J Stat Phys
https://doi.org/10.1007/s10955-018-2076-z

Tackling Information Asymmetry in Networks: A New Entropy-Based Ranking Index

Paolo Barucca[1,2] · **Guido Caldarelli**[2,3] · **Tiziano Squartini**[3]

Received: 17 October 2017 / Accepted: 1 June 2018

Abstract Information is a valuable asset in socio-economic systems, a significant part of which is entailed into the network of connections between agents. The different interlinkages patterns that agents establish may, in fact, lead to asymmetries in the knowledge of the network structure; since this entails a different ability of quantifying relevant, systemic properties (e.g. the risk of contagion in a network of liabilities), agents capable of providing a better estimation of (otherwise) inaccessible network properties, ultimately have a competitive advantage. In this paper, we address the issue of quantifying the information asymmetry of nodes: to this aim, we define a novel index—InfoRank—intended to rank nodes according to their information content. In order to do so, each node ego-network is enforced as a constraint of an entropy-maximization problem and the subsequent uncertainty reduction is used to quantify the node-specific accessible information. We, then, test the performance of our ranking procedure in terms of reconstruction accuracy and show that it outperforms other centrality measures in identifying the "most informative" nodes. Finally, we discuss the socio-economic implications of network information asymmetry.

Keywords Complex networks · Shannon entropy · Information theory · Ranking algorithm

1 Introduction

Recognizing the most relevant nodes in a networked system represents a topic of growing interest. This translates into identifying nodes with key features, be they structural or functional. Depending on the system under study, in fact, possessing certain features may translate into accessing a competitive advantage or prominent position in the system. The problem

✉ Tiziano Squartini
tiziano.squartini@imtlucca.it

[1] Department of Banking and Finance, University of Zürich, Zurich, ZH, Switzerland

[2] London Institute for Mathematical Sciences, 35a South St, Mayfair, London W1K 2XF, UK

[3] IMT School for Advanced Studies, P.zza S.Francesco 19, 55100 Lucca, Italy

Published online: 11 June 2018

has been tackled by defining a plethora of indices, aiming at quantifying the importance of a node in a given system: the so-called centrality measures [1–4].

The latter are intended to capture the role played by each node within the network by optimizing an opportunely-defined objective function: examples are provided by the degree-centrality (defined by the number of neighbors of each vertex) [2], the closeness-centrality (defined by the average distance of the reachable nodes from any, given, node) [5], the PageRank-centrality (defined by the number of "authoritative" nodes pointing at the vertex under consideration) [6], etc.

Differently from all the centrality measures above, which are partial by definition, our method focuses on the information content of the (different) interlinkages patterns of each node. We refer to this difference as to the *network information asymmetry* and we will show how it allows nodes to obtain a significantly-better estimation of the (otherwise unaccessible) network properties.

Our novel index, in other words, measures the reduction of uncertainty that the knowledge of the ego-network of each node allows to be gained: the node whose accessible information provides the largest uncertainty reduction will be identified as being the "most informative" one. More quantitatively, such a reduction is computed by comparing the Shannon entropy benchmark value—measurable by all nodes—with the one obtained by conditioning on the ego-network information on top of it.

Several attempts to define entropy-based indices have been made [8–11]; however, the measures that have been proposed so far are based on specific definitions of Shannon entropy, an evidence that severely affects their applicability. As we will show in what follows, InfoRank can be understood as a generalization of these measures, applicable to any maximum-entropy ensemble and to any subset of nodes.

The paper is organized as follows. In section "Methods", we show how InfoRank can be computed in the simplest case of the configuration model (the theoretical details of the derivation of our method can be found in Appendix). In section "Results", we measure InfoRank on a number of real-world networks and verify its correlation with the reconstruction accuracy achieved by each node. Finally, in section "Discussion" we comment on the role of network information asymmetry in social, economic and financial systems.

2 Methods

Quantifying the benchmark information. In order to measure the information gain coming from enforcing each node-specific ego-network, we first need to quantify the common, benchmark information accessible by all nodes in the network $\mathbf{A}$. To this aim, let us suppose it to be represented by the degree sequence, which amounts at considering the usual configuration model (CM) [12] as our benchmark model. Being the CM defined by the following system of equations [13]

$$k_i(\mathbf{A}) = \sum_{j(\neq i)} \frac{x_i x_j}{1 + x_i x_j} \equiv \sum_{j(\neq i)} p_{ij}, \ \forall\, i \tag{1}$$

the amount of information encoded into the degree sequence can be quantified by calculating the value of the Shannon entropy defined by the chosen constraints, i.e.

$$S_0 = \frac{1}{2} \sum_i S_0^{(i)}$$

$$= -\frac{1}{2}\sum_{i}\sum_{j(\neq i)}\left[p_{ij}\ln p_{ij} + (1-p_{ij})\ln(1-p_{ij})\right] \tag{2}$$

with $S_0^{(i)}$ indicating the contribution of node i to the benchmark entropy S_0 (the subscript 0 stresses the benchmark-like value of this functional, encoding a kind of information which is accessible to all nodes). Intuitively, the closer the S_0 value to zero, the larger the explanatory power of the degree sequence.

Quantifying the node-specific information. The second step of our procedure prescribes to constrain the ego-network of each specific node. For the sake of illustration, let us focus on node i: the aforementioned prescription amounts at posing

$$p_{ij} = \frac{x_i x_j}{1 + x_i x_j} = a_{ij},\ \forall\, j \tag{3}$$

i.e. treating as deterministic the links constituting the ego-network of i. As an example, let us suppose that node i is linked only with nodes 2 and 3 out of the N constituting our ideal network, i.e. that $\frac{x_i x_2}{1+x_i x_2} = \frac{x_i x_3}{1+x_i x_3} = 1$ and $\frac{x_i x_1}{1+x_i x_1} = \frac{x_i x_4}{1+x_i x_4} = \cdots = \frac{x_i x_N}{1+x_i x_N} = 0$. This implies that the system of equations that node i has to solve becomes

$$\begin{aligned}
k_1(\mathbf{A}) &= \frac{x_1 x_2}{1 + x_1 x_2} + \cdots\ 0\ \cdots + \frac{x_1 x_N}{1 + x_1 x_N}\\
k_2(\mathbf{A}) &= \frac{x_2 x_1}{1 + x_2 x_1} + \cdots\ 1\ \cdots + \frac{x_2 x_N}{1 + x_2 x_N}\\
k_3(\mathbf{A}) &= \frac{x_3 x_1}{1 + x_3 x_1} + \cdots\ 1\ \cdots + \frac{x_3 x_N}{1 + x_3 x_N}\\
k_4(\mathbf{A}) &= \frac{x_4 x_1}{1 + x_4 x_1} + \cdots\ 0\ \cdots + \frac{x_4 x_N}{1 + x_4 x_N}\\
&\ \ \vdots
\end{aligned} \tag{4}$$

(notice that we have omitted the equation controlling for the value of the i-th degree, since trivially satisfied). The system above can be rearranged by moving at the left hand side the known entries of the adjacency matrix. More explicitly:

$$\begin{aligned}
k_1(\mathbf{A}) &= \sum_{j(\neq 1,i)} \frac{x_1 x_j}{1 + x_1 x_j} \equiv \sum_{j(\neq 1,i)} q_{1j}^{(i)}\\
k_2(\mathbf{A}) - 1 &= \sum_{j(\neq 2,i)} \frac{x_2 x_j}{1 + x_2 x_j} \equiv \sum_{j(\neq 2,i)} q_{2j}^{(i)}\\
k_3(\mathbf{A}) - 1 &= \sum_{j(\neq 3,i)} \frac{x_3 x_j}{1 + x_3 x_j} \equiv \sum_{j(\neq 3,i)} q_{3j}^{(i)}\\
k_4(\mathbf{A}) &= \sum_{j(\neq 4,i)} \frac{x_4 x_j}{1 + x_4 x_j} \equiv \sum_{j(\neq 4,i)} q_{4j}^{(i)}\\
&\ \ \vdots
\end{aligned} \tag{5}$$

where the superscript (i) stresses that the numerical value of the probability coefficients $\{q_{jk}^{(i)}\}$ is induced by specifying the pattern of connections of node i (and, in general, $q_{jk}^{(i)} \neq p_{jk}$).

