INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

Enhanced network reconstruction from irreducible local information

Rossana Mastrandrea*
Tiziano Squartini§
Giorgio Fagiolo*
Diego Garlaschelli§

*Institute of Economics and LEM, Scuola Superiore Sant'Anna, Pisa, Italy
§Instituut-Lorentz for Theoretical Physics, University of Leiden, The Netherlands

2013/16 **July 2013**

ISSN (online) 2284-0400

Enhanced network reconstruction from irreducible local information

Rossana Mastrandrea
Institute of Economics and LEM, Scuola Superiore Sant'Anna, 56127 Pisa (Italy)

Tiziano Squartini
Instituut-Lorentz for Theoretical Physics, University of Leiden, 2333 CA Leiden (The Netherlands)

Giorgio Fagiolo
Institute of Economics and LEM, Scuola Superiore Sant'Anna, 56127 Pisa (Italy)

Diego Garlaschelli
Instituut-Lorentz for Theoretical Physics, University of Leiden, 2333 CA Leiden (The Netherlands)
(Dated: July 8, 2013)

Network topology plays a key role in many phenomena, from the spreading of diseases to that of financial crises. Whenever the whole structure of a network is unknown, one must resort to reconstruction methods that identify the least biased ensemble of networks consistent with the partial information available. A challenging case is when there is only local (node-specific) information available. For binary networks, the relevant ensemble is one where the degree (number of links) of each node is constrained to its observed value. However, for weighted networks the problem is much more complicated. While the naive approach prescribes to constrain the strengths (total link weights) of all nodes, recent counter-intuitive results suggest that in weighted networks the degrees are often more informative than the strengths, and as 'fundamental' as the latter. This implies that the reconstruction of weighted networks would be significantly enhanced by the specification of both quantities, a computationally hard and bias-prone procedure. Here we solve this problem by introducing an analytical and unbiased maximum-entropy method that works in the shortest possible time and does not require the explicit generation of reconstructed samples. We consider several real-world applications and show that, while the strengths alone give poor results, the additional knowledge of the degrees yields accurately reconstructed networks. Information-theoretic criteria rigorously confirm that the binary information is irreducible to the weighted one. Our results have strong implications for the analysis of motifs and communities and whenever the reconstructed ensemble is required as a null model to detect higher-order patterns.

A range of phenomena of critical importance, from the spread of infectious diseases to the diffusion of opinions and the propagation of financial crises, is highly sensitive to the topology of the underlying network that mediates the interactions [1]. This sensitivity implies that, whenever it is not possible to have a complete empirical knowledge of the network, one should make an optimal use of the partial information available and try to reconstruct the most likely network, or rather an ensemble of likely networks, in the least biased way. Formally, this task can be regarded as a constrained entropy maximization problem, where the constraints represent the available information and the maximization of the entropy ensures that the reconstructed ensemble of networks is maximally random, given the enforced constraints [2, 3].

Among the possible types of incomplete topological information (e.g. missing links, missing nodes, etc.), one of the most frequently encountered situations is when only a *local* knowledge of the network is available [4–9]. For instance, in binary networks knowing the *number* of links (or 'degree') of each node is typically much easier than knowing the *identity* of all neighbours (the nodes at the other end of those links). Similarly, in weighted networks knowing the total intensity of links connected to each node (or 'strength') is much easier than knowing the identity of all neighbours and the intensity of all links separately. A typical example is that of interbank networks, where it is relatively easy to know the total exposures of each bank, while privacy issues make it much more difficult to know *who* is lending to whom, and *how much* [5, 6, 8, 9]. When the available information is just local, one only knows $O(N)$ quantities (e.g. the degrees of all nodes) instead of the total $O(N^2)$ ones (e.g. all entries of the adjacency matrix) fully describing the network. This makes the network reconstruction problem very challenging, since the number of missing variables is still $O(N^2)$, i.e. of the same order of the total number.

Even when the real network is entirely known, it is still necessary to reconstruct the most likely network from local properties in order to have a benchmark (i.e. a *null model*) to assess the statistical significance of any higher-order pattern, e.g. *assortativity* [10], *rich-club* effect [11], existence of *network motifs* [12, 13] and *communities* [14]. Null models correctly filter out the intrinsic and unavoidable heterogeneity of nodes, e.g. the fact that more popular people naturally have a larger degree in social networks. The simplest and most extensively used null model is the *Configuration Model* (CM), defined as an ensemble of random graphs with given *degree sequence* (the vector of degrees of all nodes) [2, 3]. It was recently shown that, despite its conceptual simplicity, the CM already poses significant problems of *bias*: it is very

difficult to implement the model in such a way that each network in the reconstructed ensemble is assigned the correct probability, thus leading to unbiased ensemble-averaged expectations [3, 15]. Once these solutions are appropriately implemented, many binary networks turn out to be reconstructed remarkably well from the knowledge of their degree sequence alone [3, 16–18].

