Corrections on Scientific Report: “A detailed characterization of complex networks using Information Theory”

Cristopher G. S. Freitas, Andre L. L. Aquino, Heitor S. Ramos,
Alejandro C. Frery, Osvaldo A. Rosso

December 2019

Abstract

Dear Chief Editor, recently, we have been informed that our publication “A detailed characterization of complex networks using Information Theory” has an important problem that must be considered before calculating the **Network Fisher Information Measure** for complex networks. Dr. Sang Hoon Lee and Dr. Vinko Zlatic pointed out that, since the quantifier we are proposing for networks consider the node indices to build the random-walk-based distribution, if we permute these indices without altering the network structure, our results for Network Fisher may change. This is, indeed true, and we acknowledge that this is problematic when using our proposal. Nonetheless, we have found an adequate solution that enhances our results for both synthetic and real networks; and does not compromise the initial discussion, only improves it, but it comes with additional computational cost. Thus, as our paper did not address this issue at first, we write you to add the discussion of the dependency of Fisher Information Measure on node indices, and to update the results considering this preprocessing before the calculations.

1 Problem Description

In this work, we use two Information Theory quantifiers, namely Network Entropy and Network Fisher Information Measure, to characterize complex networks accurately. We evaluate our results in a wide range of different network varieties from synthetic to real-world systems. All the codes and datasets used in the paper are public since its publication date. During the preparation of this work, we conducted a wide variety of experiments to show the robustness of our results. Thus, our proposal works well in the networks with labels ordered in its conception. For instance, when we take the same graph with different labels order, the results are different.

Vinko Zlatic and Sang Hoon Lee indicated that our measure would fail for a random permutation of the node labels, as it changes the adjacency matrix $\mathbf{A}$ that the quantifier relies on. Besides that, they suggested that the Network Fisher Information Measure $\mathcal{F}$ does not follow the standard definition of the former discrete Fisher for time series:

> “Fisher information in principle measure the amount of new information needed to describe the system if the parameters changed.”

Also, they stated that the changes in node labels should not, in any way, affect the amount of information needed to describe the system. We respectfully disagree. The system’s representation may change the amount of information that we can extract from it, and for this, we understand that there are $n!$ matrices $\mathbf{A}$ that describe the same network, but there may be an optimal representation $\mathbf{A}^*$ that can reveal the most patterns, if they do exists [Behrisch et al., 2016]. Finally, Zlatic and Lee suggested that we checked the labeling issue, reproducing the results for WS and BA, after shuffling the node indices before calculating FIM.

We evaluated WS networks, the rewiring probability β, and the initial average degree k that control the system We expect a k-ring network for $\beta = 0$, and a random network such as ER graphs for $\beta = 1$. Initially, for k-ring networks, the algorithm creates a ring order (Fig. 1a) (e.g., 1–2, 2–3, 3–4, ... , n–1), which results in a very organized adjacency matrix (Fig. 1b), where varying k and β will result in the behavior originally shown in our paper.

The connections start to change when we increase the rewiring probability. Hence, the adjacency matrix will suffer from some "disorder," and the values of $\mathcal{F}$ will increase. Such "disorder" effect, as we observed based on the comments raised, can also be achieved easily by shuffling the node labels (Fig. 1c,1d): when we shuffle all the nodes, the values change from $\mathcal{F} = 0.5$ to $\mathcal{F} = 0.97$ as an example for k-ring with $k = 2$ and $N = 1000$. Nevertheless, these results become elusive, and we cannot conclude if the rewiring is causing the growth in $\mathcal{F}$, or it is merely the fact that we are reordering the matrix.