The problem of quantifying the informativeness of the ego-network of each node can, in fact, be restated by imagining that the node itself is *removed* from the network. In this way, a reduced adjacency matrix remains naturally defined, inducing, in turn, a reduced system of equations.

Calculating the node-specific InfoRank. Let us now come to the third step of our procedure, i.e. the computation of the InfoRank index. Once i has been removed from the network, the entropy of the "surviving" topological structure can be measured by employing the novel probability coefficients defined by the system of equations in (5), i.e.

$$S_{(i)} = -\frac{1}{2}\sum_{j}\sum_{k(\neq j)}\left[q_{jk}^{(i)}\ln q_{jk}^{(i)} + \left(1-q_{jk}^{(i)}\right)\ln\left(1-q_{jk}^{(i)}\right)\right]. \tag{6}$$

Since removing different nodes will, in general, impact on the benchmark entropy S_0 differently, a ranking is naturally induced by the amount of "uncertainty reduction" caused by the removal of each node. Since our aim is identifying the node(s) possessing the largest amount of information, in order to define a novel ranking index, let us divide $S_{(i)}$ by S_0 and take the complement to 1:

$$I_i = 1 - \frac{S_{(i)}}{S_0}; \tag{7}$$

as apparent from the definition, the larger the entropy reduction, the higher the rank of the node causing it. In what follows, the index I_i will be referred to as to the InfoRank index.

Approximating the node-specific InfoRank Although formally similar, the quantities $S_0^{(i)}$ and $S_{(i)}$, respectively defined in Eqs. (2) and (6), are conceptually very different and must not be confused. In fact, while $S_0^{(i)}$ just represents the contribution of node i to the benchmark entropy S_0, the second index $S_{(i)}$ quantifies the (residual) uncertainty about the network structure after the removal of node i. Whenever the effect of this removal on the remaining vertices can be ignored (i.e. when diminishing the nodes degree by one unit doesn't affect much the magnitude of the surviving probability coefficients), $S_{(i)}$ can be indeed approximated by $S_0 - S_0^{(i)}$, further implying that

$$I_i \simeq \frac{S_0^{(i)}}{S_0}. \tag{8}$$

3 Results

3.1 Ranking Nodes in Synthetic Networks

In order to better illustrate the meaning of the InfoRank index, let us consider two extreme cases, i.e. the removal of either an isolated or a fully-connected node. It is intuitive that, in both cases, the knowledge of the connections of the considered nodes adds no information or, equivalently, that removing these nodes doesn't lead to any uncertainty reduction. This is readily seen by comparing the systems (1) and (5). In presence of a hub, in fact, the system of equations $k_i(\mathbf{A}) = \sum_{j(\neq i)}\frac{x_i x_j}{1+x_i x_j}$, $\forall\, i$ can be rewritten as $\tilde{k}_i(\mathbf{A}) + 1 = \sum_{j(\neq i,h)}\frac{x_i x_j}{1+x_i x_j} + 1$, $\forall\, i(\neq h)$ (with h denoting the hub). Since solving the latter system with respect to $\{x_i\}_{i\neq h}$ is equivalent to solve the former system, removing a hub doesn't change the information content of the network configuration; analogously, when considering an isolated node. On the other hand, the value $I_i = 1$ characterizes a node whose removal induces a configuration which is perfectly deterministic (i.e. composed by isolated nodes, cliques or both). Naturally, in the very special case of a star graph, the central node is both the hub and the vertex with largest InfoRank value.

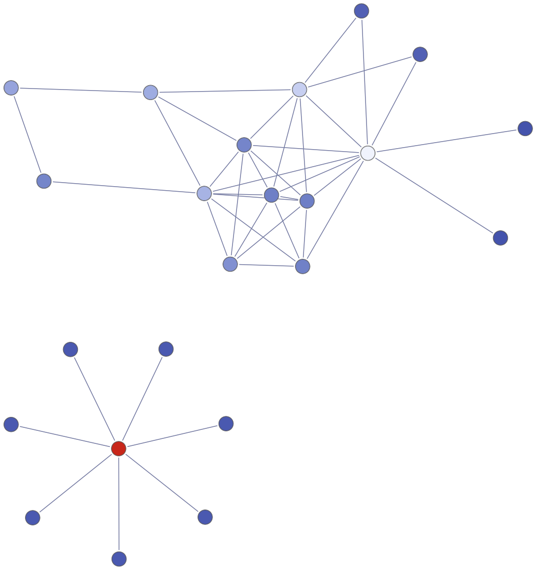

Fig. 1 Toy network, whose nodes have been ranked according to InfoRank (red nodes are ranked higher than blue nodes). Since the node with the largest score is the one maximally reducing the residual uncertainty of the network, InfoRank is not completely determined by the nodes degrees: the center of the star, in fact, has exactly the same number of neighbors of other nodes (i.e. 7); differently from them, however, its removal would cause and entire portion of the network to become deterministic (color figure online)

A relationship between InfoRank and the node degree, nonetheless, exists. In order to understand it, let us, first, consider the approximate definition provided in Eq. (8). The node with largest (approximate) InfoRank is also the one maximizing $S_0^{(i)}$, i.e. the one bringing the largest contribution to the benchmark entropy S_0. This is achieved by letting each of the addenda in Eq. (2) contribute with an average coefficient $\overline{p}_{ij} = \frac{k_i}{N-1} \simeq \frac{1}{2}$, further implying that a ranking based on the naïve contributions to S_0 would privilege nodes with $k_i \simeq (N-1)/2$ neighbors.

Let us consider the synthetic network shown in Fig. 1: upon computing the vector $\{S_0^{(i)}\}$, one finds that the largest contribution to S_0 comes from the hub, consistently with the discussion above (notice, in fact, that $k_h = 10 \simeq (N-1)/2 = 11$). InfoRank, instead, also accounts for the effect that constraining the pattern of connections of a given node has on the connections of the neighboring ones. Let us compare the consequences of removing the hub and the center of the star, from the network in Fig. 1: while removing the latter would cause an entire portion of the network to become deterministic (7 nodes would become isolated), removing the former, on the contrary, would just disconnect two more nodes. InfoRank correctly assigns the highest score to the center of the star, pointing it out as the vertex establishing the most informative set of interconnections: our index, in other words, encodes higher-order corrections to the naïve contribution $S_0^{(i)}$, by including the "effects" of the additional constraints on the neighboring vertices.