In this paper we address the problem of the effective reconstruction, from local properties alone, of *weighted* networks. We first show that, in contrast with what is generally believed, the reconstruction of weighted networks does not merely involve a one-to-one mapping of the corresponding methodology that works well for binary networks. Specifically, inferring the structure of a weighted network only from the knowledge of its *strength sequence* (the vector of strengths of all nodes) can lead to a very bad reconstruction, even for the networks that, at a binary level, can be reproduced extremely well from their degree sequence [3, 16, 18]. We then conjecture that the reason is the fact that the knowledge of the strengths does not merely include or improve that of the degrees, since the binary information is completely lost once purely weighted quantities are measured. This leads us to the expectation that the reconstruction of weighted networks would be significantly enhanced by the specification of both strengths and degrees. We therefore introduce an analytical and unbiased maximum-entropy technique to reconstruct unbiased ensembles of weighted networks from the knowledge of both strengths and degrees.

In applying our enhanced method to several networks of different nature, we show that it leads to a significantly improved reconstruction, while remaining completely feasible since the required information is still local and the number of known variables is still $O(N)$. We finally introduce rigorous information-theoretic criteria confirming that the joint binary and weighted local information cannot be reduced to the weighted information alone. The resulting self-consistent picture is that the reconstruction of weighted networks is dramatically enhanced by the use of the irreducible set of joint degrees and strengths.

Our results also have strong implications for the identification of higher-order patterns in real networks. In particular, many of the observed properties that are unexplained by local weighted information turn out to be consistent with the enhanced, but still entirely local, information that includes both strenghts and degrees.

I. RESULTS

A. Naive reconstruction of weighted networks

Naively, the most natural generalization of the CM to weighted networks is a reconstructed ensemble with given *strength sequence*, and is sometimes referred to as the *Weighted Configuration Model* (WCM) [3, 20, 21]. The WCM is widely used both as a reconstruction method and as the most important null model to detect communities. In both cases, if s_i denotes the strength of node i and N is the number of nodes, the expected weight of the link between nodes i and j predicted by the WCM is routinely written in the form

$$\langle w_{ij}\rangle = \frac{s_i s_j}{\sum_{m=1}^{N} s_m} \tag{1}$$

or in a slightly different way if the network is directed (for simplicity, in this paper we will only consider undirected networks). For instance, the above expression represents one of the standard procedures to infer interbank linkages from the total exposures of individual banks [5], or the fundamental null model used by most algorithms aimed at detecting densely connected *communities* in weighted networks [14].

Unfortunately, despite its widespread use, eq.(1) is however incorrect, and differs from the unbiased expression derived within a rigorous maximum-entropy approach [3, 22, 23]. A simple signature of this inadequacy is the fact that, although eq.(1) is treated as an expected value, there is no indication of the probability distribution from which it is derived. Therefore, it is impossible to derive the expected value of topological properties which are nonlinear functions of the weights (i.e. the weighted clustering coefficient that we will introduce later). This problem has been solved only recently with the introduction of an analytical maximum-likelihood approach that leads to the correct expressions for the weight probability and any function of the expected weights [3].

But a more profound limitation of the WCM persists even when it is correctly implemented. It should be noted that the motivation for using the WCM as the natural generalization of the CM to weighted networks is the implicit assumption that the strength is an improved node-specific property, superior to the degree because it encapsulates the extra information provided by link weights. However, recent counter-intuitive results have shown that, while the *complete* knowledge of a weighted network conveys of course more information than the complete knowledge of just its binary projection, the strength sequence is often surprisingly less informative than the degree sequence [3, 16–18]. In particular, several *purely topological* properties of real weighted networks turn out to be reproduced much better by applying the CM to the binary projection of the graph, than by applying the WCM to the original weighted network [3, 16, 18]. The reason is that the strength sequence gives a very bad prediction of purely topological properties, and particularly the degrees: in fact, out of the many, possible ways to redistribute each node's strength among the remaining vertices irrespectively of the number of new links created, the WCM selects those predicting much denser networks than the real ones [18].

As a preliminary step of our analysis, we now confirm and extend these non-obvious findings to various networks of different nature. We consider the Italian Interbank network in year 1999 [24], three 'classic' social

networks collected in [25], seven food webs from [26], and finally the aggregated World Trade Web (WTW) in year 2002 [18]. The latter example, where nodes are world countries and links are their trade relationships (amount of imports and exports), is the system for which the role of strengths and degrees, when considered separately, has been studied in greatest detail [16–18].

From the above discussion, it is clear that in order to assess the performance of the network reconstruction method one should monitor not only the reconstructed properties that depend entirely on link weights, but also those that depend on the binary topology. For this reason, in fig.1 we compare, for all networks in the sample, the empirical and reconstructed values of various structural properties, including both purely topological properties and their weighted counterparts. If the full weighted matrix is denoted by $\mathbf{W}$ (where w_{ij} is the weight of the link between node i and node j), the purely topological quantities are calculated on the binary projection $\mathbf{A}$ (adjacency matrix) of $\mathbf{W}$, with entries $a_{ij} = 1$ if $w_{ij} > 0$ and $a_{ij} = 0$ if $w_{ij} = 0$ (compactly, we can write $a_{ij} \equiv w^0_{ij}$ with the convention $0^0 = 0$).