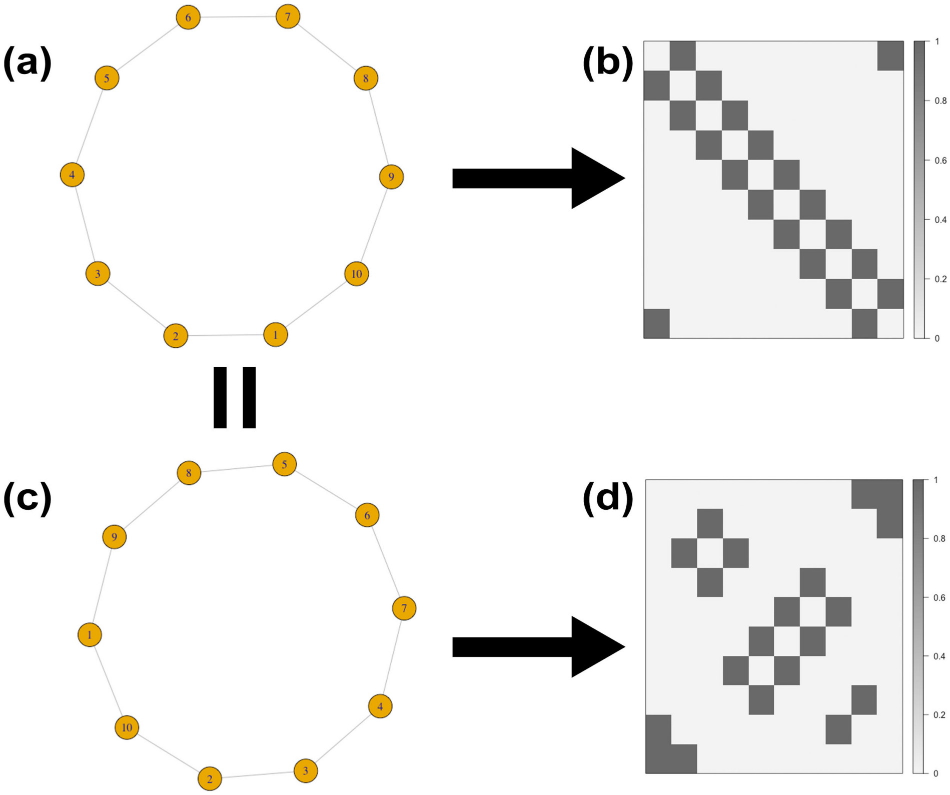

Figure 1: (1a) shows the 1-ring topology with $N = 10$ and original order from the algorithm available at *igraph*. (1b) shows a very ordered adjacency matrix for 1-ring with ordered labels. (1d) shows the 1-ring topology with $N = 10$ after randomly permuting the node labels. (1d) shows the resulting matrix after shuffling the node labels for the 1-ring topology.

There is a natural ordering in the algorithm for BA networks: the first node to be inserted into the system is node 1; the second is node 2, and so on. Thus, the results are always consistent with the analysis, but if after generating the network we shuffle the node labels, then the same effect happens. Although, as this model produces sparse networks, with a link density of around 0.003, the disturbance is less likely than what happens for k-ring with $k = 2$. Still, it does happen, and our results become inconclusive.

Initially, we did not find a single example of a synthetic network generated by those algorithms that failed the proposed characterization without the shuffling of node labels. Although, real-world networks produced some inconsistent results, all of them were reported in our paper.

2 Proposed solution

As the Fisher Information Measure is sensitive to the ordering of node labels, we want to find an order φ that maximizes the amount of information that we can extract from the system using the adjacency matrix $\mathbf{A}$, i.e., an optimal representation $\mathbf{A}^*$ that it can reveal the most patterns, if they do exist.

An ordering, or order, is a bijection $\varphi(v) \to i$ from $v \in V$ to $i \in N = \{1, \ldots, n\}$ that associates a unique index to each vertex. We denote one specific ordering from the set of all possible orderings as φ^*. Usually, a network comes with an arbitrary ordering that we call *initial order*, denoted $\varphi_0(v)$ to distinguish from a computed order. A *transformation* from one ordering to another is called a permutation $\boldsymbol{\pi}$. Formally, a permutation is a bijection $\boldsymbol{\pi}(x) \to y$ such that:

$$\boldsymbol{\pi}(x_i) = y_i, \ (x, y) \in N^2 \text{ where } y_i = y_j \implies i = j. \tag{1}$$

Each permutation is implemented as a vector containing n distinct indices in N. We denote S the set of all possible permutations $n!$ for n. A reordering of an undirected network G consists in computing one permutation $\boldsymbol{\pi} \in S$ that maximizes or minimizes an objective function $q(\boldsymbol{\pi}, G)$, such that:

$$\underset{\boldsymbol{\pi} \in S}{\arg\min}\, q(\boldsymbol{\pi}, G). \tag{2}$$

For each permutation $\boldsymbol{\pi}$, we may have a different value of Network Fisher Information Measure $\mathcal{I}(\boldsymbol{\pi}, G) \in \mathcal{F}$, where $\mathcal{F}$ is redefined as the set of all possible FIM for a given network G and permutation $\boldsymbol{\pi}$.