3.2 Ranking Nodes in Real-World Networks

Let us now employ InfoRank to analyse real-world configurations. The core of our analysis will consist in a thorough comparison of a number of alternative ranking indices (in what follows, binary, *directed* networks will be considered, since one of the chosen indicators becomes trivial in the undirected case): in order to consistently compare the ranking scores output by the selected algorithms, the former ones are normalized in order to let them range within the same interval. More specifically, if we let $R_i^{(a)}$ represent the rank of node i according to the chosen algorithm a, the applied transformation reads $\overline{R}_i^{(a)} = (R_i^{(a)} - \min\{R_i^{(a)}\})/(\max\{R_i^{(a)}\} - \min\{R_i^{(a)}\}) \in [0, 1]$ and ensures that nodes with minimum rank are assigned a value $\overline{R}_i^{(a)} = 0$ (in blue, according to the color scale adopted throughout the paper); viceversa, nodes with maximum rank are assigned a value $\overline{R}_i^{(a)} = 1$ (in red, according to the color scale adopted throughout the paper).

The first alternative index is represented by the *degree-centrality*, identifying the nodes characterized by the largest degree as the most important (i.e. central) ones. A first limitation of such an index lies in the nature of the connectivity concept, which lacks an obvious generalization to, e.g. the directed case we are considering in the present paper. In what follows we will adopt the following definition

$$D_i = k_i^{out} \tag{9}$$

which ranks the nodes according to the number of their out-neighbors. As evident from the first panel of Figs. 2, 3 and 4, the (out-)degree-centrality trivially identifies the hubs as the most central nodes.

The second indicator we have considered is the so-called *closeness-centrality* [5], defined as

$$C_i = \frac{1}{\overline{d}_i} = \frac{\kappa_i}{\sum_j d_{ij}} \tag{10}$$

i.e. as the reciprocal of the average topological distance of a vertex from the other, connected ones (κ_i is the number of nodes that can be reached from i—following the links direction—and d_{ij} is the topological distance separating i from any reachable node j). Intuitively, any two nodes are said to be "close" when their distance is "small", i.e. few links must be walked to reach one from the other. Naturally, the nodes with $C_i = 0$ are the ones with zero out-degree, while a node with exactly $N - 1$ connections will be also the most central one.

Generally speaking, however, nodes with small degree do not necessarily have a small closeness-centrality value: an example is provided by the second panel of Fig. 3, where nodes behaving like "local hubs" (e.g. at the center of star-like subgraphs) are, in fact, characterized by a large C_i independently from their degree. On the other hand, nodes with a large degree do not necessarily have a large closeness-centrality value: in fact, the first panel of Fig. 4 shows that although a large number of nodes can be reached from the hub, many lie at a large distance from it. Interestingly, as the second panel of Figs. 2 and 4 shows, the nodes minimizing the (average) topological distances are the ones connected (but not necessarily belonging) to the strongly connected component (SCC) that is present in these systems: in the case of the *C. Elegans* neural network, then, its large reciprocity ($\simeq 0.43$) further levels out the differences between the C_i values of the SCC vertices.

The third indicator considered in the present analysis is the *PageRank-centrality* [6]. It is computed by solving to iterative equation

$$P_i = \frac{1-\alpha}{N} + \alpha \sum_j \left(\frac{a_{ji}}{k_j^{out}} \right) P_j \tag{11}$$

(where we have set $\alpha = 0.85$) which can be imagined to describe a Markov chain: if $a_{ji} = 1$ a walker moves from j to i with probability $\frac{1-\alpha}{N} + \frac{\alpha}{k_j^{out}}$; if $a_{ji} = 0$ such a probability becomes $\frac{1-\alpha}{N}$ (in a sense, the walker "jumps" from j to i). The introduction of the addendum accounting for jumps guarantees the convergence of the formula above to the stationary distribution of this dynamical process (which exists and is unique—its Markov chain, in fact,

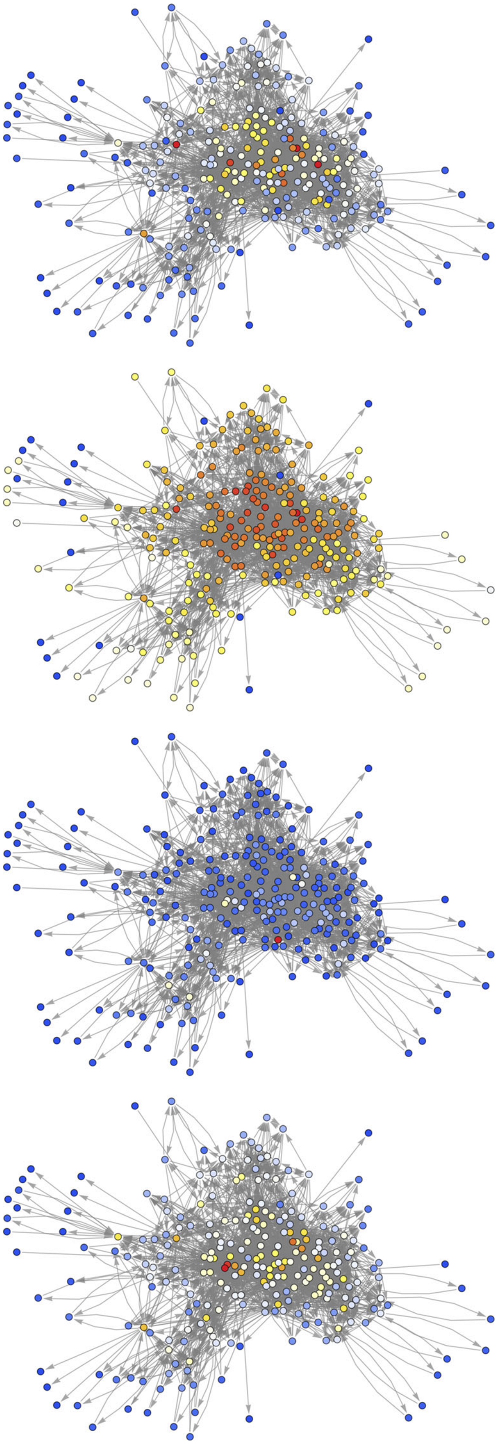

Fig. 2 *C. Elegans* neural network [14]. From top to bottom, nodes are ranked according to their (out-)degree-centrality, closeness-centrality, PageRank-centrality, InfoRank (red nodes are ranked higher than blue nodes). Notice that, according to PageRank, (only) the node with largest in-degree is ranked first; the same node, however, is characterized by a zero out-degree which, in turn, causes its closeness-induced score to be zero as well (color figure online)

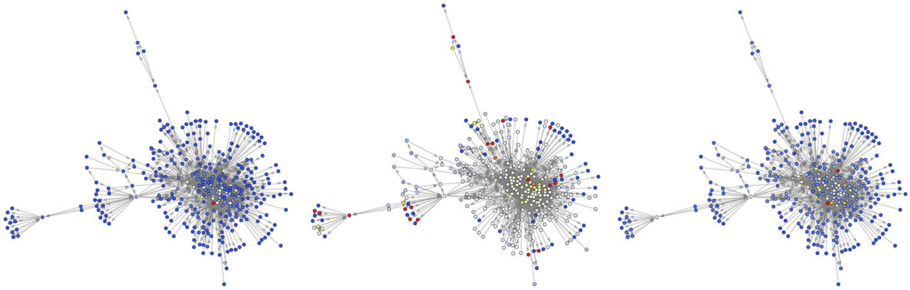

Fig. 3 US airports network in 1997 [15]. Nodes are ranked according to their value of (out-)degree-centrality (left), closeness-centrality (center) and InfoRank (right—red nodes are ranked higher than blue nodes). Nodes with a large (out-)degree-centrality (hubs) do not necessarily coincide with the nodes characterized by a large value of closeness-centrality: in fact, although many nodes can be reached by a walker leaving the hubs, these may lie at a large distance from it (color figure online)

becomes strongly connected and aperiodic by construction), providing the searched ranking scores.