The binary quantities we choose are the simplest non-local ones, i.e. those involving paths going two and three steps away from a node. The *average nearest neighbor degree* (ANND), which is a measure of correlation between the degrees of adjacent nodes, is defined as

$$k^{nn}_i(\mathbf{W}) \equiv \frac{\sum_{j\neq i} a_{ij}k_j}{k_i} = \frac{\sum_{j\neq i}\sum_{k\neq j} w^0_{ij}w^0_{jk}}{\sum_{j\neq i} w^0_{ij}} \qquad (2)$$

(where $k_i = \sum_{j\neq i} a_{ij} = \sum_{j\neq i} w^0_{ij}$) and the *clustering coefficient*, which measures the fraction of triangles aroung node i, is defined as

$$c_i(\mathbf{W}) = \frac{\sum_{j\neq i}\sum_{k\neq i,j} w^0_{ij}w^0_{jk}w^0_{ki}}{\sum_{j\neq i}\sum_{k\neq i,j} w^0_{ij}w^0_{ki}} \qquad (3)$$

The corresponding weighted quantities are the *average nearest neighbor strength* (ANNS) [18] defined as

$$s^{nn}_i(\mathbf{W}) \equiv \frac{\sum_{j\neq i} a_{ij}s_j}{k_i} = \frac{\sum_{j\neq i}\sum_{k\neq j} w^0_{ij}w_{jk}}{\sum_{j\neq i} w^0_{ij}} \qquad (4)$$

(where $s_i = \sum_{j\neq i} w_{ij}$) and the *weighted clustering coefficient* [18, 19] defined as

$$c^w_i(\mathbf{W}) = \frac{\sum_{j\neq i}\sum_{k\neq i,j}(w_{ij}w_{jk}w_{ki})^{1/3}}{\sum_{j\neq i}\sum_{k\neq i,j} w^0_{ij}w^0_{ki}} \qquad (5)$$

In each panel of fig. 1, we show the measured value of one of the quantity defined above, for all nodes and for all networks, and we compare it with the corresponding reconstructed value predicted by the WCM[30]. In this type of plot, every point is a node. Therefore the target of a good reconstruction method is that of placing all the points along the identity. By constrast, in most cases we find that the reconstructed values for all nodes of a given network lie along horizontal lines, i.e. they are nearly

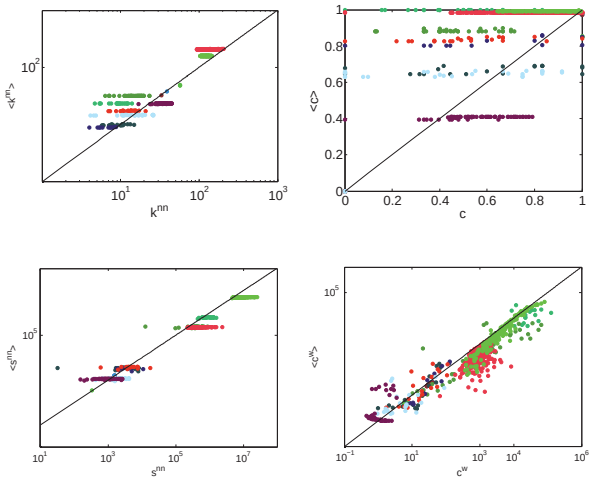

FIG. 1. Naive network reconstruction from node strengths (WCM), showing that purely weighted local properties are poorly informative. In each panel we compare the reconstructed (y axis) and real (x axis) value of a node-specific network property, for all nodes of the following 12 networks: Office social network (●), Research group social network (●), Fraternity social network (●), Maspalomas Lagoon food web (●), Chesapeake Bay food web (●), Crystal River (control) food web (●), Crystal River food web (●), Michigan Lake food web (●), Mondego Estuary food web (●), Everglades Marshes food web (●), Italian Interbank network in year 1999 (●), aggregated World Trade Web in year 2002 (●). Top left: average nearest neighbour degree (k^{nn}_i). Top right: binary clustering coefficient (c_i). Bottom left: average nearest neighbour strength (s^{nn}_i). Bottom right: weighted clustering coefficient (c^w_i).

equal to each other and totally unrelated to the 'target' real values. It could also be noted the better performance achieved by the WCM in reproducing the weighted clustering coefficient of the WTW than the other networks' one: however, as already pointed out in [18], this result is not robust to disaggregation (the sparser the commodity the worse the agreement), proving that the WCM is, generally speaking, a very bad reconstruction method.

At this point, the typical interpretation of a result like the above one is that the reconstruction of networks from local node-specific information is intrinsically problematic, presumably because of higher-order mechanisms involved in the formation of real networks, thus taking a difference between real data and the WCM as an important signature of non-local patterns [3, 20, 21]. Most community detection methods are indeed entirely based on this difference, and use it to define the so-called *modularity* guiding the detection algorithm [14]. However, as we show in the following, all the above results and the corresponding interpretations are completely reversed if we consider an enhanced reconstruction method.