From our previous results, we observe that there exists a pattern of transitions between k-ring and random networks, where k-ring are the most ordered matrices with *block-diagonal patterns* and the lowest FIM values, with the exception of $k = 1$. Therefore, when the number of connections increases, the adjacency matrix $\mathbf{A}$ starts to saturate with ones, and the Network Fisher decreases. Thus, we choose a permutation $\boldsymbol{\pi}^*$ that results in the smallest FIM $\mathcal{I}^*$ for a given network G, such that:

$$\mathcal{I}^* = \underset{\boldsymbol{\pi}^* \in S}{\arg\min}\, \mathcal{I}(\boldsymbol{\pi}^*, G). \tag{3}$$

Now that we know what to find, we need to define how to do it. Finding the best possible solution for our problem is immediate if we run all the possible permutations, and we choose the one with the lowest possible value of FIM. However, this is not feasible, as we have $n!$ permutations for each undirected network G. For this, there are several algorithms for matrix reordering or *seriation*, as it is called.

The *block-diagonal pattern* is one of the most sought-after matrix patterns. It consists of coherent retangular areas that appear in ordered matrix whenever strongly connected components or cliques are present in the underlying topology. Initially, we are focused on finding the best possible solution, and for this task, the *Optimal-Leaf-Ordering* is the best algorithm as it finds an exact solution [Brandes, 2007, Bar-Joseph et al., 2001]. However, is the most expensive technique with a time of complexity of $\mathcal{O}(n^2 \log(n))$ and memory complexity of $\mathcal{O}(n)$.

Thus, we use this exact solution for $N < 10000$, but for $N > 10000$, we chose a sub-optimal algorithm that focuses on the angular order of eigenvectors Friendly [2002]; as we evaluate many networks throughout our study, it would be unfeasible to wait the exact reordering solution for a network with $N = 10000$ that takes around 6 hours to complete, while the sub-optimal solution takes around 30 minutes. Although the exact solution (obviously) performs better than the sub-optimal algorithm, the sub-optimal algorithm produces a very consistent result, and it is our understanding that this algorithm gives us more information than the natural ordering of the system, then, enhancing our results.

3 Results and Discussion

Considering these new findings, we ran all the synthetic networks experiments again. The new approach either improved many aspects of the original proposal of the Shannon-Fisher plane, or produced very similar results. Although, the computational cost of our approach went much higher, and the analysis of huge networks may become impracticable. Erdős-Rényi networks continue with a similar behavior: there is a transition in between disconnected and connected networks, and a later saturation of the Fisher Information Measure. However, the gap in between the different network sizes increased as we can observe in Figure 2.

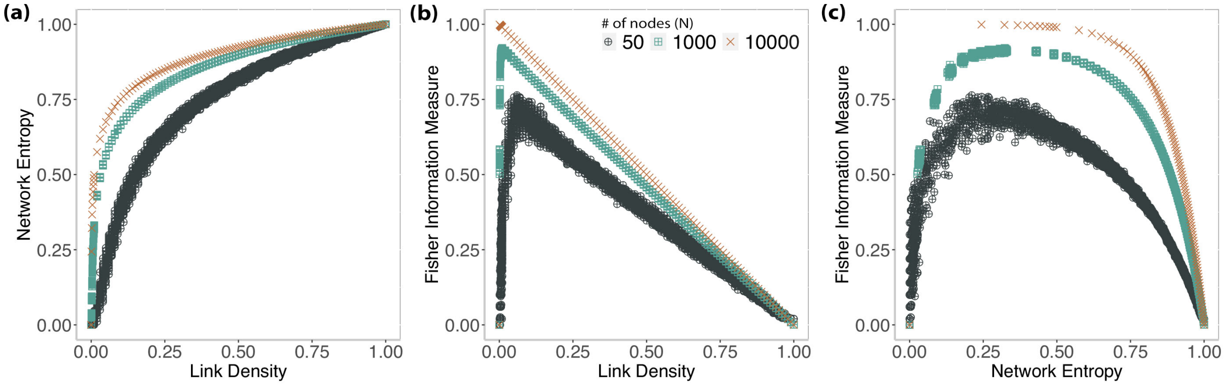

Figure 2: Results showing the relationship of Shannon Entropy and Fisher Information Measure with link density ((**a**) and (**b**)), and between Fisher Information Measure and Shannon Entropy (**c**) for 50 independent Erdős-Rényi networks except when $N = 10000$. The dark-green circles correspond to $N = 50$; the light-green squares to $N = 1000$; and the orange crosses to $N = 10000$.