By oversimplifying a bit, PageRank scores higher the so-called "authoritative" nodes, i.e. nodes that are pointed by a large number of vertices which, in turn, have low out-degree [6]. The evidence that nodes with a large PageRank value do not necessarily coincide with the nodes having a large in-degree is provided by the Little Rock food web; in this particular case, a couple of species predated by a limited number of predators can be, indeed, observed: the former, however, constitute the only preys of the latter. In all the other cases the correlation coefficient between the vectors $\{P_i\}$ and $\{k_i^{in}\}$ is quite large: 0.70 for the US airports network (in 1997), 0.82 for the *C. Elegans* neural network, 0.99 for both the World Trade Web and the e-MID interbank network (notice that upon lowering α the two vectors become less correlated, since the random contribution to the dynamics becomes the prevalent one). Such a correlation has been also noticed elsewhere [17].

Let us now consider our novel InfoRank index. As a first observation, the ranking induced by it shows a little overlap with the one provided by the other indices, thus confirming its degree of novelty. The intuitive idea according to which the nodes with largest InfoRank are the ones disconnecting the largest number of subgraphs is confirmed upon looking at the fourth panel of Fig. 2 and the third panel of Fig. 3: when considering either the *C. Elegans* neural network or the US airport networks, in fact, vertices acting as "junctions" between a group of leaves and the remaining part of the network are often assigned an InfoRank value that is larger than the one assigned to the "most internal" nodes. Notice that when directed networks are considered, reducing uncertainty does not necessarily imply isolating nodes: information can be gained, in fact, also by determining just the out- or the in-degrees.

3.3 Exploring the InfoRank Degree-Dependence

Let us now consider the World Trade Web (WTW) [18]. The main reason to include it in our analysis lies in the value of its link density: being much denser than the other networks considered so far, it also allows us to better understand the relationship between the InfoRank value of a node and its degree(s).

Since the WTW topological structure can be deduced with great accuracy from the knowledge of its degree-sequence(s) [19], we may also expect the latter to be correlated with the ranking indices considered for the present analysis. This is indeed the case. As the fourth

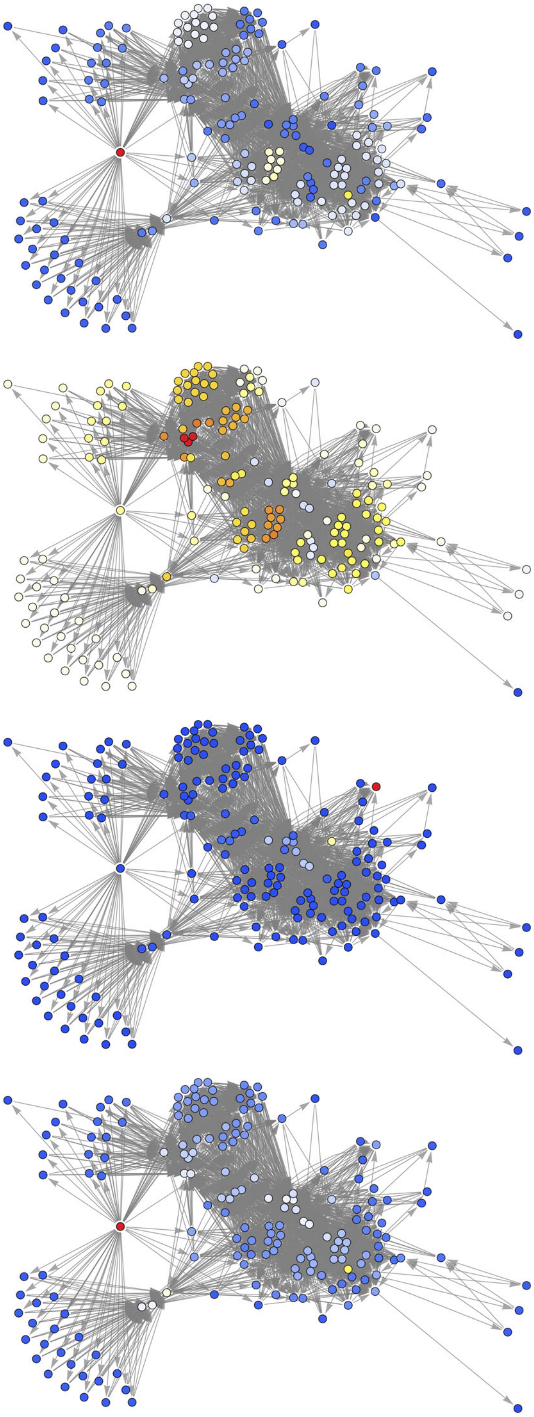

Fig. 4 Little Rock food web [16]. From top to bottom, nodes are ranked according to their (out-)degree-centrality, closeness-centrality, PageRank-centrality, InfoRank (red nodes are ranked higher than blue nodes). Nodes with large in-degree do not necessarily coincide with nodes having a large value of PageRank: this is evident in the case of food-webs, where species exist that are predated by a limited number of predators of which constitute the only preys. In other networks however, the correlation between the PageRank value and the in-degree is quite large (color figure online)