B. The irreducibility conjecture

In what follows, we propose a different interpretation of the above findings. We conjecture (and rigorously prove later) that, in general, the poor reconstruction achieved by the WCM might be largely due to fact that the strength sequence discards purely topological information, and in particular the degrees. This hypothesis is consistent with previous results on the role of strengths and degrees in the WTW [16–18]. While, at a binary level, the assortativity and clustering properties of the WTW can be excellently reproduced by the CM [17], the corresponding weighted quantities turn out to be very different from the ones predited by the WCM on the basis of the strength sequence alone [18]. These results are very robust and hold true over time, on different datasets, and for various resolutions of the WTW (i.e. for different levels of aggregation of traded commodities) [16–18]. They perfectly illustrate that the naive expectation that weighted quantities are *per se* more informative than the corresponding binary ones is fundamentally incorrect. According to our conjecture, the degrees are to be considered a 'fundamental' local structural property of weighted networks, irreducible to the knowledge of the strengths and thus at least as important as the latter.

We should at this point clarify that by 'irreducible' we do not refer to the *numerical values* of strengths and degrees, but to the different *functional roles* that the two quantities play in determining or constraining the network's structure. In fact, strengths and degrees are typically highly correlated in real networks [10], which means that we can reasonably infer the values of one quantity from those of the other (in this sense, strengths and degrees are 'reducible' to each other). However, this is, generally speaking, only true from an *empirical* point of view. What is of interest to us is a deeper form of irreducibility, encountered when the joint specification of strengths and degrees *constrains the network in a fundamentally different way* than the specification of only one of the two. As an example, nothing guarantees that their observed correlation is preserved by the randomization procedure in the latter case (i.e. that $s_i \propto f(k_i)$ implies $\langle s_i \rangle \propto f(\langle k_i \rangle)$), as proved by the bad performance of the WCM in reproducing the degree sequence of the WTW [16, 18].

So, our conjecture leads us to the expectation that an enhanced reconstruction method (or null model) of weighted networks from purely local information should build on the simultaneous specification of strengths and degrees. Unfortunately, no satisfactory way to implement such method has been proposed so far. Moreover, no rigorous criterion has been defined to assess whether the introduction of the degree sequence as an additional constraint in the WCM is indeed non-redundant (i.e. not over-fitting the network).

In what follows, we fill both gaps by first defining a fast and unbiased approach to realize the enhanced network reconstruction method, and then introducing an information-theoretic criterion to check *a posteriori* whether the addition of degrees is non-redundant, confirming the irreducibility conjecture.

C. Unbiased ensembles with given strengths and degrees

For simplicity, we will refer to the ensemble of networks with given strenghts and degrees as the 'Mixed Configuration Model' (MCM). Early attempts to generate the MCM were either based on computational randomizations [27] or on theoretical arguments [21]. However, analytical calculations later showed that these approaches are statistically biased [23]. We now develop a maximum-entropy formalism, starting with the exact analytical results available for the MCM [23] (we only consider the case of undirected networks, although the generalization to the directed case is straightforward). Formally, an ensemble of weighted networks with N nodes can be characterized by a collection $\{\mathbf{W}\}$ of $N \times N$ matrices and by an appropriate probability $P(\mathbf{W})$ [23]. On each network $\mathbf{W}$, the strength is defined as $s_i(\mathbf{W}) \equiv \sum_{j \neq i} w_{ij}$ and the degree is defined as $k_i(\mathbf{W}) \equiv \sum_{j \neq i} w_{ij}^0$. We assume that each w_{ij} is a non-negative integer number (again, with the convention $0^0 = 0$).

We look for a probability that, besides being normalized ($\sum_{\mathbf{W}} P(\mathbf{W}) = 1$), ensures that the (expected) degree and strength of each node can be set equal to their known observed values, while leaving the ensemble maximally random otherwise (thus not biasing the probability). This is achieved by requiring that $P(\mathbf{W})$ maximizes Shannon's entropy $S \equiv -\sum_{\mathbf{W}} P(\mathbf{W}) \ln P(\mathbf{W})$ with a constraint on the expected degree and strength sequences $\langle \vec{k} \rangle$, $\langle \vec{s} \rangle$ [23]. The fundamental result [23] of this constrained maximization is the probability

$$P(\mathbf{W}|\vec{x},\vec{y}) = \prod_{i<j} q_{ij}(w_{ij}|\vec{x},\vec{y}) \tag{6}$$

where $\vec{x}$ and $\vec{y}$ are two N-dimensional Lagrange multipliers controlling for the expected degrees and strengths respectively (with $x_i \geq 0$ and $0 \leq y_i < 1$ $\forall i$), and

$$q_{ij}(w|\vec{x},\vec{y}) = \frac{(x_i x_j)^{\Theta(w)} (y_i y_j)^w (1 - y_i y_j)}{1 - y_i y_j + x_i x_j y_i y_j} \tag{7}$$

is the probability that a link of weight w exists between nodes i and j. In the above expression, $\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ otherwise. Note that $\sum_{w=0}^{+\infty} q_{ij}(w|\vec{x},\vec{y}) = 1$ $\forall i, j$.