Watts-Strogatz networks remain with a very similar behavior to what we initially observed. The reordering of the adjacency matrix actually solved a problem that we had in networks with $k = 1$: they overlapped with BA networks. Besides that, as we see in Figure 3, the upper bound for random networks ($WS_{k\approx 1}$) has lower values than the original, as it no longer approaches one (1).

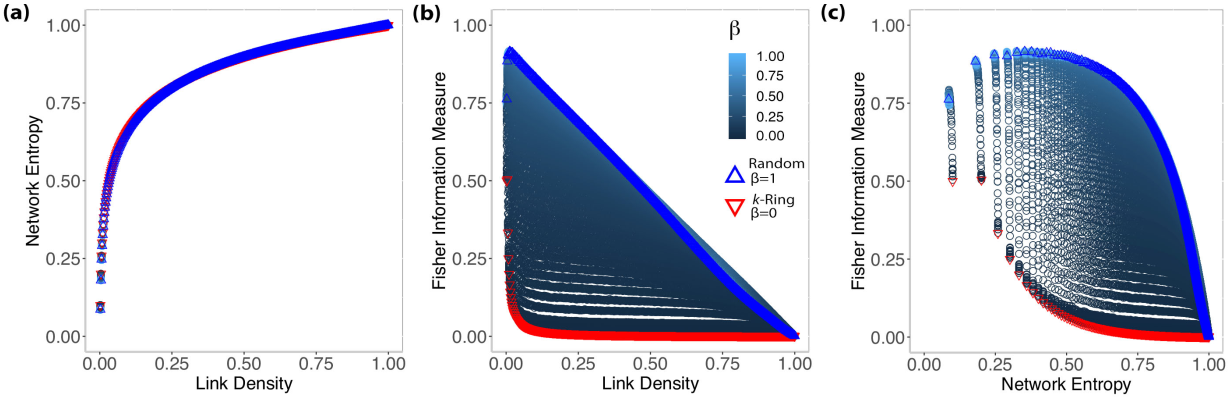

Figure 3: Relationship between Shannon Entropy and Fisher Information Measure with link density ((**a**) and (**b**)), and between Fisher Information Measure and Shannon Entropy for Watts-Strogatz networks (**c**). We restricted the analysis to $N = 1000$, $k \in \{1, 2, 3, \ldots, 499, 500\}$ and $\beta \in \{0, 0.001, 0.002, \ldots, 0.99, 1\}$; the downward red triangles correspond to k-rings ($G_{N,k}$ with $\beta = 0$); the upwards blue triangles are random graphs ($G_{N,k}$ with $\beta = 1$). The blue gradient from dark to light corresponds to the rewiring probability β: the intensity of the blue color is inversely proportional to the value of β.

Figure 4 shows no intersections between Barabási-Albert networks (BA) and any other model, as we previously had. Also, there is no oscillation between $\mathcal{F} = 0.999$ and $\mathcal{F} = 0.5$ when increasing the α; we only have network FIM approaching $\mathcal{I}^{(*)} = 0.5$, cf.Fig. 4b, removing another source of possible confusion. Figures 5a,b remain very similar to the first results, but again, the values of FIM no longer approach one.

