

Research Article

Chengcheng Shao, Pengshuai Cui, Peng Xun, Yuxing Peng, and Xinwen Jiang*

Rank correlation between centrality metrics in complex networks: an empirical study

<https://doi.org/10.1515/phys-2018-0122>

Received Sep 20, 2018; accepted Dec 07, 2018

Abstract: Centrality is widely used to measure which nodes are important in a network. In recent decades, numerous metrics have been proposed with varying computation complexity. To test the idea that approximating a high-complexity metric by a low-complexity metric, researchers have studied the correlation between them. However, these works are based on Pearson correlation which is sensitive to the data distribution. Intuitively, a centrality metric is a ranking of nodes (or edges). It would be more reasonable to use rank correlation to do the measurement. In this paper, we use degree, a low-complexity metric, as the base to approximate three other metrics: closeness, betweenness, and eigenvector. We first demonstrate that rank correlation performs better than the Pearson one in scale-free networks. Then we study the correlation between centrality metrics in real networks, and find that the betweenness occupies the highest coefficient, closeness is at the middle level, and eigenvector fluctuates dramatically. At last, we evaluate the performance of using top degree nodes to approximate three other metrics in the real networks. We find that the intersection ratio of betweenness is the highest, and closeness and eigenvector follows; most often, the largest degree nodes could approximate largest betweenness and closeness nodes, but not the largest eigenvector nodes.

Keywords: network centrality, centrality measures, rank correlation, complex networks

PACS: 02.50.Sk, 05.10.-a

***Corresponding Author: Xinwen Jiang:** MOE Key Laboratory of Intelligent Computing and Information Processing, Xiangtan University, Xiangtan, 411105, Hunan, China; Email: jiangxinwen@yeah.net
Chengcheng Shao: College of Computer, National University of Defense Technology, Deya Road No.109, Kaifu District, 410073, Changsha, Hunan, China

Pengshuai Cui: College of Computer, National University of Defense Technology, Deya Road No.109, Kaifu District, 410073, Changsha, Hunan, China

1 Introduction

Centrality is used to address the research question “which are the most important or central nodes or vertices in a network”. In the study of networks, sociologists have perhaps done the longest work to establish the best traditions in both quantitative and empirical ways [1]. The idea of centrality was, of course, first introduced by a socialist, Bavelas, in the 1950s, when he tried to characterize the human communication in small groups of people [2, 3]. Since then, centrality has been used to investigate the adoption of innovation [4], robustness of networks [5, 6], the engagement in higher education [7] and more. Whilst a number of metrics have been proposed, the definition of centrality itself is still not beyond the general descriptors such as node prominence or structural importance [8]. This ambiguous definition of centrality leads the interpretations of centrality metrics varied, including autonomy, control, risk, exposure, influence, independence, power and so on.

To answer the question “what is centrality”, Freeman [9] first reviewed a number of published metrics and reduced them into three basic concepts with the canonical formulation. These three canonical metrics were degree, closeness, and betweenness that are still widely used today. Borgatti [8, 10] provided a more comprehensive answer in both network flow and graph structure perspectives. In the network flow context, centrality is centered on the outcomes of the nodes through which the network traffic goes. On the other hand, the graph-theoretic perspective is centered on how centrality metrics are calculated.

Today, the research of network centrality falls into two main fields: application of centrality concept into new realms and enhancement of the computing performance for specified metrics. In the application field, as where it was first introduced, the concept of centrality is broadly

Peng Xun: College of Computer, National University of Defense Technology, Deya Road No.109, Kaifu District, 410073, Changsha, Hunan, China

Yuxing Peng: National Key Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Deya Road No.109, Kaifu District, 410073, Changsha, Hunan, China

used in social networks. Besides the early investigation of influence, power, control in the context of organization or small group networks [4, 11–13], today centrality is still an important tool to quantify social impact and to identify the most influential people in large online social networks. For instance, Weng et al. [14] took advantages of topical interests in Twitter to identify individual social influence; Yafang et al. proposed a parameter-free community detection method based on centrality. Beyond the social networks, the centrality is also introduced into technological networks. For example, Hypertext Induced Topic Search (HITS) is a very famous web page ranking method proposed by Jon Kleinberg [15]. Its sibling PageRank [16] has been used by Google search engine to index web pages for years. In biological networks, the concept of centrality is widely utilized to identify the most influential person in epidemic networks [17], to validate the drug target in protein-protein interaction networks [18], and to distinguish the most important nodes in weighted functional brain networks [18].

