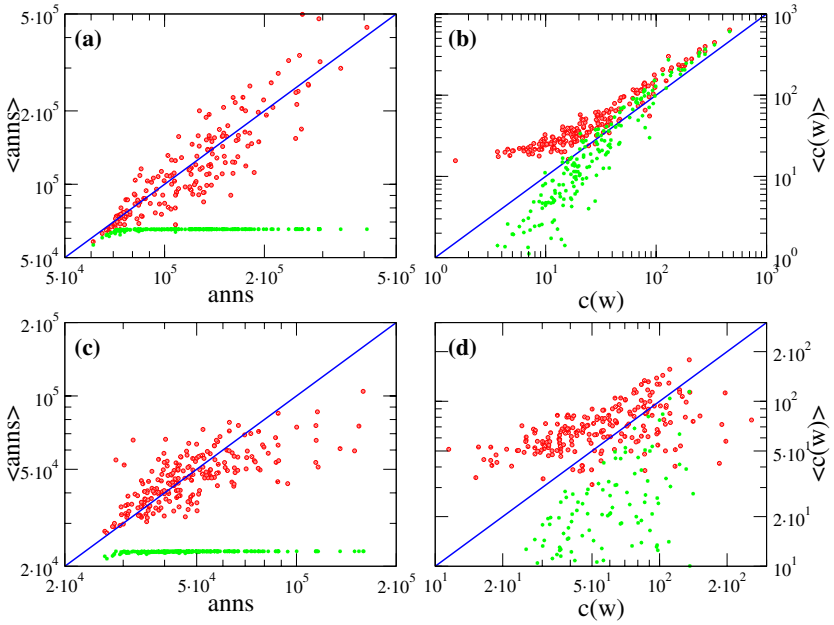


**Figure 2. Reconstruction of two higher order properties of directed weighted networks: the average nearest neighbor strength *anns* [panels (a,c)] and the weighted clustering coefficient *c*(*w*) [panels (b,d)] (refer to[37] for their formal definition).** Scatter plots of these quantities observed for the real network $G_0$ and their ensemble averages obtained from the degree-corrected gravity model of eq. (8) (red circles) or from the standard gravity model of eq. (1) (green asterisks). Upper panels (**a**,**b**) refer to WTW, lower panels (**c**,**d**) to E-mid. Remarkably, our degree-corrected gravity allows to obtain fairly accurate estimates for the *anns*, whereas, the standard gravity model completely fails in this respect as the resulting reconstructed network is fully connected. The degree-corrected gravity model outperforms the standard gravity model also in the reconstruction of $c(w)$. In this latter case, note that $\langle c(w)\rangle$ systematically overestimates the real $c(w)$, because in the definition of this quantity the number of reciprocal links plays an important role, yet it is slightly underestimated by the method (see Figs 5 and 6, and refer to the discussion in[26]).

- Using such $z$, we generate the FiCM ensemble $\Omega$ by placing a directed link from any node $i$ to any node $j$ with probability $p_{i\to j}$ of eq. (3), and assigning it with the corresponding weight $\tilde{w}_{i\to j}$ of eq. (7)—provided its existence;
- We compute the estimate of $X(G_0)$ as $\langle X\rangle_\Omega \pm \sigma_X^\Omega$ in the FiCM ensemble, typically numerically (*i.e.*, by measuring it on networks drawn from $\Omega$).

**Empirical Dataset.** In order to test our network reconstruction method, we use two representative empirical systems of economic and financial nature. The first one is the international trade network of the World Trade Web (WTW)[23], *i.e.*, the network whose nodes are the countries and links represent trade volumes between them: thus, $w_{i\to j}$ is the monetary flux from country $i$ to country $j$ (the "amount" of the export from $j$ to $i$). The second one is the (E-mid) Electronic Market for Interbank Deposits[24]: in this case, the nodes are banks and a link $w_{i\to j}$ from bank $i$ to bank $j$ represents the amount of the loan that $i$ granted to $j$.

In the following analysis we will use and show results for WTW trade volume data of year 2000, and E-mid aggregated transaction data of year 1999 (both temporal snapshots correspond to the largest size of the network). Analyses for other annual snapshots are reported in the Supplementary Information, and bring to comparable results. In the light of the above observations, we will use as fitnesses $\chi_i(\psi_i)$ the real node in-strength $s_i^{in} = \sum_{j\in V} w_{j\to i}$ (out-strength $s_i^{out} = \sum_{j\in V} w_{i\to j}$), *i.e.*, the total import (export) volumes of countries for WTW, and with the total liquidity borrowed (lent) by banks for E-mid. Note that the goodness of any choice for the fitness values must be first validated according to hypothesis II of our method (as discussed in the first part of section Results).