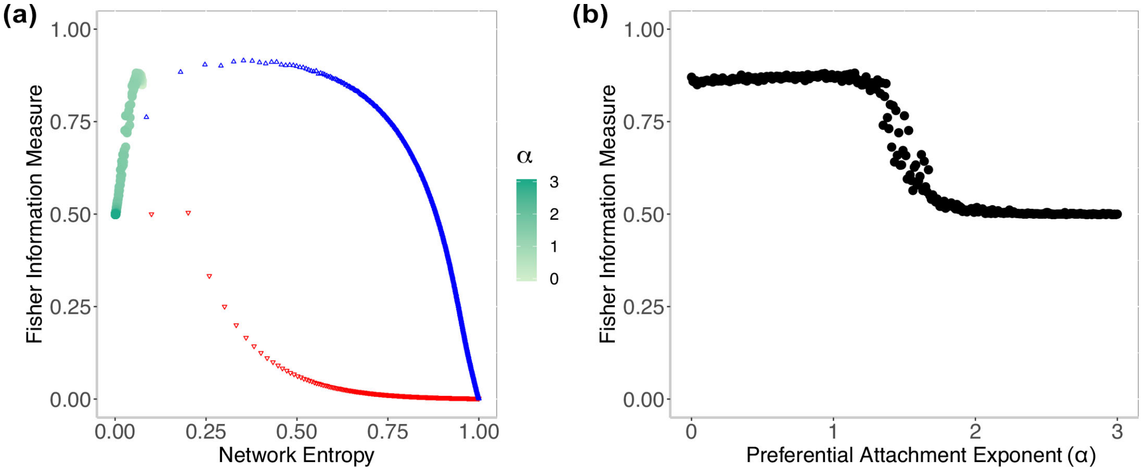

Figure 4: (4a) Barabási-Albert networks, non-linear preferential attachment with $N = 1000$, and $\alpha \in [0, 3]$. For the sake of visualization, we plot the red downward triangles representing G^{WS} with $\beta = 0$, i.e., k-ring graphs; blue upward triangles are G^{WS} with $\beta = 1$, i.e., random graphs. (4b) shows how changing α causes disturbances in the Fisher Information Measure, when evaluating the Barabási-Albert model with non-linear PA.

Figure 6 shows that the results for the Fitness model fall into a region similar to BA, but with some networks closer to Random rather than Scale-Free.

Figure 7 presents the same behavior we saw initially, although now the values of FIM are a bit lower, as it happened to BA and Random networks. The transition between Random and Hub-and-Spoke still exists, and now it is even clearer: there are no longer oscillation for Hub-and-Spoke.

It is interesting to assess the configuration model; now, the networks built using a power-law with $\gamma \in [2, 3]$ represent a small area of the Shannon-Fisher plane (Fig. 8c), where the allegedly Scale-Free networks present themselves. This behavior provides interesting an interesting perspective when evaluating real networks.

Figure 9 shows that the results changed significantly in comparison with the first version without matrix reordering. We can now identify three separated clusters and two networks (6 and 14) a bit further from the others.