Equation (7) defines the 'mixed' Bose-Fermi distribution [23] where, due to the presence of $\Theta(w)$, the establishment of a link of unit weight between two nodes requires a different (higher if $x_i x_j > 0$) 'cost' than the reinforcement (by a unit of weight) of an already existing link. This feature is due to the mixed binary and weighted constraints and makes the MCM potentially very appropriate to model real networks.

To achieve this, we now apply the maximum-likelihood approach [3, 28] to the model. We consider a particular real weighted network $\mathbf{W}^*$, whose only degrees $k_i^* \equiv k_i(\mathbf{W}^*)$ and strengths $s_i^* \equiv s_i(\mathbf{W}^*)$ are known. The log-likelihood of the MCM defined by eqs.(6) and (7) reads

$$\mathcal{L}(\vec{x},\vec{y}) \equiv \ln P(\mathbf{W}^*|\vec{x},\vec{y}) = \sum_{i<j} \ln q_{ij}(w_{ij}^*|\vec{x},\vec{y}) = \sum_{i=1}^{N} (k_i^* \ln x_i + s_i^* \ln y_i) + \sum_{i<j} \ln\left(\frac{1-y_iy_j}{1-y_iy_j+x_ix_jy_iy_j}\right) \quad (8)$$

We now look for the specific parameter values $\vec{x}^*, \vec{y}^*$ that maximize $\mathcal{L}(\vec{x},\vec{y})$. A direct calculation, analogous to the simpler ones encountered in other null models [3, 28], shows that $\vec{x}^*, \vec{y}^*$ can be obtained as the real solution to the following system of $2N$ equations:

$$\langle k_i \rangle = \sum_{j\neq i} \frac{x_ix_jy_iy_j}{1-y_iy_j+x_ix_jy_iy_j} = k_i^* \quad \forall i \quad (9)$$

$$\langle s_i \rangle = \sum_{j\neq i} \frac{x_ix_jy_iy_j}{(1-y_iy_j)(1-y_iy_j+x_ix_jy_iy_j)} = s_i^* \quad \forall i \quad (10)$$

Therefore, we find that the likelihood-maximizing values $\vec{x}^*, \vec{y}^*$ are precisely those ensuring that the expected degree and strength sequences coincide with the observed sequences $\vec{k}^*$ and $\vec{s}^*$, thus solving our initial problem.

As we show below, the values $\vec{x}^*, \vec{y}^*$ contain all the information necessary to reconstruct the network. Thus the maximum-likelihood approach translates the time-consuming and bias-prone problem of the computational generation of several reconstructed networks into the much simpler problem of solving the $2N$ equations (9-10), or equivalently maximizing the function $\mathcal{L}(\vec{x},\vec{y})$ of $2N$ variables[31]. Consistently with our problem, in either case finding $\vec{x}^*$ and $\vec{y}^*$ only requires the knowledge of the observed strengths and degrees, and not that of the entire network $\mathbf{W}^*$.

D. Reconstructed properties

Once the solutions $\vec{x}^*$ and $\vec{y}^*$ are found, they can be used to obtain the reconstructed (ensemble-averaged) network properties analytically, with no need to actually measure such properties on any sampled network. Specifically, given a topological property $X(\mathbf{W})$ whose 'true' (but in general unknown) value is $X^* \equiv X(\mathbf{W}^*)$, the reconstructed value can be calculated analytically as $\langle X \rangle \equiv \sum_{\mathbf{W}} X(\mathbf{W}) P(\mathbf{W}|\vec{x}^*,\vec{y}^*)$. For most topological properties of interest, this involves calculating the expected product of (powers of) distinct matrix entries, which simply reads

$$\left\langle \sum_{i\neq j\neq k,\ldots} w_{ij}^{\alpha} \cdot w_{jk}^{\beta} \cdot \ldots \right\rangle = \sum_{i\neq j\neq k,\ldots} \langle w_{ij}^{\alpha} \rangle \cdot \langle w_{jk}^{\beta} \rangle \cdot \langle \ldots \rangle \quad (11)$$

with the generic term given by

$$\langle w_{ij}^{\gamma} \rangle = \sum_{w=0}^{+\infty} w^{\gamma} q_{ij}(w|\vec{x}^*,\vec{y}^*) = \frac{x_i^*x_j^*(1-y_i^*y_j^*)\mathrm{Li}_{-\gamma}(y_i^*y_j^*)}{1-y_i^*y_j^*+x_i^*x_j^*y_i^*y_j^*} \quad (12)$$

where $\mathrm{Li}_n(z) \equiv \sum_{l=1}^{+\infty} z^l/l^n$ is the nth polylogarithm of z. The simplest and most useful cases $\gamma = 1$ and $\gamma = 0$ yield the expected weight $\langle w_{ij} \rangle$ and the connection probability $p_{ij} = \langle \Theta(w_{ij}) \rangle = \langle w_{ij}^0 \rangle$, respectively. Therefore the reconstructed value $\langle X \rangle$ can be calculated in the same time as that required to calculate the real (if known) value $X(\mathbf{W}^*)$ (i.e. the shortest possible time), by simply replacing w_{ij}^{γ} with $\langle w_{ij}^{\gamma} \rangle$ in the definition of $X(\mathbf{W})$.