**Topological Properties.** As stated in the Introduction, we will test our network reconstruction method focusing on the network properties (each playing the role of $X$ in the discussion of section Methods) which are commonly regarded as the most significant for describing the network resilience to

systemic shocks and crashes. We first consider two properties defined for undirected networks (in order to reconstruct these properties, we use the undirected version of the method[10]):

- Degree of the main core $k^{main}$ and size of the main core $S^{main}$, where a $k$-core is defined as the "largest connected subgraph whose nodes all have at least $k$ connections" (within this subgraph), and the main core is the $k$-core with the highest possible degree ($k^{main}$)[35]. The main core is relevant to our analysis as it consists of the most influential spreaders (of, *e.g.*, an infection or a shock) in a network[20].
- Size of the giant component $S_{GC}$ at the bond percolation threshold $p^* = \overline{k}^{-1}$ ($\overline{k}$ is the mean degree of the network), where bond percolation is the process of occupying each link of the network with probability $p$, and $p^*$ is the critical value of $p$ at which a percolation cluster containing a finite fraction of all nodes first occurs[21]. Note that the percolation threshold at $p^* = \overline{k}^{-1}$ (that we take as reference value) is a feature proper of homogeneous graphs in the infinite volume limit, whereas, for scale-free networks in the same limit it is $p^* \to 0$. Note also that a bond percolation process can be mapped into a SIR model with infection rate $\beta$ and uniform infection time $\tau$. In fact, by defining the trasmissibility $T = 1 - e^{-\tau\beta}$ as the probability that the infection will be transmitted from an infected node to at least a susceptible neighbor before recovery takes place, the set of nodes reached by a SIR epidemic outbreak originated from a single node is statistically equivalent to the cluster of the bond percolation problem (with $p \equiv T$) the initial node belongs to[36].

We then move to properties defined for directed graphs:

- Link reciprocity $r$, measuring the tendency of node pairs to form mutual connections. It is defined as the ratio between the number of bidirected links and the total number of network connections: $r = \left(\sum_{ij} a_{i\to j} a_{j\to i}\right) / \left(\sum_{ij} a_{i\to j}\right)$. Reciprocity is considered a sensible parameter for systemic risk, giving a measure of direct mutual exposure among nodes.
- Average shortest path length $\lambda$[22], where the shortest path length $\lambda_{i\to j}$ from node $i$ to node $j$ is the minimum number of links required to connect $i$ to $j$ (following link directions), and $\lambda = N(N-1) / \left(\sum_{i\neq j} \lambda_{i\to j}^{-1}\right)$ (the harmonic mean is commonly used to avoid problems caused by pairs of nodes that are not reachable from one to another, and for which $\lambda$ diverges). This quantity measures the number of steps that are required, on average, for a signal or a shock to propagate between any two nodes of the network.
- The Group DebtRank $DR$[4], a measure of the total economic value in the network that is potentially affected by a distress on all nodes amounting to $0<\Phi<1$, with $\Phi = 1$ meaning default. In a nutshell, $DR$ is based on computing the recursive impact (*i.e.,* the reverberation on the network) of the initial distress, and is defined as:

$$DR = \sum_i (h_i^* - h_i^0)\nu_i \tag{9}$$

where $h_i^*$ is the final amount of distress on $i$ ($h_i^0 = \Phi$) and $\nu_i$ is the relative economic value of $i$. We refer to the original paper[4] for the details on how to compute $DR$, recalling here that $DR$ builds upon the detailed information on individual link weights in the network.

## Results

**Test of FiCM modeling.** When testing our network reconstruction procedure it is important to keep in mind that the method is subject to three different kind of errors. The first one comes from hypothesis I that the real network $G_0$ can be properly described by a CM, whose Lagrange multipliers are obtained by constraining the whole in-degree and out-degree sequences[13]. The second one derives instead from hypothesis II that the node fitnesses $\{\chi, \psi\}_{i\in V}$ are proportional to the CM's Lagrange multipliers $\{x, y\}_{i\in V}$, *i.e.*, from imposing a FiCM. Finally, the third one is due to the limited information available for calibrating the FiCM and obtain the true value of $z$—namely, the partial knowledge of the in-degree and out-degree sequences. Note however that the first source of mistakes cannot be controlled for in our context, as finding the CM that describes the data requires the knowledge of the whole in-degree and out-degree sequences (which is not accessible for our case studies). This is exactly why we have to make hypothesis II and impose a FiCM by assigning *ad hoc* values to the Lagrange multipliers. In this section we thus concentrate on the second source of errors.