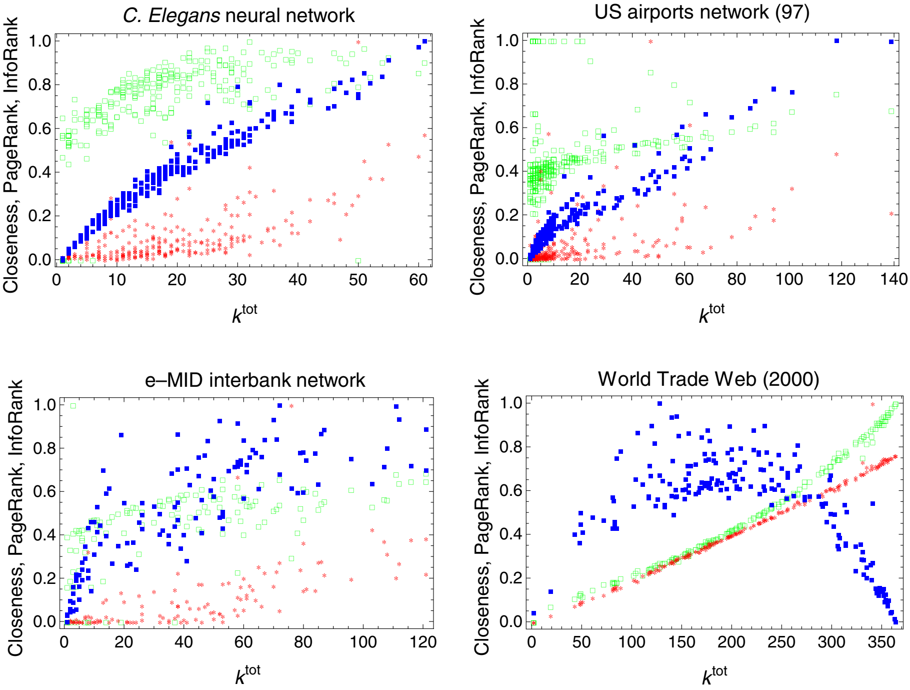

Fig. 5 Dependence of the (rescaled) ranking indices considered for the present analysis (closeness-centrality open green square, PageRank centrality red asterisk, InfoRank filled blue square) on the nodes total degree. Notice how both e-MID and the World Trade Web are characterized by a strongly positive correlation between the closeness-centrality and the total degree and the PageRank centrality and the total degree; InfoRank, instead, is characterized by a bell-shaped trend for the same systems, whose point-of-maximum lies close to the value $k_i^{tot} \simeq (N-1)/2 + (N-1)/2 = N-1$. Although the nodes providing the largest contribution to the entropy reduction overlap with the ones maximizing $S_0^{(i)}$, this doesn't imply that the removal of a given node has a small impact on the other vertices (see also Fig. 6). For what concerns sparser systems, instead, InfoRank shows an overall increasing trend while a clear functional dependence between closeness-centrality and total degree and PageRank centrality and total degree is not visible (a weakly positive correlation between closeness-centrality and total degree is, however, present in the *C. Elegans* neural network) (color figure online)

panel of Fig. 5 shows, both the closeness-centrality and the PageRank indices are highly correlated with the total degree (i.e. $k_i^{tot} = k_i^{out} + k_i^{in}$). The monotonic, increasing, relationship between total degree and closeness-centrality can be explained by supposing that all countries have established a direct connection with the nodes that can be reached by them via some other (indirect) path. This is not true, for example, for the Little Rock food web shown in Fig. 4: in that case, the node with largest out-degree is *directly* connected to only some of the *reachable* nodes (as a consequence, the overall distance from the set of such vertices increases).

The monotonic, increasing, relationship between total degree and PageRank, instead, rests upon a double (empirical) evidence: countries with a large out-degree are 1) also characterized by a large in-degree and are usually 2) "pointed" by countries with a small out-degree.

InfoRank, on the other hand, shows an overall bell-shaped trend with a maximum in correspondence of the values $k_i^{tot} \simeq (N-1)/2 + (N-1)/2 = N-1$. This means that the nodes providing the largest contribution to the entropy reduction overlap with the ones

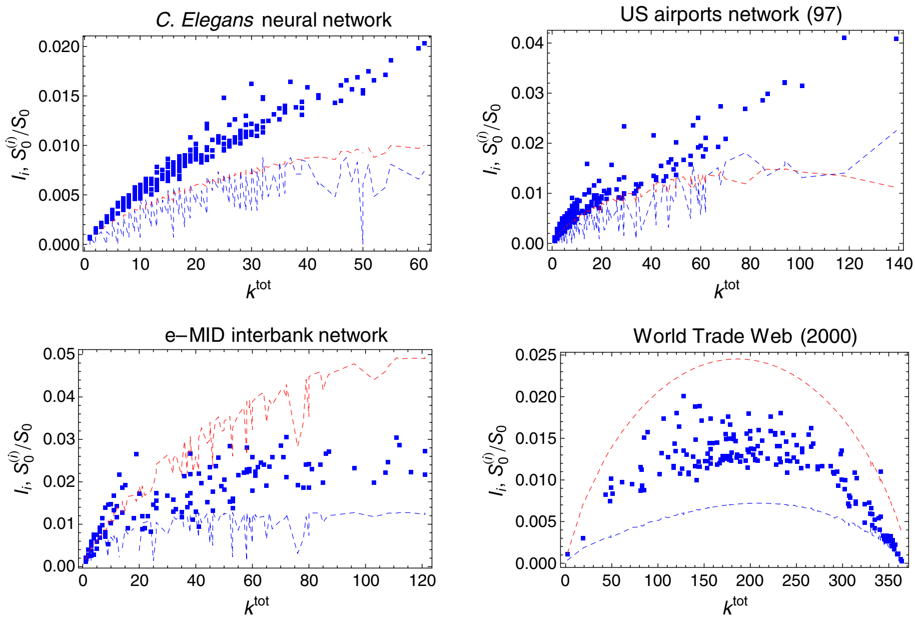

Fig. 6 Comparison between InfoRank I_i (filled blue square) and $S_0^{(i)}$ (blue, dashed line). We have also tested the agreement between the latter and the two approximations derived in Appendix A (both indicated with a red, dashed line). While, for sparse networks (upper panels), $S_0^{(i)} \simeq \sum_{(b)=in,out}\left[-k_i^{(b)}\ln(k_i^{(b)}/\sqrt{L})+k_i^{(b)}\right]$, for dense networks (lower panels) a non-trascurable difference exists between $S_0^{(i)}$ and $-(N-1)\sum_{(b)=in,out}\left[\overline{p}_{ij}^{(b)}\ln\overline{p}_{ij}^{(b)}+\left(1-\overline{p}_{ij}^{(b)}\right)\ln\left(1-\overline{p}_{ij}^{(b)}\right)\right]$ with $\overline{p}_{ij}^{(b)}=k_i^{(b)}/(N-1)$ (color figure online)

maximizing $S_0^{(i)}$. However, as evident upon inspecting Fig. 6, evident deviations from the $S_0^{(i)}$ trend are clearly visible: InfoRank adjusts the estimation provided by exclusively accounting for the nodes degrees, although its *functional dependence* on them is, overall, similar to the one characterizing $S_0^{(i)}$.

3.4 Exploring the Relationship Between InfoRank and the Reconstruction Accuracy

As we have seen, InfoRank individuates the node(s) reducing the network residual uncertainty to the largest extent. We may, thus, suspect InfoRank to also "select" the nodes able to provide the best reconstruction of the network itself. In order to verify our conjecture, we have explicitly tested the agreement between the reconstruction achieved by each node and the observed network structure. In order to do so, we have computed an index often employed to test the (global) goodness of a reconstruction algorithm: the *accuracy*, defined as $\langle A\rangle = \frac{\langle TP\rangle+\langle TN\rangle}{N(N-1)}$ where $\langle TP\rangle$ is the expected number of true positives, i.e. $\langle TP\rangle=\sum_i\sum_{j(\neq i)}a_{ij}p_{ij}$, $\langle TN\rangle$ is the expected number of true negatives, i.e. $\langle TN\rangle=\sum_i\sum_{j(\neq i)}(1-a_{ij})(1-p_{ij})$ and N is the total number of vertices [7]. We have then summarized our findings by calculating the correlation between the vector $\{A_i\}$ and the vector $\{I_i\}$.