Though we have a flourish of centrality applications, the enhancement of computation of many difficult metrics moves slowly. For example, computing betweenness of large networks is still a hard task. Even though there are no big breaks in mathematical theories, researchers have made significant progress in other aspects. As observed that large networks are often sparse networks, Brandes [19] proposed a betweenness computing algorithm for large sparse networks, which has been the de facto algorithm that widely used in many network analysis tools. And there are many following works, e.g. [20]. In addition, many parallel computing techniques [21–23] have been brought into the graph algorithms, which make them much faster.

In the early stage of network analysis, it is often hard to determine which centrality metric is the most suitable. Moreover, it is never an easy task to calculate metrics with high-complexity, e.g. betweenness for large networks. In computation perspective [8], some metrics could be aligned into one family. For example, when constraining the counts of walks from one to infinite, we could get different metrics such as Freeman degree, Bonacich eigenvector, Katz status and others, indicating a possible correlation among these metrics. This possible correlation motivates us to think is it feasible to approximate a high-complexity metric by a low-complexity one? In fact, some studies [24, 25] showed that metrics like degree and betweenness are indeed highly correlated. However, these works use Pearson correlation that is sensitive to the data distribution. As we know that Pearson correlation is a parametric measurement that is restricted by several as-

sumptions, e.g. the bivariate should be approximately normally distributed; while the assumptions for rank correlation, e.g. Spearman correlation, are quite loose. Considering many real networks are scale-free networks, we believe that rank correlation would be a better choice to use.

In this paper, we study the correlation between centrality metrics, especially the rank correlation. We first demonstrate that rank correlation coefficients perform better than the Pearson's in scale-free networks. Then we investigate the correlation between centrality metrics in real networks. We find that betweenness occupies the highest coefficient, closeness is in the middle level, and eigenvector fluctuates dramatically. And at last, to evaluate the performance of approximating high-complexity metrics by degree, we conduct two experiments in real networks. We find that the intersection ratio of betweenness is the highest, followed by closeness and eigenvector; most often, the largest degree nodes could approximate the largest betweenness and closeness nodes, but not the eigenvector ones.

The remaining part of the paper is organized as follows. We first describe the data and methods in section 2. Then in section 3, we show the results for both scale-free networks and real networks. And finally, we make the conclusion of our analysis in section 4.

2 Data and methods

Networks are always represented as a graph $G(V, E)$, where V is the vertex set and E is the edge set. The total number of vertices is denoted as N and the total number of edges is denoted as M . In this section, we would first review the four widely used centrality metrics in a manner of time computation complexity: degree, closeness, betweenness, and eigenvector; then we describe the difference between Pearson and rank correlation; and finally we introduce how we prepare the data.

2.1 Definition of centrality metrics

(1) Degree D_c : The degree centrality is the simplest metric, which is defined as the number of links incident upon a node (i.e. the number of ties that a node has). The degree of vertex i , $D_c(i)$, can be formulated as $C_d(i) = \sum_j a_{ij}$, where a_{ij} is a element of its adjacency matrix A . The normalized formulation is $C'_d(i) = C_d(i)/(N - 1)$. The computing of degree of all nodes takes $O(|V|^2)$ in a dense adjacency matrix

representation. However, for sparse large real networks, the computation complexity could be reduced to $O(|E|)$.