Indeed, real networks are not perfect realizations of the FiCM and can only be approximated by it[12]. In order to assess qualitatively how well this FiCM describes the real network $G_0$, one can compare the observed in-degrees and out-degrees of $G_0$ with their averages $\langle k_i^{in}\rangle_\Omega$ and $\langle k_i^{out}\rangle_\Omega$ computed on the FiCM ensemble $\Omega$. Figure 3 shows such comparison when the average degrees are obtained through eq. (3) for a fully informed FiCM, *i.e.*, with the value of $z$ computed via eq. (4) using the knowledge of in- and out-degrees for all nodes. We indeed observe a remarkable agreement between these quantities for our empirical networks: the real degrees are scattered around the functional form of their expected values.

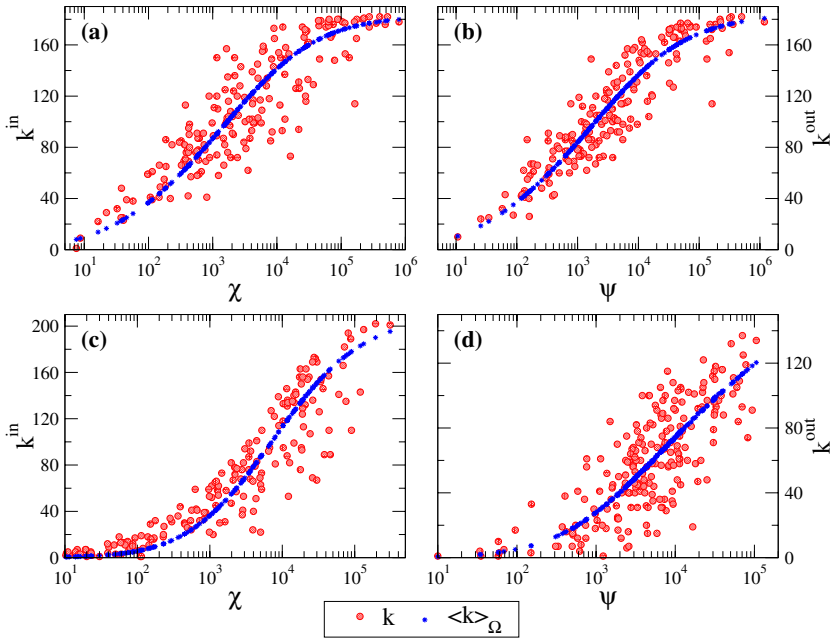


**Figure 3. Qualitative assessment for the FiCM description of the real network $G_0$.** Scatter plots of node fitnesses $\{\chi, \psi\}$ versus real node in- and out-degrees $\{k^{in}, k^{out}\}$ of $G_0$ (red circles) and their ensemble averages computed via the FiCM (blue asterisks). Upper panels (**a**,**b**) refer to WTW, lower panels (**c**,**d**) to E-mid.

The amount of deviations from perfect correlation (which would correspond to an actual realization of the FiCM) gives an indication of how well our model describes the real network. Note that the validity of hypothesis II can be evaluated also in the case of partial information by performing such comparison on the subset $I$ of nodes whose topological properties are available.

In the following, in order to have a quantitative global assessments of the errors caused by hypothesis II, we will test our network reconstruction method both on real networks and on benchmark synthetic networks numerically generated with the fully informed FiCM through eq. (3). In the latter case, the errors made by the method will be due only to the limited information available about the degree sequences. It is then interesting to check whether such generated synthetic networks are equivalent to the real networks in term of systemic risk. Figure 4 shows that bond percolation properties, shortest path length distribution and DebtRank values of synthetic networks are in excellent agreement with those of their real counterparts (the correlation coefficients between real and synthetic curves are all above 0.99). FiCM thus proves itself to be a proper framework for modeling our empirical networks.

**Test against limited information.** In this section we finally proceed to the key testing of the method against the third (and more relevant) source of errors: the limitedness of the information available on the degree sequences for calibrating the FiCM. In order to obtain a quantitative estimation of the method's effectiveness in reconstructing a topological property $X$ of a given a network $G_0$ (which can be either the real one or its synthetic version), we implement a procedure consisting in the following operative steps:

- Choose a value of $n < N$ (the number of nodes for which the in- and out- degrees are known).
- Build a set of $M = 100$ subsets $\{I_\alpha\}_{\alpha=1}^{M}$ of $n$ nodes picked at random from $G_0$.
- For each subset $I_\alpha$, use the degree sequences from $G_0$ to evaluate $z$ from eq. (4), and name such value $z_\alpha$.
- Build the ensemble $\Omega(z_\alpha)$ using the linking probabilities from eq. (3): generate $m = 100$ networks from $\Omega(z_\alpha)$, and compute the average value $X_\alpha$ of property $X$ on this ensemble.
- Compute the relative root mean square error (rRMSE) of property $X$ over the subsets $\{I_\alpha\}_{\alpha=1}^{M}$:

$$\rho[X] = rRMSE_X \equiv \sqrt{\frac{1}{M}\sum_{\alpha=1}^{M}\left[\frac{X_\alpha}{X(G_0)} - 1\right]^2} \tag{10}$$

where $X(G_0)$ is the value of $X$ measured on $G_0$.

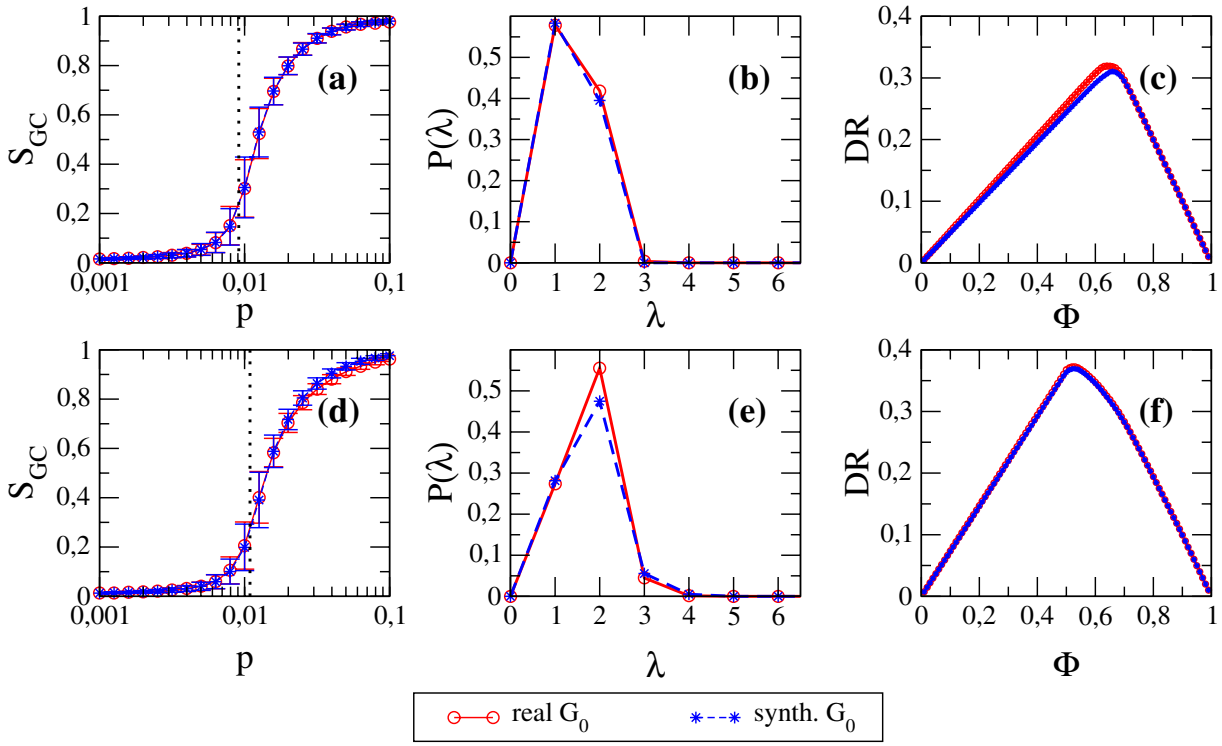


**Figure 4. Properties of real and synthetic networks.** Left plots (**a**,**d**): dependence of the size of the giant component $S_{GC}$ on the occupation probability $p$ (the vertical dotted line indicates $p^*$). Central plots (**b**,**e**): probability distribution of the directed shortest path length $\lambda$. Right plots (**c**,**f**): dependence of $DR$ on the initial distress $\Phi$. Top panels (**a**–**c**) refer to WTW, bottom panels (**d**–**f**) to E-mid. Correlation coefficient values $c$ between real and synthetic curves: (**a**) $c = 0.999$, (**b**) $c = 0.999$, (**c**) $c = 0.994$, (**d**) $c = 0.999$, (**e**) $c = 0.989$, (**f**) $c = 0.998$.