The results are reported in Table 1: while the correlation between InfoRank and accuracy is almost 1, when comparing the latter with the ranking values obtained via alternative indices a worse agreement is found. In particular, when e-MID and the WTW are considered, negative

Table 1 Table showing the Pearson correlation coefficient between the vector of accuracy values A_i and the vector of (rescaled) ranking scores $\overline{R}_i$

$r_{A_i,\overline{R}_i}$	D_i	C_i	P_i	I_i
Little Rock food web	0.44	0.34	0.27	**0.97**
C. Elegans neural network	0.82	0.60	0.65	**0.98**
US airports network (1997)	0.89	0.39	0.89	**0.99**
e-MID interbank network	0.52	0.44	0.57	**0.99**
World Trade Web (1950)	0.097	0.008	0.098	**0.99**
World Trade Web (1970)	− 0.1	− 0.23	− 0.15	**0.99**
World Trade Web (2000)	− 0.39	− 0.52	− 0.42	**0.99**

Bold numbers indicate the largest correlation coefficient among the ones characterizing the considered algorithms

The transformation (which doesn't affect the correlation value) reads $\overline{R}_i^{(a)} = (R_i^{(a)} - \min\{R_i^{(a)}\})/(\max\{R_i^{(a)}\} - \min\{R_i^{(a)}\}) \in [0, 1]$ with $R_i^{(a)}$ representing the rank of node i according to the chosen algorithm a

correlation values are observed: this is due to the bell-shaped trend recovered when scattering the accuracy values versus any of the chosen ranking indicators, as shown in Fig. 7.

3.5 Exploring the Relationship Between InfoRank and the Systemic Risk Estimation

In this subsection, we provide evidence of how a better reconstruction accuracy can, in turn, lead to a better estimation of relevant properties of a financial system. Let us focus on a real network of transaction in an interbank money market. Interbank money markets are essential for financial institutions as sources of liquidity provision. In such markets, information asymmetry [20] translates into a better estimation of the expected payments, widely recognized as a measure of systemic risk in networks of interbank liabilities [21]. Here, we focus on data from the e-MID platform (the electronic Market of Interbank Deposit), that served a significant percentage ($\sim$ 17%) of the unsecured money market in the Euro Area before the 2007–2008 crisis [23].

We apply the clearing mechanism originally proposed in [21], in the generalization discussed in [22], and compute the payment vector, whose components represent the amount a financial institutions able to repay its creditors. When the payment of a bank is less then its corresponding obligation, that bank is considered insolvent: hence, computing the payment vector corresponds to identify insolvencies and estimate systemic risk in a financial network (a detailed discussion of such measures of systemic risk is found in [24,25]). Insolvency of bank occurs when its equity, the difference between assets and liabilities, becomes negative. The external cash flow is given by the external assets A_e, affected by fire sales in case of insolvency, and external liabilities L_e. Both are sampled from a Gaussian distribution, with parameters $\mu_a = 10$, $\sigma_a = 0.1$ and $\mu_l = 1$, $\sigma_l = 0.1$, respectively.

First, we compute the payment vector $\{p_i^{(r)}\}$—that entails the information on systemic risk losses—starting from the real e-MID network; second, we sample networks from the specific ensemble of each node and compute the payment vector on each drawn configuration, in order to evaluate the normalized squared error of these "sampled" payment vectors $\{p_i^{(s)}\}$ with respect to the real one. Finally, we calculate the mean over the set of sampled payment

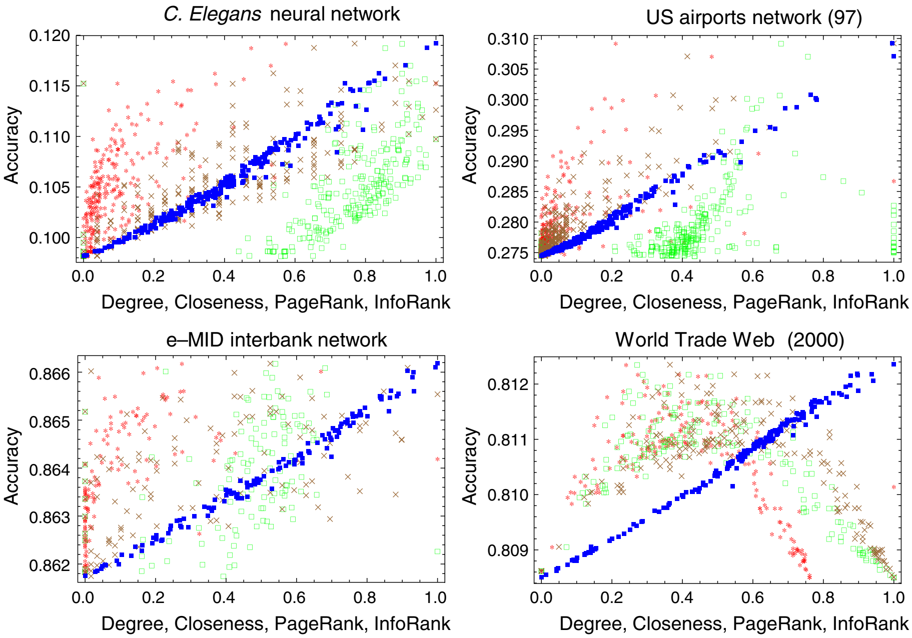

Fig. 7 Dependence of the accuracy value on the (rescaled) ranking indices considered for the present analysis ((out-)degree-centrality brown times symbol, closeness-centrality open green square, PageRank centrality red asterisk, InfoRank filled blue square). Notice the clear, increasing, trend describing the functional dependence of the accuracy value on the InfoRank value, further confirming that the node(s) establishing the most informative sets of interconnections are the ones characterized by the largest InfoRank value(s)

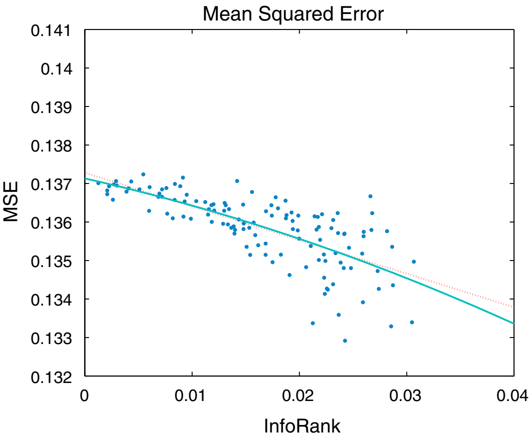

Fig. 8 Mean squared error over the payment vector of the financial clearing process on the e-MID network. The parameters that account for fire sales effect and insolvency costs in the Rogers and Veraart [22] clearing mechanism are $\alpha = \beta = 0.9$. The dotted line is a linear fit of the data $y = -0.087 * x + 0.14$ (RSS = 0.0070645), while the solid line is a quadratic fit $y = -0.79 * x^2 - 0.063 * x + 0.14$ (RSS = 0.007046) (color figure online)

vectors and obtain the mean squared error affecting each node estimation of the chosen risk measure: as Fig. 8 shows, errors are smaller for nodes with larger InfoRank. This, in turn, sheds light on the relationship between financial risk and network topology, proving that a better knowledge of the latter indeed leads to a better estimation of the former.

4 Discussion

InfoRank represents a novel measure of the relevance of nodes in a network. We have approached this problem from an information-theoretic point of view, by quantifying the information content of each node-specific pattern of interconnections.

Differently from other existing indices, InfoRank can both be employed to analyze any kind of network, be it directed, weighted, etc.; and can quantify the informativeness of *whole subsets* of nodes: this is usually a major limitation for the other centrality indicators, tailored to provide single-nodes estimates.

Such an approach allows us to explore the relationship between the proposed index and the much more general concept of *information asymmetry* which is supposed to affect the interactions between agents in financial systems. As our example about financial contagion shows, the competitive advantage represented by a larger amount of *information about the network* leads to a better estimation of systemically-important properties.

It is evident, however, that computing the whole InfoRank vector requires the knowledge of the entire network.