(2) Closeness C_c : The concept of closeness metric comes from the term distance that is how close the distance between node i and other nodes. In graph theory, the distance d_{ij} is measured by the shortest path length between i and j . The closeness $C_c(i)$ is defined as $C_c(i) = 1/\sum_j d_{ij}$ and the normalized one is defined as $C'_c(i) = (N-1)/\sum_j d_{ij}$. The essential to compute closeness is to solve the shortest paths problem, which has been studied for decades. Depending on the types of the graph (directed, undirected, weighted, unweighted), the computation complexity of the shortest path problem varies from $|O(E)|$ to $O(V^2)$.

(3) Betweenness B_c : The betweenness metric is defined as $C_b(i) = \sum_{j \neq i \neq k \in V} \sigma_{jk}(i)/\sigma_{jk}$, where σ_{jk} is the total number of shortest paths from j to k and $\sigma_{jk}(i)$ is the number of those paths that pass through i . The normalized betweenness is defined by dividing the number of pairs of vertices excluding i , which is $(n-1)(n-2)$ for directed graphs and $(n-1)(n-2)/2$ for undirected graphs. The state-of-the-art algorithm for unweighted graphs is Brandes' algorithm, of which the computation complexity is $O(|V||E|)$.

(4) Eigenvector E_c : The concept of eigenvector metric is that the importance of a vertex not only depends on the number of its neighbors but also relies on their importance. If the importance of vertex i is noted as x_i , this idea could be defined as matrix formulation $\mathbf{x} = c\mathbf{A}\mathbf{x}$. This equation means that $\mathbf{x}$ is the eigenvector of matrix $\mathbf{A}$ with corresponding eigenvalue $\lambda = c^{-1}$. Though there could be multiple eigenvalues for a certain eigenvector, however, with the restriction that $\mathbf{A}$ is nonnegative, only the largest magnitude eigenvalue λ_1 makes the corresponding eigenvector meaningful as a metric of centrality. Other metrics like HITS and PageRank also borrowed the idea from eigenvector. When calculating E_c , we usually take the power iteration method, which runs at the time complexity $O(|V|^2)$ for each iteration and follows a linear convergence.

2.2 Correlation coefficients

Correlation is a bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. If the coefficient goes to ± 1 , then it is said they are perfectly associated. If the coefficient tends to 0, there is almost no relationship between them.

Usually, there are two types of correlation analysis, shown in Table 1. While the parametric statistical procedures are more powerful as they use more underlying in-

Table 1: Parametric vs. non-parametric correlations¹

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	mean	median
Correlation test	Pearson	Spearman, Kendall

formation from the normal distribution, we believe rank correlation is more suitable for the centrality metrics analysis, because the underlying distribution is not necessary to be normal.

Here we present the measurement of assortativity as a good example to support that rank correlation is better than the Pearson one for this particular network parameter. The assortativity is a way of measuring mixing patterns that refer to the extent for nodes to connect to other similar or different nodes. We often start to examine the assortativity in terms of degree. That is to say, that degree assortativity is used to answer questions like do the high-degree vertices in a network associate preferentially with other high-degree vertices or low-degree one. In social networks, it is also called homophily to explain phenomena like rich clubs. The assortativity coefficient was first introduced by Newman [26], which is, in fact, a Pearson correlation coefficient. However, Litvak et al. [27] argued that this measurement suffers a problem: for disassortative networks, with the increase of network size, the coefficient decreases significantly. Furthermore, they explained this problem mathematically, proposed a new method by using rank correlation measures, e.g. Spearman's rho, and their experiments proved that Spearman's rho performs better.

In this paper, we would examine the correlation between centrality metrics by Spearman and Kendall coefficients, which shows some different results from the Pearson one.

¹ Please visit http://changingminds.org/explanations/research/analysis/parametric_non-parametric.htm

2.3 Scale-free networks

If the degree distribution of a network follows a power-law distribution (or at least asymptotically), we call it a scale-free network. Many real-world networks have been observed that they have power-law degree distributions. Preferential attachment and the fitness model have been proposed as mechanisms to explain conjectured power-law degree distributions in real networks. The BA model [28] is an algorithm that uses the preferential attachment. There are two key general concepts in the BA model.