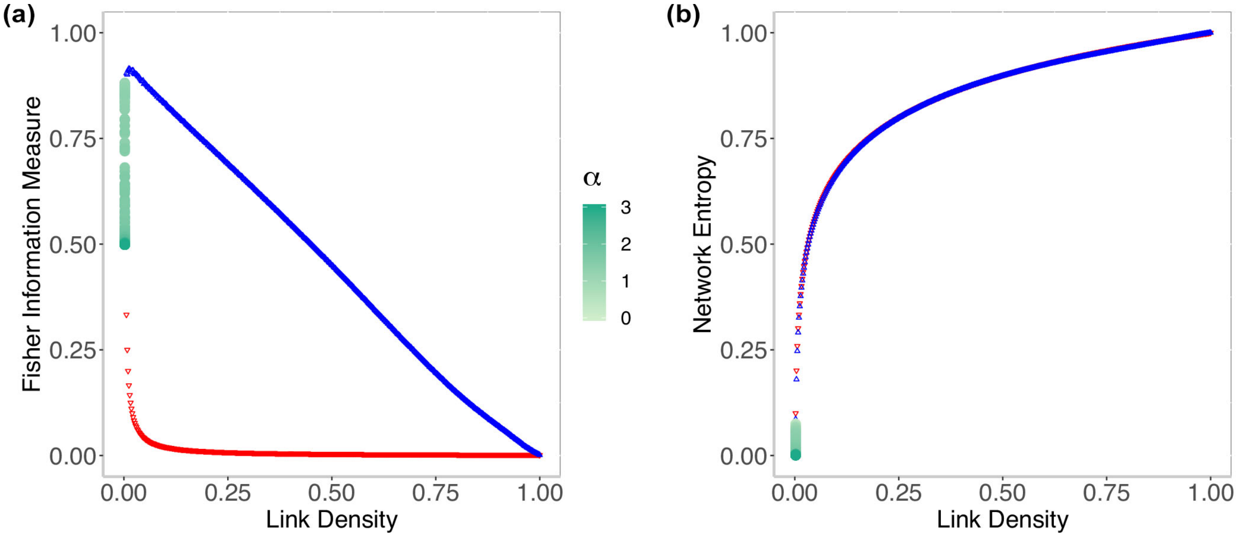

Figure 5: (5a) Relationship between link density and Fisher Information Measure for Barabási-Albert networks using a non-linear preferential attachment; the gradient indicates how the preferential attachment exponent α changes. (5b) Relationship between the Network Entropy and link density, where $\xi = 0.002$ for any α. To help the visualization of the region where Barabási-Albert networks stand in relation to the other synthetic networks, red downward triangles represent G^{WS} with $\beta = 0$, i.e., k-ring graphs; blue upward triangles are G^{WS} with $\beta = 1$, i.e., random graphs.

4 Conclusions

Most of the initial conclusions hold, but now we have stronger evidences. The ordering of these networks strengthens the validity of the Shannon-Fisher plane, considering the given synthetic networks. The Network Fisher Information Measure reveals information about the network when given a good representation to work with but, when evaluated alone, it becomes hard to draw strong conclusions. When the number of connections in the network grows, the adjacency matrix saturates, and FIM starts losing information about the network structure; the lower the values of FIM are, the more organized we expect the system's representation to be.

Although real networks analysis requires more caution, as the results in comparison from the initial ordering to an optimal order changed a bit, our contribution remains. Although there are differences in our numerical results (in between the natural ordering and the optimal one); the discussion is still valid.

For these reasons, we write to you, Chief Editor of Scientific Reports, to ask your advice on how to proceed with this issue. Our metric, as it is defined in our paper, may be affected by reordering the node labels, but the results are consistent. Adding the reordering into the discussion enhances our contribution. Reading through the politics of Scientific Reports, we believe there is room for a correction to alert the readers of the need of using the improved metric.

Finally, a whole new discussion can be conducted considering different matrix reordering techniques and their impacts on the FIM. Nevertheless, we believe that our work is still relevant.

References

Ziv Bar-Joseph, David K Gifford, and Tommi S Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1):S22–S29, 2001.

Michael Behrisch, Benjamin Bach, Nathalie Henry Riche, Tobias Schreck, and Jean-Daniel Fekete. Matrix

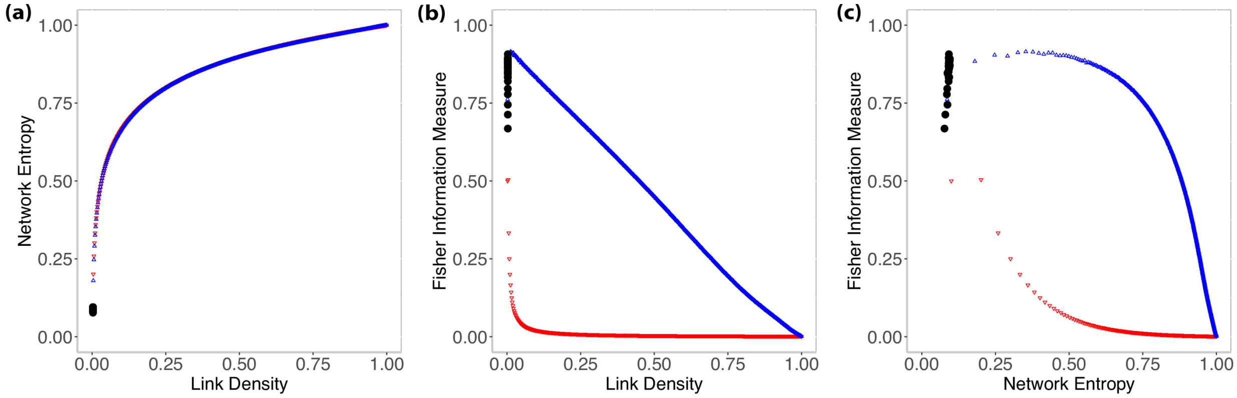

Figure 6: Relationship between Shannon Entropy and Fisher Information Measure with link density (**(a)** and **(b)**), and between Fisher Information Measure and Shannon Entropy (**c**) for Biaconi-Barabási (Fitness model). Black points indicate the thirty networks with $N = 1000$ generated using a uniform distribution for the fitness scores of each node.