E. Enhanced reconstruction of real weighted networks

We can now apply our general methodology to the reconstruction of real-world networks. We consider again the assortativity and clustering properties defined in eqs.(2)-(5). The result is illustrated in fig. 2 for all the networks shown previously in fig. 1. We clearly see that our enhanced method achieves a dramatic improvement over the standard approach. Now most points lie in the vicinity of the identity, meaning that our method is able to successfully reconstruct, for each vertex, the structure of the network two and three steps away from it. Note that the noisiest property is the binary clustering coefficient; however if we compare our results with the naive ones we find that the improvement achieved for this quantity is perhaps the most significant one.

The above findings completely reverse the conclusions one would draw from the previous interpretation of the naive results. First, network reconstruction from purely local properties is now shown to be possible to a highly satisfactory level, at least for the networks considered here. Second, the assortativity and clustering properties of these networks turn out to be well explained by purely local, even if augmented, properties. So, there is no need to invoke non-local mechanisms in order to explain such properties in these networks. We similarly expect that, if one considers the MCM as an improved null model to detect communities or other higher-order patterns, the result will be dramatically different from what is obtained by using the WCM prediction in the definition of the modularity [14].

F. Information-theoretic tests of irreducibility

We now confirm the superiority of our method using a rigorous goodness-of-fit approach that compares the performance of the WCM and MCM in reproducing the whole network. At the same time, this approach will automatically allow us to test our initial conjecture that the degrees are irreducible to the strengths. Indeed, both

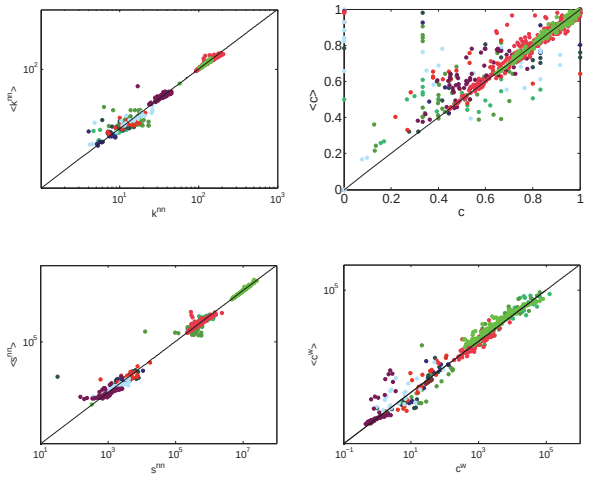

FIG. 2. Enhanced network reconstruction from strengths and degrees (MCM), showing dramatic improvements over the standard approach. In each panel we compare the reconstructed (y axis) and real (x axis) value of a node-specific network property, for all nodes of the following 12 networks: Office social network (●), Research group social network (●), Fraternity social network (●), Maspalomas Lagoon food web (●), Chesapeake Bay food web (●), Crystal River (control) food web (●), Crystal River food web (●), Michigan Lake food web (●), Mondego Estuary food web (●), Everglades Marshes food web (●), Italian Interbank network in year 1999 (●), aggregated World Trade Web in year 2002 (●). Top left: average nearest neighbour degree (k_i^{nn}). Top right: binary clustering coefficient (c_i). Bottom left: average nearest neighbour strength (s_i^{nn}). Bottom right: weighted clustering coefficient (c_i^w).

problems can be equivalently stated within a model selection framework, where one is interested in determining not only which of the two models achieves the best fit to the data, but also whether the introduction of the degrees as extra parameters in the MCM is really non-redundant.

To start with, we need to compare the likelihood of the ordinary WCM with that of MCM. Note that the WCM can be obtained as a particular case of the MCM by setting $\vec{x} = \vec{1}$. The log-likelihood of the WCM is therefore the reduced function $\mathcal{L}(\vec{1}, \vec{y})$ of N variables, and is maximized by a new vector $\vec{y}^{**} \neq \vec{y}^{*}$ which is also the solution of eq.(10) with $\vec{x} = \vec{1}$. The predictions of the WCM are still obtained as in eqs.(11) and (12), by replacing x_i^* with 1 and y_i^* with y_i^{**} in the latter. This is how the reconstructed properties in Fig.1 were computed.