We then study how the rRMSE for the various network properties we consider varies as a function of the size $n$ of the subset of nodes used to calibrate the FiCM (*i.e.*, for which in- and out- degree information is available). Results are shown in Figs 5 and 6. We observe that in most of the cases there is a rapid decrease of the relative error as the number of nodes $n$ used to reconstruct the topology increases. For instance, generally the error drops to half of the starting rRMSE (for $n = 1$) at $n/N = 5\%$, and to one quarter for $n/N = 10\%$—a value that is rather close to that of the final error made at $n \equiv N$. This is an indication of the goodness of the estimation provided by our method. As expected, the rRMSE is higher for real networks than for synthetic networks, and the difference between the two curves gives a quantitative estimation of the error made in modeling real networks with the FiCM. The fact that such a difference is higher for E-mid than for WTW is directly related to a slightly better correlation between real and expected degrees observed in the latter case (Fig. 3). Note that the various rRMSE for synthetic networks do not necessarily tend to zero, because the generated synthetic configuration might be highly improbable—in some cases, the synthetic network can be even more atypical than the real one. We thus indicate with error bars the range of performance of our method for different choices of synthetic $G_0$.

Generally, $S_{GC}$, $\lambda$, $k_{main}$ and $S_{main}$ are the properties which are reconstructed better: for instance, with the knowledge of only 10% of the nodes, all the relative errors become smaller than 10%, and they decrease for increasing $n$. The rRMSE for $r$ and $DR$ show instead a behavior almost flat in $n$. The fact that the rRMSE for $r$ computed for real networks remains steadily high is probably due to the fact that reciprocity is hardly reproduced by a directed CM, and is better suited as additional imposed constraint[26]. The rRMSE for $DR$ is instead remarkably small for real networks (with values around 0.5%), and we can thus conclude that our method is efficient in estimating $DR$ also when the available information is minimal. This is particularly relevant to our analysis, since we are estimating $DR$ at its peak (*i.e.*, at its maximum, and thus mostly fluctuating, value), where the details of the weighted topology play a fundamental role in the process of risk propagation. Besides, and more importantly, the value of $DR$ for the real network is computed using the original weighted topology, whereas, the computation of $DR$ in the reconstructed network builds on link weights obtained by the degree-corrected gravity prescription of eq. (7).

In conclusion, the outcome of this analysis is that our network reconstruction method is able to estimate the network properties related to systemic risk with good approximation, by using the information