On one hand, complete knowledge about the network structure can be accessible to an external authority, e.g. a central bank or a regulatory agency, interested in monitoring information asymmetry. The authors argue that, in a similar scenario, this kind of authority could be interested in the knowledge of the extent of information asymmetry in the market, as an indicator of market (in)efficiency. Furthermore, the general definition of InfoRank would allow it to identify the *minimal subset* of nodes allowing for the complete knowledge of the network structure, potentially constituting a warning signal for the emergence of cartels. This, however, requires additional analysis which constitutes the subject of ongoing research.

On the other hand, in case of missing information about the actual network structure, a node-specific InfoRank value can be compared with an expected one, as obtained by implementing a benchmark null model [7,13]. In this way, individual nodes could, in any case, test their *competitive advantage* against some null hypothesis.

Finally, the ability of identifying highly informed nodes—characterized by high InfoRank values—may also provide strategies to optimally sample networks, when gathering information on individual nodes is costly (e.g. when surveying a financial system for regulatory purposes).

Acknowledgements PB and TS acknowledge support from: FET Project DOLFINS No. 640772 and FET IP Project MULTIPLEX No. 317532.

Appendix A

Here we show how the computation of $S_0^{(i)}$ can be simplified in two cases of general interest. The first one concerns sparse networks: since, in this case, the probability coefficients defined by Eq. (1) satisfy the requirement $p_{ij} \ll 1$, the following factorization holds $p_{ij} \simeq x_i x_j$, further implying that

$$S_0^{(i)} \simeq -\sum_{j(\neq i)} [p_{ij} \ln p_{ij} - p_{ij}] = -k_i \ln \left(\frac{k_i}{\sqrt{2L}} \right) + k_i. \qquad (12)$$

The second approximation is valid whenever the node i-specific probability coefficients are well represented by their average value, i.e. $p_{ij} \simeq \frac{k_i}{N-1} \equiv \overline{p}_{ij}$; in this case,

$$S_0^{(i)} \simeq -(N-1)\left[\overline{p}_{ij} \ln \overline{p}_{ij} + \left(1-\overline{p}_{ij}\right) \ln \left(1-\overline{p}_{ij}\right)\right]. \tag{13}$$

Appendix B

This second appendix collects the details of the derivation of our proposed methodology. Let us focus on the simplest case of a single node (hereafter indexed by l): in order to calculate InfoRank it can be imagined to solve two different problems. The first one concerns the maximization of the functional

$$\begin{aligned} S_0 = &-\sum_{\mathbf{G}} P(\mathbf{G}) \ln P(\mathbf{G}) + \\ &-\sum_i \eta_i \left[\sum_{\mathbf{G}} P(\mathbf{G}) C_i(\mathbf{G}) - C_i^*\right] \end{aligned} \tag{14}$$

i.e. the *constrained* Shannon entropy, constraints encoding the benchmark information accessible by all nodes (represented by the vector of M constraints $\vec{C}^*$—notice that the normalization condition of the probability distribution, $P(\mathbf{G}|\vec{\eta})$, to be determined can be re-written as an $M+1$-th constraint of the kind $C_{M+1}(\mathbf{G}) = C_{M+1}^* = 1$) [13]. By solving the constrained-optimization problem in (14), node l finds that

$$S_0 = \vec{\eta} \cdot \vec{C}^* + \ln Z(\vec{\eta}). \tag{15}$$

(where $Z(\vec{\eta}) = \sum_{\mathbf{G}} e^{-\vec{\eta}\cdot\vec{C}(\mathbf{G})}$ is the so-called *partition function* and depends on the unknown Lagrange multipliers $\vec{\eta}$). On the other hand, the second optimization problem node l has to solve concerns the functional

$$S_{(l)} = S_0 - \sum_m \psi_{lm} \left[\sum_{\mathbf{G}} P(\mathbf{G}) a_{lm}(\mathbf{G}) - a_{lm}^*\right] \tag{16}$$

with $S_{(l)}$ being nothing else than the functional in (14) further constrained by imposing the ego-network of node l as well (i.e. the values of the link-specific variables a_{lm}^*—either 0 or 1). Upon solving the second problem, the expression

$$S_{(l)} = \vec{\theta} \cdot \vec{C}^* + \sum_m \psi_{lm} a_{lm}^* + \ln Z'(\vec{\theta}, \vec{\psi}) \tag{17}$$

(where $Z'(\vec{\theta}, \vec{\psi}) = \sum_{\mathbf{G}} e^{-\vec{\theta}\cdot\vec{C}(\mathbf{G}) - \sum_m \psi_{lm} a_{lm}(\mathbf{G})}$) is found. Notice that although $S_{(l)}$ and S_0 are defined by the same vector of constraints, $\vec{C}$, the numerical values of the Lagrange multipliers ensuring that $\langle\vec{C}\rangle = \vec{C}^*$ will, in general, differ, whence the use of different symbols, i.e. $\vec{\eta}$ and $\vec{\theta}$.

Both functionals achieve a minimum in their stationary point (consistently with our attempt to minimize each node—residual—uncertainty). This can be easily proven, upon noticing that the Hessian matrix of both S_0 and $S_{(l)}$ is the covariance matrix of the constraints and, as such, positive-semidefinite. In order to find the stationary point of $S_{(l)}$, node l must solve the equations

$$\frac{\delta S_{(l)}}{\delta \theta_i} = 0,\ \forall\, i \quad \text{and} \quad \frac{\delta S_{(l)}}{\delta \psi_{lm}} = 0,\ \forall\, m \tag{18}$$

which lead to the system of equations in (4). More explicitly, the second group of conditions reads

$$\sum_{\mathbf{G}} \left(\frac{e^{-\vec{\theta}\cdot\vec{C}(\mathbf{G})-\sum_m \psi_{lm} a_{lm}(\mathbf{G})}}{Z'(\vec{\theta},\vec{\psi})} \right) a_{lm}(\mathbf{G}) = a^*_{lm}, \ \forall\, m; \tag{19}$$

in order to numerically evaluate the parameters $\vec{\psi}$, let us focus on a specific value, e.g. ψ_{l1} controlling for the value of the entry a_{l1}. Let us now explicitly distinguish the configurations characterized by $a_{l1} = 0$ from the ones with $a_{l1} = 1$: upon doing so, condition (19) can be rewritten as

$$\sum_{\mathbf{G}_1} \left(\frac{e^{-\vec{\theta}\cdot\vec{C}(\mathbf{G}_1)-\psi_{l1}-\sum_{m(\neq 1)} \psi_{lm} a_{lm}}}{Z'(\vec{\theta},\vec{\psi})} \right) = a^*_{l1} \tag{20}$$

i.e. as a sum over only the configurations with $a_{l1} = 1$ (indicated with the symbol $\mathbf{G}_1$). Analogously, we can split $Z'(\vec{\theta},\vec{\psi})$ into the sum of two terms, i.e. $Z'(\vec{\theta},\vec{\psi}) = Z'_0(\vec{\theta},\vec{\psi}) + e^{-\psi_{l1}} Z'_1(\vec{\theta},\vec{\psi})$, where the first sum

$$Z'_0(\vec{\theta},\vec{\psi}) = \sum_{\mathbf{G}_0} e^{-\vec{\theta}\cdot\vec{C}(\mathbf{G}_0)-\sum_{m(\neq 1)} \psi_{lm} a_{lm}} \tag{21}$$

runs over the networks having $a_{l1} = 0$ and the second sum

$$Z'_1(\vec{\theta},\vec{\psi}) = \sum_{\mathbf{G}_1} e^{-\vec{\theta}\cdot\vec{C}(\mathbf{G}_1)-\sum_{m(\neq 1)} \psi_{lm} a_{lm}} \tag{22}$$

runs over the networks having $a_{l1} = 1$.