1. Growth: starting with a small number of vertices (m_0), at every timestep we add a new vertex with Δm ($\Delta m \leq m_0$) edges that link the new vertex to Δm different vertices already present in the system.
2. Preferential attachment: the probability $P(k_i)$ that a new vertex will be connected to vertex i depends on the connectivity k_i of that vertex, such that $P(k_i) = \frac{k_i}{\sum_j k_j}$.

After t timesteps the model leads to a random network with $n = t + m_0$ vertices and mt edges.

In this paper, we use the BA model to generate scale-free networks. In each experiment, we would give the parameter values we used to generate the scale-free network instance. In addition, we also use the *Configuration* model to generate random networks with power-law distribution. In the Configuration model, we use the generated BA network as input and rewire the edges while keeping the degree sequences.

2.4 Real networks

Thanks to the Internet, we have a chance to get rich network datasets from other researchers. The datasets in this paper are downloaded from four sites: personal site of Newman, SNAP, Pajek, and CCNR (please see Appendix Table A1). We know that network could be divided as undirected and directed. The interpretation and computation of network centrality between undirected and directed are a little different. To make it simple and comparable, we treat all of our networks as undirected.

In [29], Newman adopted a loose classification that divided real networks into three categories: social, technological and biological. In [30], he added a new category, information networks, consisting of knowledge networks like citation network and World Wide Web (WWW) network. Appendix Table A1 shows the real networks we analyzed in this paper. Please note that we adopt the former

classification and do not distinguish the technological and informational networks.

3 Results

3.1 Correlation between centrality metrics in scale-free networks

In section 2.2, we compared Pearson correlation with rank correlation (e.g. Spearman). Here we would conduct an experiment in scale-free networks to demonstrate that rank correlation performs better in such a context. The scale-free networks in the experiment are generated by BA model and Configuration model. We choose four different parameter settings for BA model: $\Delta m = 1, 2, 5, 10$ (see section 2.3, we set $m_0 = \Delta m$). The Configuration model uses BA network instances as input and rewires edges randomly while keeping the degree sequences. Thus we have four groups of different network settings. For each group, we compare the three correlation coefficients in three pairs of centrality metrics. To make our results robust, the size of the network grows from 2^{10} to 2^{20} , and at each exponent integer, we generate 20 samples for each network settings.

The results are shown in Figure 1 and Figure 2. For example, in Figure 1, the vertical panel represents four groups of different network settings and they are: BA model with $\Delta m = 1$, Configuration model inputting from BA model with $\Delta m = 1$, BA model with $\Delta m = 2$ and Configuration model inputting from BA model with $\Delta m = 2$. The horizontal panel represents three correlations of different pairs of metrics: the degree with betweenness ($\text{corr}(D, B)$), closeness ($\text{corr}(D, C)$) and eigenvector ($\text{corr}(D, E)$). In each subfigure, we show the results of three correlation coefficient with the growth of network sizes. Each point represents the mean values of the coefficient from the 20 samples. The error bar represents 95% confidence interval. To better refer to these subfigures, we locate them by the panel coordinate, e.g. panel (0, 0) represents the first subfigure in the top left directions (the first row and first column in the panel).

From these two figures, we have the following observations.

1. For each subfigure, compared with Pearson coefficients, rank correlation coefficients drop very little along the growing of networks. In some cases, e.g. Panel (1, 2) and (2, 2), r goes down dramatically.
2. For almost all cases, the deviations of rank correlation coefficients are almost invisible, while we could often observe a large deviation for r , especially in the