Now, if we simply compare the maximized likelihoods of the two reconstruction methods, we trivially obtain $\mathcal{L}(\vec{x}^*, \vec{y}^*) \geq \mathcal{L}(\vec{1}, \vec{y}^{**})$ since the MCM includes the WCM as a particular case. However, information-theoretic criteria exist [29] to assess whether the increased accuracy of a model with more parameters is a result of over-fitting, in which case a more parsimonious model should be preferred. The most popular choices are the Likelihood-ratio test (LRT) and Akaike's Information Criterion (AIC), showing that the optimal trade-off between accuracy and parsimony is achieved by discounting the number of free parameters from the maximized likelihood [29]. For our two competing null models,

$$\mathrm{AIC}_{MCM} \equiv -2\mathcal{L}(\vec{x}^*, \vec{y}^*) + 4N \tag{13}$$

$$\mathrm{AIC}_{WCM} \equiv -2\mathcal{L}(\vec{1}, \vec{y}^{**}) + 2N \tag{14}$$

and the optimal model is the one minimizing AIC. However, if the AIC difference is small, the two models will still be comparable. A correct quantitative criterion is given by the so-called AIC Weights [29], which in our case read

$$w_{MCM}^{AIC} \equiv \frac{e^{-AIC_{MCM}/2}}{e^{-AIC_{MCM}/2} + e^{-AIC_{WCM}/2}} \tag{15}$$

$$w_{WCM}^{AIC} \equiv 1 - w_{MCM}^{AIC} \tag{16}$$

and quantify the weight of evidence in favour of a model, i.e. the probability that the model is the best one among the two.

TABLE I. AIC weights and BIC weights for the considered null models.

Network	w_{WCM}^{AIC}	w_{MCM}^{AIC}
● Office social network [25]	1	0
● Research group social network[25]	1	0
● Fraternity social network [25]	0	1
● Maspalomas Lagoon food web [26]	0	1
● Chesapeake Bay food web [26]	0	1
● Crystal River (control) food web [26]	0	1
● Crystal River food web [26]	0	1
● Michigan Lake food web [26]	0	1
● Mondego Estuary food web [26]	0	1
● Everglades Marshes food web [26]	0	1
● Italian interbank network (1999) [24]	0	1
● World Trade Web (2000)[18]	0	1

The AIC weights of the two reconstruction methods are shown in table I for all networks [32]. Moreover, the LRT response is the same of AIC, at both 5% and 1% significance levels. We see that, apart from two social networks, the enhanced method is always superior to the naive one, and achieves unit probabilty (within machine precision) of being the best among the two models. A closer inspection of the two networks for which the opposite result holds reveals that they are (almost) fully connected. This explains why the specification of the degree sequence, which in this case is close to the almost fully connected prediction of the WCM, is redundant for these networks. In such cases, the relevant local constraints effectively reduce to the strength sequence, so the 'standard' WCM is preferable. Our method correctly indentifies this situation. However, whenever the topology is nontrivial (as in most real-world networks), the local constraints are irreducible to the strength sequence alone and the degrees must be separately specified. We should therefore expect that, for the vast majority of real-world networks, the degree sequence is irreducible to the

strength sequence. In such cases, the inclusion of degrees is non-redundant, explaining why our method retrieves significantly more information.

II. CONCLUSIONS

Motivated by recent findings showing that the local binary properties of weighted networks can be surprisingly more informative than their weighted counterparts, in this work we have introduced an improved, fast and unbiased method to reconstruct weighted networks from the joint set of strengths and degrees. We compared our enhanced method (MCM) with the simpler one that naively uses only the strength sequence to reconstruct the network (WCM). We confirmed an extremely bad agreement between real network properties and their WCM-reconstructed counterparts, implying that the strength sequence is in general uninformative about the higher-order properties of the network. The typical interpretation of this result is the conclusion that the network is shaped by non-local mechanisms, irreducible to local formation rules. By contrast, we showed that the MCM provides accurate reconstructed properties, indicating that the combination of strengths and degrees is extremely informative. In other words, the real networks in our analysis turned out to be typical members of the MCM ensemble and not of the WCM ensemble. This has important consequences for critical problems like the reconstruction of interbank linkages from bank-specific information: the analysis of the interbank network shows that the standard approach is systematically uninformative.

Moreover, information-theoretic criteria confirmed that the inclusion of the degrees as additional constraints is non-redundant. An important consequence is that our MCM should be regarded as a more appropriate, and still sufficiently parsimonious, null model of weighted networks with local constraints. The agreement of this stricter null model with the networks in our sample implies that the higher-order properties considered here are well explained by local constraints, thus completely inverting the conclusions following from the use of the naive approach.

ACKNOWLEDGMENTS

D.G. acknowledges support from the Dutch Econophysics Foundation (Stichting Econophysics, Leiden, the Netherlands) with funds from beneficiaries of Duyfken Trading Knowledge BV, Amsterdam, the Netherlands.
G.F. gratefully acknowledges financial support received by the research project "The international trade network: empirical analyses and theoretical models" funded by the Italian Ministry of Education, University and Research (Scientific Research Programs of National Relevance 2009).