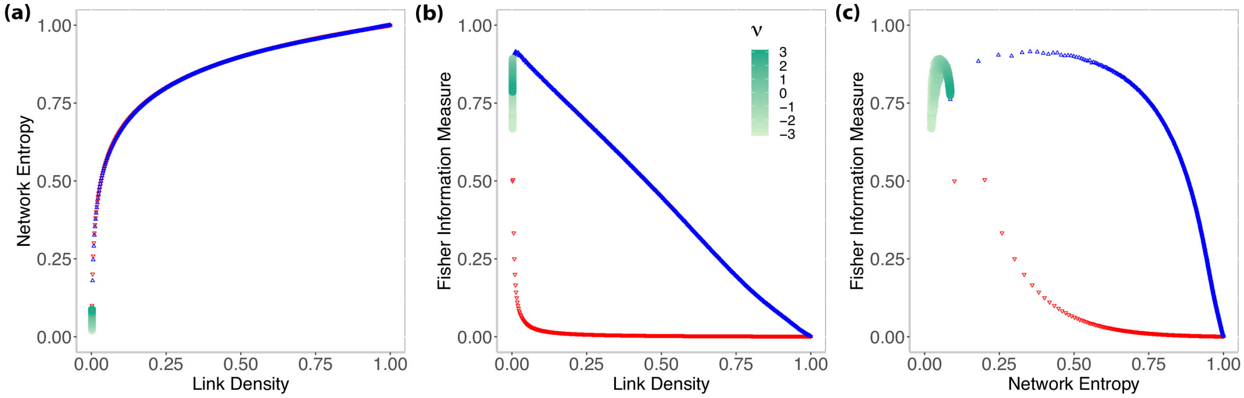

Figure 7: Relationship between Shannon Entropy and Fisher Information Measure with link density (**(a)** and **(b)**), and between Fisher Information Measure and Shannon Entropy **(c)** for the Aging model. The gradient indicates the aging exponent $\nu \in [-3, 3]$ and how its growth controls the network scaling regimes.

reordering methods for table and network visualization. In *Computer Graphics Forum*, volume 35, pages 693–716. Wiley Online Library, 2016.

Ulrik Brandes. Optimal leaf ordering of complete binary trees. *Journal of Discrete Algorithms*, 5(3):546–552, 2007.

Michael Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56 (4):316–324, 2002.

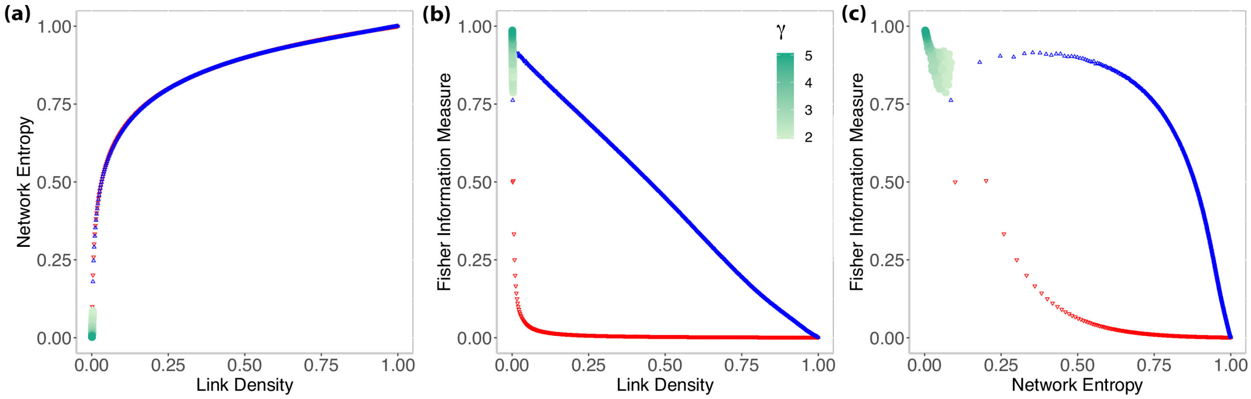

Figure 8: Relationship of Shannon Entropy and Fisher Information Measure with link density **(a,b)**, and between Fisher Information Measure and Shannon Entropy **(c)** for the configuration model with a degree distribution following a pure power-law $P(k) \sim k^{-\gamma}$. The gradient indicates the degree exponent $\gamma \in [2, 5]$ and how it controls the network scaling regimes.