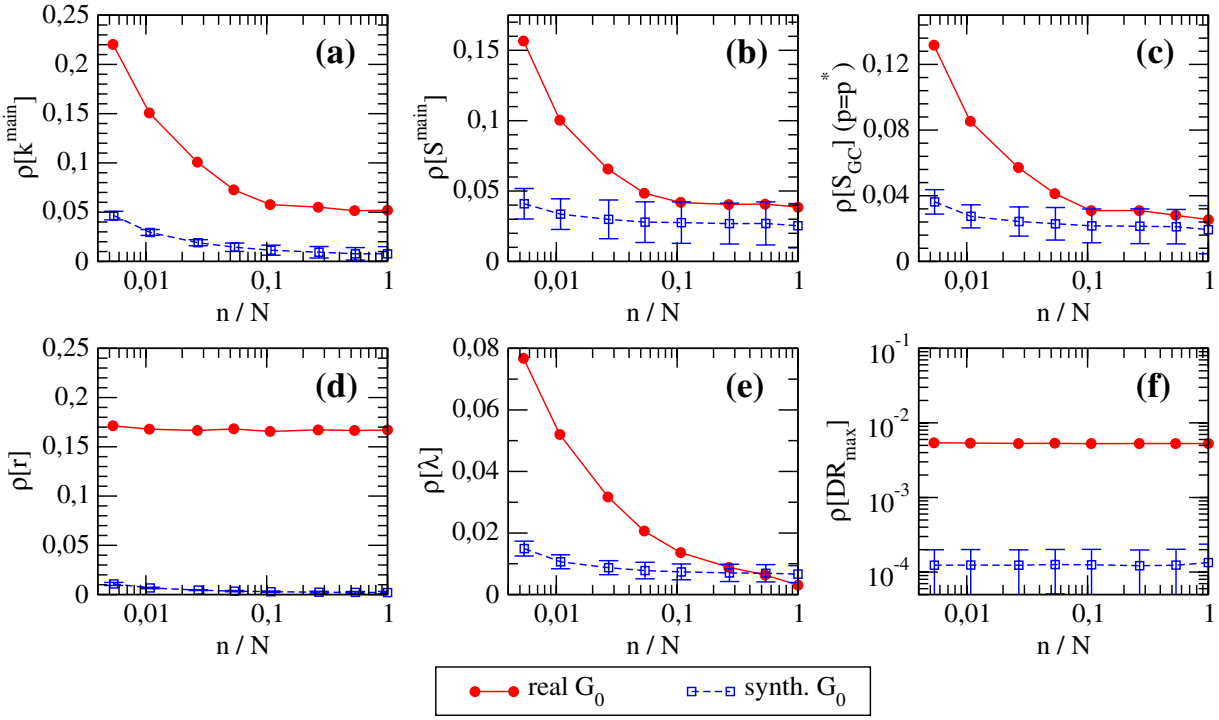


**Figure 5. rRMSE of various topological properties versus *n* for the reconstructed $G_0$ of the WTW (both the real network and its synthetic version).** rRMSE for: (**a**) degree of the main core $k^{main}$, (**b**) size of the main core $S^{main}$, (**c**) size of the giant component $S_{GC}$ at the bond percolation threshold $p^* = \overline{k}^{-1}$, (**d**) link reciprocity $r$, (**e**) mean shortest path length $\lambda$, (**f**) maximum value of the group DebtRank $DR_{max}$.

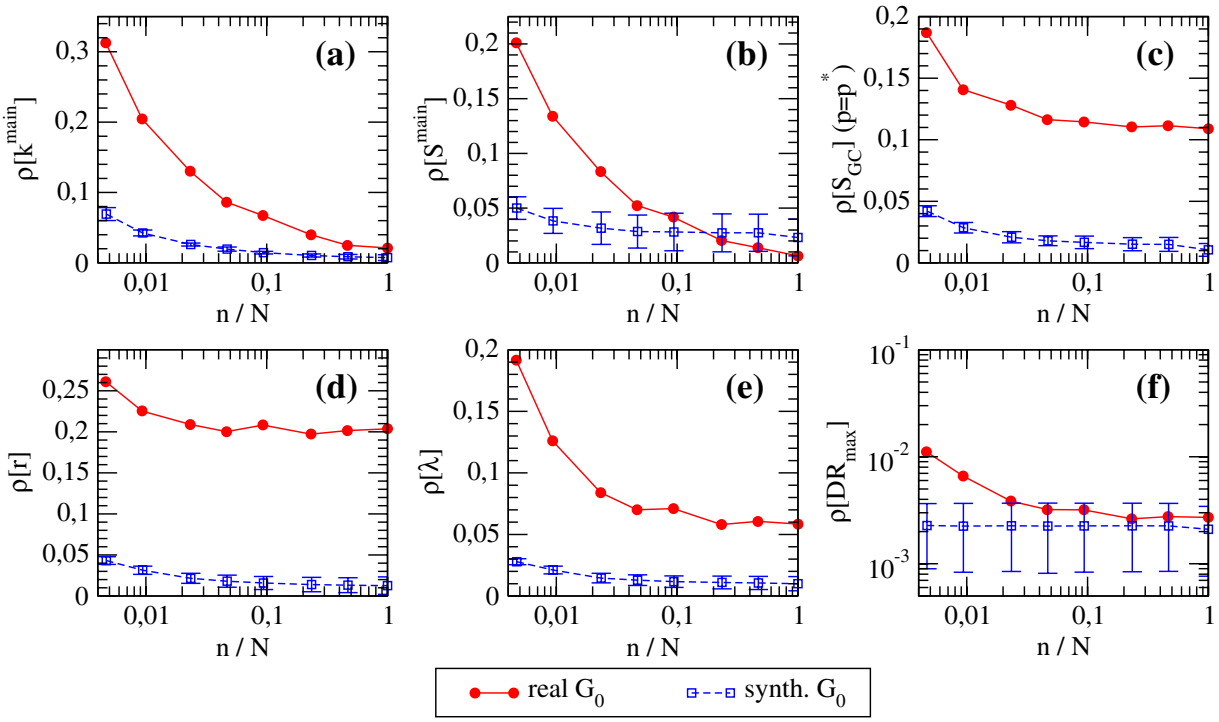


**Figure 6. rRMSE of various topological properties versus *n* for the reconstructed $G_0$ of the E-mid (both the real network and its synthetic version).** rRMSE for: (**a**) degree of the main core $k^{main}$, (**b**) size of the main core $S^{main}$, (**c**) size of the giant component $S_{GC}$ at the bond percolation threshold $p^* = \overline{k}^{-1}$, (**d**) link reciprocity $r$, (**e**) mean shortest path length $\lambda$, (**f**) maximum value of the group DebtRank $DR_{max}$.

on the number of connections of a relatively small fraction of nodes—as long as the fitnesses of all nodes is known.