[1] Barrat, A., Barthlemy, M., & Vespignani, A. (2008). Dynamical processes on complex networks. Cambridge University Press.
[2] Park, J., & Newman, M. E. (2004). Statistical mechanics of networks. Physical Review E, 70(6), 066117.
[3] Squartini, T., & Garlaschelli, D. (2011). Analytical maximum-likelihood method to detect patterns in real networks. New Journal of Physics, 13(8), 083001.
[4] Garlaschelli, D., & Loffredo, M. I. (2004). Fitness-dependent topological properties of the world trade web. Physical review letters, 93(18), 188701.
[5] Wells, S. (2004). Financial interlinkages in the United Kingdom's interbank market and the risk of contagion, Bank of England Working Paper, No. 230/2004.
[6] Bargigli, L., & Gallegati, M. (2011). Random digraphs with given expected degree sequences: A model for economic networks. Journal of Economic Behavior & Organization, 78(3), 396-411.
[7] Musmeci, N., Battiston, S., Caldarelli, G., Puliga, M., & Gabrielli, A. (2012). Bootstrapping topology and systemic risk of complex network using the fitness model. arXiv preprint arXiv:1209.6459.
[8] Caldarelli, G., Chessa, A., Pammolli, F., Gabrielli, A., & Puliga, M. (2013). Reconstructing a credit network. Nature Physics, 9(3), 125-126.
[9] Mastromatteo, I., Zarinelli, E., & Marsili, M. (2012). Reconstruction of financial networks for robust estimation of systemic risk. Journal of Statistical Mechanics: Theory and Experiment, 2012(03), P03011.
[10] Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101(11), 3747-3752.
[11] Zlatic, V., Bianconi, G., Díaz-Guilera, A., Garlaschelli, D., Rao, F., & Caldarelli, G. (2009). On the rich-club effect in dense and weighted networks. The European Physical Journal B, 67(3), 271-275.
[12] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science Signaling, 298(5594), 824.
[13] Squartini, T., & Garlaschelli, D. (2012). Triadic motifs and dyadic self-organization in the World Trade Network. In Self-Organizing Systems (pp. 24-35). Springer Berlin Heidelberg.
[14] Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3), 75-174.
[15] Roberts, E. S., & Coolen, A. C. C. (2012). Unbiased degree-preserving randomization of directed binary networks. Physical Review E, 85(4), 046103.
[16] Fagiolo, G., Squartini, T., & Garlaschelli, D. (2011). Null models of economic networks: the case of the world trade web. Journal of Economic Interaction and Coordination, 1-33.
[17] Squartini, T., Fagiolo, G., & Garlaschelli, D. (2011). Ran-

domizing world trade. I. A binary network analysis. Physical Review E, 84(4), 046117.
[18] Squartini, T., Fagiolo, G., & Garlaschelli, D. (2011). Randomizing world trade. II. A weighted network analysis. Physical Review E, 84(4), 046118.
[19] Fagiolo, G. (2007). Clustering in complex directed networks. Physical Review E, 76(2), 026107.
[20] Serrano, M. Á., & Boguñá, M. (2005, June). Weighted Configuration Model. In AIP Conference Proceedings (Vol. 776, p. 101).
[21] Serrano, M. Á., Boguñá, M., & Pastor-Satorras, R. (2006). Correlations in weighted networks. Physical Review E, 74(5), 055101.
[22] Bianconi, G. (2009). Entropy of network ensembles. Physical Review E, 79(3), 036114.
[23] Garlaschelli, D., & Loffredo, M. I. (2009). Generalized bose-fermi statistics and structural correlations in weighted networks. Physical review letters, 102(3), 038701.
[24] De Masi, G., Iori, G., & Caldarelli, G. (2006). Fitness model for the Italian interbank money market. Physical Review E, 74(6), 066112.
[25] Killworth, P. D., & Bernard, H. R. (1976). Informant accuracy in social network data. Human Organization, 35(3), 269-286.
[26] http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm
[27] Bhattacharya, K., Mukherjee, G., Saramäki, J., Kaski, K., & Manna, S. S. (2008). The international trade network: weighted network analysis and modelling. Journal of Statistical Mechanics: Theory and Experiment, 2008(02), P02002.
[28] Garlaschelli, D., & Loffredo, M. I. (2008). Maximum likelihood: extracting unbiased information from complex networks. Physical Review E, 78(1), 015101.
[29] Burnham, K. P., & Anderson, D. R. (2002). Model selection and multi-model inference: a practical information-theoretic approach. Springer Verlag.
[30] The methodology used is described in refs. [3, 18] and briefly summarized later.
[31] To find $\vec{x}^*$ and $\vec{y}^*$, we chose to solve eqs.(9-10) using MatLab. The code is available on request.
[32] We also used the Bayesian Information Criterion (BIC) [29], that puts a higher penalty on the number of parameters. We found that BIC weights are identical to the AIC ones (within machine precision) for all networks in our samples.