Table 1: Real networks and their descriptors.

ID	Network	N	$\langle k \rangle$	ξ	L	C^{Δ}	S^{Δ}	γ	p-value	$\mathcal{H}$	$\mathcal{F}$
1	Email Network	1133	9.622	0.009	3.606	0.166	17.146	6.775	1.000	0.253	0.868
2	Adolescent Health	2539	8.236	0.003	4.559	0.142	38.118	8.244	0.996	0.248	0.858
3	Arxiv AstroPh	18772	21.101	0.001	4.194	0.318	239.693	4.496	0.980	0.231	0.900
4	NetScience Collaborations	1461	3.754	0.003	5.823	0.693	271.946	3.607	0.401	0.141	0.659
5	Science Collaborations	23133	8.078	0.000	5.352	0.264	737.276	3.426	0.310	0.160	0.915
6	Slucene	2956	7.336	0.002	4.499	0.057	22.710	2.187	0.896	0.174	0.704
7	AS Caida	16301	4.043	0.000	3.771	0.008	66.782	2.124	1.000	0.070	0.986
8	Power Grid	4941	2.669	0.001	18.989	0.103	92.461	7.629	1.000	0.094	0.840
9	Amazon pages	2879	2.700	0.001	3.433	0.023	56.123	3.257	0.000	0.016	0.831
10	Roget's Thesaurus	1010	7.224	0.007	4.075	0.134	17.770	6.246	0.919	0.250	0.885
11	Autobahn	1168	4.257	0.002	19.419	0.003	0.745	7.050	0.000	0.099	0.709
12	Protein Interactions	2018	2.681	0.001	5.611	0.024	23.907	2.782	1.000	0.077	0.871
13	Drosophila Medulla 1	1781	10.007	0.006	2.911	0.069	14.828	3.957	0.986	0.198	0.892
14	Mouse Retina 1	1076	168.794	0.157	1.861	0.400	2.526	2.312	0.000	0.693	0.443

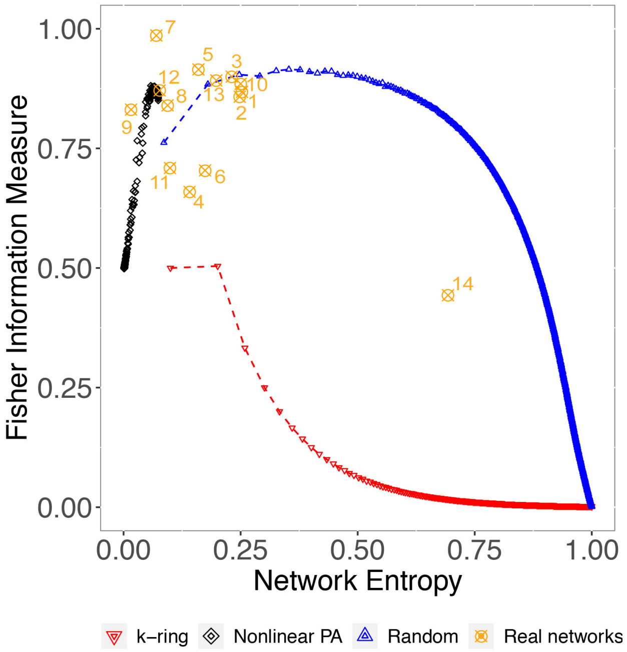

Figure 9: Relationship between Fisher Information Measure and Shannon Entropy for the real world networks. Blue upward triangles represent the ER graphs; the dashed-blue line indicates the upper limit of the small-world region delimited by graphs G^{WS} with $\beta = 1$; the downward red triangles represent k-ring graphs, as the dashed-red line indicates a "rough" lower limit for the small-world region.