Solving Eq. (20) in the case $a^*_{l1} = 0$ leads to $\psi_{l1} = +\infty$. As a consequence, in this case $S_{(l)} = \vec{\theta}\cdot\vec{C}^* + \ln Z'_0(\vec{\theta},\vec{\psi})$ since the term $Z'_1(\vec{\theta},\vec{\psi})$ is suppressed by the coefficient $e^{-\psi_{l1}}$ that converges to zero. On the other hand, solving Eq. (20) in the case $a^*_{l1} = 1$ leads to $\psi_{l1} = -\infty$ and $S_{(l)} = \vec{\theta}\cdot\vec{C}^* + \ln Z'_1(\vec{\theta},\vec{\psi})$ since the term $Z'_0(\vec{\theta},\vec{\psi})$ is now suppressed by the coefficient $e^{\psi_{l1}}$ (this is readily seen by multiplying both the numerator and the denominator at the left-hand side of Eq. (20) by $e^{\psi_{l1}}$). Specifying the node-specific ego-networks, in other words, leads to reducing the number of configurations over which the estimation of the constraints is carried out: $Z'(\vec{\theta})$, thus, runs over a smaller number of configurations than $Z(\vec{\eta})$. The estimation of the other parameters $a_{l2} \dots a_{lN}$ proceeds in an analogous way, by applying the same line of reasoning to the "surviving" partition functions.

Let us now evaluate the expressions $Z(\vec{\eta})$ and $Z'(\vec{\theta})$ for the *same value* of the parameters (say $\vec{\mu}$): since the number of addenda in $Z(\vec{\mu})$ is larger than the number of addenda in $Z'(\vec{\mu})$, it also holds true that $\ln Z(\vec{\mu}) \geq \ln Z'(\vec{\mu})$, in turn implying the inequivalence $S_0(\vec{\mu}) \geq S_{(l)}(\vec{\mu})$ to be true as well. Let us now choose a particular value of the parameters, i.e. the point of minimum of S_0: $\vec{\mu} = \vec{\eta}^*$. Thus,

$$S_0(\vec{\eta}^*) \geq S_{(l)}(\vec{\eta}^*) \geq S_{(l)}(\vec{\theta}^*) \tag{23}$$

where the second inequality follows from the very definition of minimum. This ensures the ratio $S_{(l)}/S_0$ to be smaller than one and the InfoRank index in Eq. (7) to be always well-defined.

Our ranking procedure builds upon the evidence that, by imposing more information on top of the common one, each node further reduces its uncertainty about the unknown network structure: the one reducing the residual uncertainty to the largest extent is identified as the "most informative" one.

The same line of reasoning applies when subsets of nodes are considered, although the resolution of such a problem may be computationally demanding: given a network of size N, quantifying the InfoRank of all possible subsets of s nodes would require computing $\binom{N}{s}$ different Shannon entropies.

References

1. Newman, M.E.J.: Networks: An Introduction. Oxford University Press, New York (2010)
2. Bloch, F., Jackson, M.O., Tebaldi, P.: Centrality measures in networks (2017). arXiv:1608.05845
3. Borgatti, S.P.: Centrality and network flow. Soc. Netw. **27**, 55–71 (2005)
4. Benzi, M., Klymko, C.: A matrix analysis of different centrality measures. SIAM J. Matrix Anal. Appl. **36**, 686–706 (2013). https://doi.org/10.1137/130950550
5. Sabidussi, G.: The centrality index of a graph. Psychometrika **31**, 581–603 (1966)
6. Langville, A.N., Meyer, C.: Google's PageRank and Beyond. Princeton University Press, Princeton (2006)
7. Squartini, T., Cimini, G., Gabrielli, A., Garlaschelli, D.: Network reconstruction via density sampling. Appl. Netw. Sci. **2**(3) (2017). https://doi.org/10.1007/s41109-017-0021-8
8. Zhang, Q., Meizhu, L., Yuxian, D., Yong, D.: Local structure entropy of complex networks (2014). arXiv:1412.3910v1
9. Bianconi, G., Pin, P., Marsili, M.: Assessing the relevance of node features for network structure. PNAS **106**(28), 11433–11438 (2009). https://doi.org/10.1073/pnas.0811511106
10. Bianconi, G.: The entropy of randomized network ensembles. Europhys. Lett. **81**(2), 28005 (2007)
11. Borgatti, S.P.: Identifying sets of key players in a social network. Comput. Math. Organ. Theory **12**, 21–34 (2006). https://doi.org/10.1007/s10588-006-7084-x
12. Park, J., Newman, M.E.J.: The statistical mechanics of networks. Phys. Rev. E **70**, 066117 (2004). https://doi.org/10.1103/PhysRevE.70.066117
13. Squartini, T., Garlaschelli, D.: Maximum-Entropy Networks. Pattern Detection, Network Reconstruction and Graph Combinatorics. Springer Briefs in Complexity. Springer, Cham (2018)
14. Oshio, K., Iwasaki, Y., Morita, S., Osana, Y., Gomi, S., Akiyama, E., Omata, K., Oka, K., Kawamura, K.: Tech. Rep. of CCeP, Keio Future 3. Keio University, Tokyo (2003)
15. Colizza, V., Pastor-Satorras, R., Vespignani, A.: Reaction-diffusion processes and metapopulation models in heterogeneous networks. Nat. Phys. **3**, 276–282 (2007)
16. Martinez, N.D.: Artifacts or attributes? Effects of resolution on the Little Rock Lake food web. Ecol. Monogr. **61**(4), 367–392 (1991)
17. Fortunato, S., Boguna, M., Flammini, A., Menczer, F.: Approximating PageRank from in-Degree in Lecture Notes in Computer Science 4936. Springer, Berlin (2008)
18. Gleditsch, K.S.: Expanded trade and GDP data. J. Confl. Resolut. **46**, 712–724 (2002)
19. Squartini, T., Fagiolo, G., Garlaschelli, D.: Randomizing world trade. I. A binary network analysis. Phys. Rev. **E84**, 046117 (2011). https://doi.org/10.1103/PhysRevE.84.046117
20. Wittenberg-Moerman, R.: The role of information asymmetry and financial reporting quality in debt trading: evidence from the secondary loan market. J. Account. Econ. **46**(2), 240–260 (2008)
21. Eisenberg, L., Noe, T.H.: Systemic risk in financial systems. Manag. Sci. **47**(2), 236–249 (2001)
22. Rogers, L.C.G., Veraart, L.A.M.: Failure and rescue in an interbank network. Manag. Sci. **59**(4), 882–898 (2013)
23. Barucca, P., Lillo, F.: The organization of the interbank network and how ECB unconventional measures affected the e-MID overnight market (2015). arXiv:1511.08068
24. Glasserman, P., Young, P.H.: Contagion in financial networks. J. Econ. Lit. **54**(3), 779–831 (2016)
25. Barucca, P., Bardoscia, M., Caccioli, F., D'Errico, M., Visentin, G., Battiston, S., Caldarelli, G.: Network valuation in financial systems (2016). arXiv:1606.05164