## Discussion

In this paper we studied a novel method that allows to reconstruct a directed weighted network and estimate its topological properties by using only partial information about its connection patterns, as well as two additional intrinsic properties (interpreted as fitnesses) associated to each node. Tests on empirical networks as well as on synthetic networks generated through a fitness-induced configuration model reveal that the method is highly valuable for overcoming the lack of topological information that often hinders the estimation of systemic risk in economic and financial systems. Indeed, the information exploited by the method is minimal but is (or should be) publicly available for these kind of systems.

Our work originates from the study of Garlaschelli and Loffredo[12] and of Musmeci *et al.*[10]. The latter in particular represented a first attempt in tackling the problem of network reconstruction from partial information within the framework of fitness-induced configuration models. Here however we make fundamental improvements to the method, the key advance being that of extending it to directed weighted networks (the most general class of networks). In the present form, the method is then suited to reconstruct high-order network properties related to systemic risk, a task of primary practical importance the method was conceived to address—that was however beyond the reach of its original version. Besides, the validation of the fitness-induced configuration model approach to model real networks, as well as the reconstruction of benchmark synthetic networks generated as fitness-based counterparts of the empirical networks, are both novel ingredients that allow to assess quantitatively the accuracy of the method. Last but not least, the extensive analysis of different temporal snapshots of the real networks we provide in the Supplementary Information allows to strengthen considerably the effectiveness and robustness of our method.

We remark that the method we are proposing here, by reproducing both the binary and weighted topology of the network, represent a substantial step forward in the field of network reconstruction. In fact, most of the previous works[6–9,25] focused on reproducing the strengths of the real network to the detriment of connection patterns, whereas, only recently it has been realized that a successful reconstruction procedure must resort also on topological constraints[2,10]. Here we are proposing a method that allows to always reproduce the strengths, but also to tune the network topology through appropriate connection probabilities. In this respect, the use of probabilities derived from degree constraints represent the most general case, which include as specific instances both the dense reconstruction[6–9] and the sparse reconstruction[2] techniques.

Note that one should not be much surprised that the knowledge of a small number of nodes allows to precisely estimate a wide range of network properties, because the method assumes the additional knowledge of the fitness parameters for all the nodes. Besides, the effectiveness of the method strongly depends on the accuracy of the fitness model used to calibrate the CM in order to fit the empirical dataset. In the case of WTW and E-mid, the fitness model well describes how links are established across nodes, and our method is thus effective in reconstructing the network properties. Finally, we remark that the issue of having limited information on the system under investigation, while being typical for social, economic and financial systems (that are privacy-protected), is very relevant also for biological systems such as ecological networks, metabolic networks and functional brain networks—where, due to observational limitations and high experimental costs for collecting data, detailed topological information about connections is often missing. Notably, our method can be used to reconstruct any network representing a set of (directed and weighted) dependencies among the constituents of a complex system, and we thus believe it will find wide applicability in the field of complex networks and statistical physics of networks.

## References


1. Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 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,** 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,** 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. Fouque, J. P. & Langsam, J. A. (Eds). *Handbook on Systemic Risk* (Cambridge University Press, 2013).
6. Wells, S. Financial interlinkages in the United Kingdom's interbank market and the risk of contagion. *Bank of England's Working paper* **230** (2004).
7. van Lelyveld, I. & Liedorp, F. Interbank contagion in the dutch banking sector. *Int. J. Cent. Bank.* **2,** 99–134 (2006).
8. Degryse, H. & Nguyen, G. Interbank exposures: an empirical examination of contagion risk in the Belgian banking system. *Int. J. Cent. Bank.* **3,** 123–171 (2007).
9. Mistrulli, P. Assessing financial contagion in the interbank market: Maximum entropy versus observed interbank lending patterns. *J. Bank. Finance* **35,** 1114–1127 (2011).
10. 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,** 720–734 (2013).
11. Caldarelli, G., Chessa, A., Gabrielli, A., Pammolli, F. & Puliga, M. Reconstructing a credit network. *Nature Physics* **9,** 125 (2013).
12. Garlaschelli, D. & Loffredo, M. Fitness-dependent topological properties of the World Trade Web. *Phys. Rev. Lett.* **93,** 188701 (2004).

13. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70,** 066117 (2004).
14. Caldarelli, G., Capocci, A., De Los Rios, P. & Muñoz, M. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89,** 258702 (2002).
15. Boguñá, M. & Serrano, M. A. Generalized percolation in random directed networks. *Phys. Rev. E* **72,** 016106 (2005).
16. Meyers, L. A., Newman, M. E. J. & Pourbohloul, B. Predicting epidemics on directed contact networks. *J. Theo. Biol.* **240,** 400–418 (2006).
17. Mastrandrea, R., Squartini, T., Fagiolo, G. & Garlaschelli, D. Enhanced reconstruction of weighted networks from strengths and degrees. *New J. Phys.* **16,** 043022 (2014).
18. Fagiolo, G., Squartini, T. & Garlaschelli, D. Null models of economic networks: The case of the World Trade Web. *J. Econ. Interac. Coord.* **8,** 75–107 (2012).
19. Almog, A., Squartini, T. & Garlaschelli, D. A GDP-driven model for the binary and weighted structure of the International Trade Network. *New J. Phys.* **17,** 013009 (2015).
20. Kitsak, M., Gallos, L., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. & Makse H. Identification of influential spreaders in complex networks. *Nat. Phys.* **6,** 888–893 (2010).
21. Barrat, A., Barthékemy, M. & Vespignani, A. *Dynamical Processes On Complex Networks* (Cambridge University Press, 2008).
22. Bellman, R. On a routing problem. *Quart. Appl. Math.* **16,** 87–90 (1958).
23. Gleditsch, K. S. Expanded trade and GDP data. *J. Confl. Res.* **46,** 712–724 (2002).
24. De Masi, G., Iori, G. & Caldarelli, G. A fitness model for the Italian Interbank Money Market. *Phys. Rev. E* **74,** 066112 (2006).
25. Squartini, T. & Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *New J. Phys.* **13,** 083001 (2011).
26. Squartini, T. & Garlaschelli, D. Triadic motifs and dyadic self-organization in the World Trade Network. *Lec. Not. Comp. Sci.* **7166,** 24–35 (2012).
27. Garlaschelli, D., Battiston, S., Castri, M., Servedio, V. & Caldarelli, G. The scale-free topology of market investments. *Physica A* **350,** 491–499 (2005).
28. Garlaschelli, D. & Loffredo, M. Structure and evolution of the world trade network. *Physica A* bf **355**, 138–144 (2005).
29. Cimini, G. *et al.* Reconstructing topological properties of complex networks using the fitness model. *Lec. Not. Comp. Sci.* **8852,** 323–333 (2015).
30. Furfine, C. H. Interbank exposures: Quantifying the risk of contagion. *Journal of Money, Credit and Banking* **35,** 111–128 (2003).
31. Montagna, M. & Lux, T. Contagion risk in the interbank market: A probabilistic approach to cope with incomplete structural information. *Kiel Working Papers* **1937** (2014).
32. Battiston, S., D'Errico, M., Gurciullo, S. & Caldarelli, G. Leveraging the network: A stress-test framework based on DebtRank. arXiv:1503.00621 (2015).
33. Bardoscia, M., Battiston, S., Caccioli, F. & Caldarelli, G. DebtRank: A microscopic foundation for shock propagation. *PLoS ONE* **10(6),** e0130406 (2015).
34. Cimini, G., Squartin, T., Gabriell, A. & Garlaschelli,, D. Estimating topological properties of weighted networks from limited information. *Phys. Rev.* **E 92,** 040802(R) (2015).
35. Dorogovtsev, S. Lectures on complex networks. *Phys. J.* **9,** 51 (2010).
36. Newman, M. E. J. Spread of epidemic disease on networks. *Phys. Rev. E* **66,** 016128 (2002).
37. Fagiolo, G., Reyes, J. & Schiavo, S. On the topological properties of the world trade web: A weighted network analysis. *Physica A* **387,** 3868–3873 (2008).


## Acknowledgements

This work was supported by the EU project GROWTHCOM (611272), the Italian PNR project CRISIS-Lab, the EU project MULTIPLEX (317532) and the Netherlands Organization for Scientific Research (NWO/OCW). DG acknowledges support from the Dutch Econophysics Foundation (Stichting Econophysics, Leiden, the Netherlands) with funds from beneficiaries of Duyfken Trading Knowledge BV (Amsterdam, the Netherlands). We thank Guido Caldarelli and Stefano Battiston for useful discussion.

## Author Contributions

G.C. wrote the main manuscript text, and prepared figures 1–6 and the Supplementary Information. G.C. and T.S. performed the experiment. All authors designed the experiment, analyzed results and reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at http://www.nature.com/srep

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article**: Cimini, G. *et al.* Systemic Risk Analysis on Reconstructed Economic and Financial Networks. *Sci. Rep.* **5**, 15758; doi: 10.1038/srep15758 (2015).