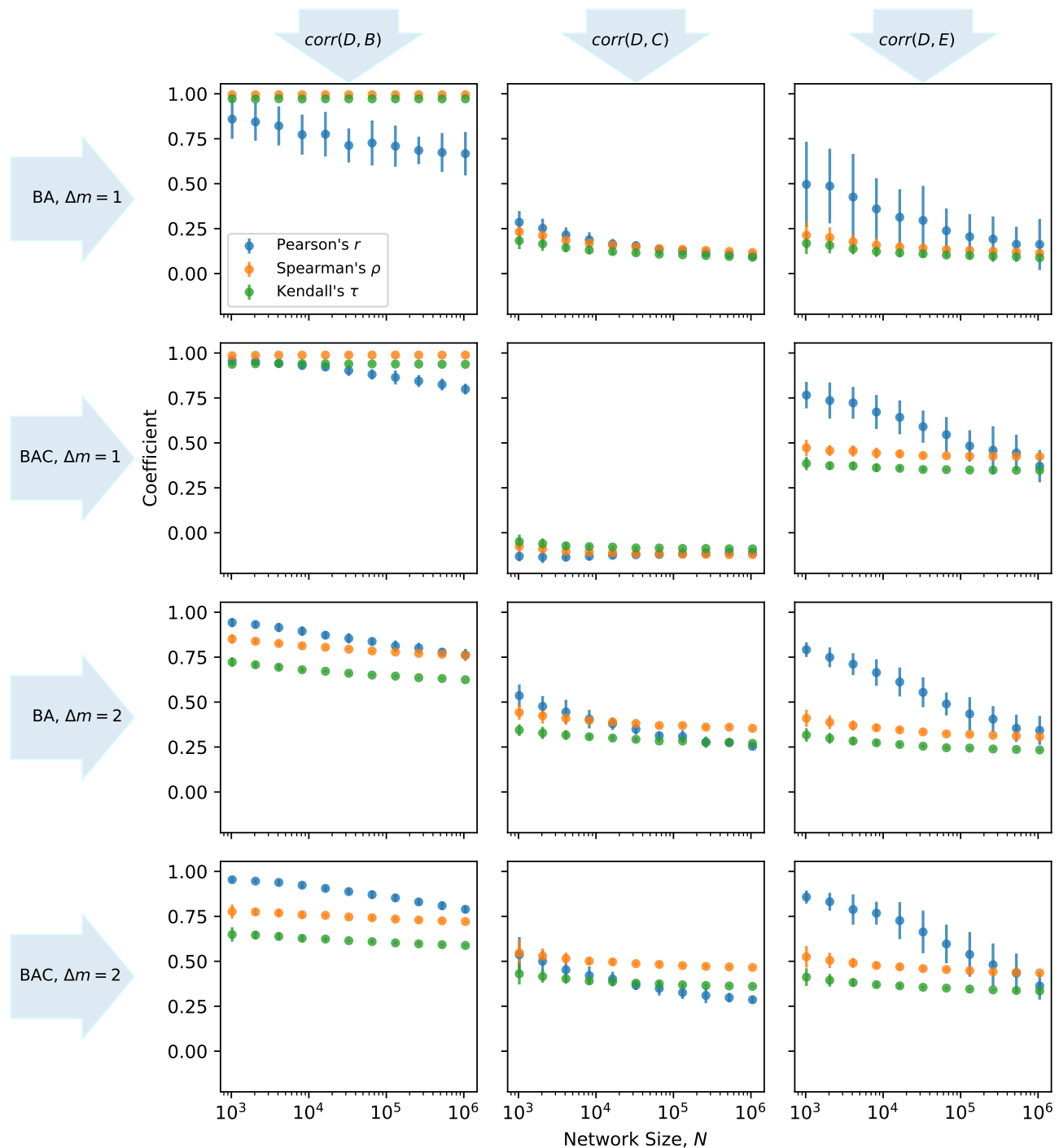


Figure 1: Correlation between centrality metrics in scale-free networks. Three correlation coefficients are used: one is Pearson's r (blue color); two are rank correlation coefficients, Spearman's ρ (orange) and Kendall's τ (green). Three pairs of centrality metrics are measured: the degree with betweenness ($\text{corr}(D, B)$), closeness ($\text{corr}(D, C)$) and eigenvector ($\text{corr}(D, E)$). We use BA model and Configuration model to generate scale-free networks. The BA model uses two parameter settings: $\Delta m = 1$ and $\Delta m = 2$. And the configuration model uses the generated BA networks as input and rewires the edges while keeping the degree sequence. In total, we have four groups of network settings, shown in vertical panel; three pairs of centrality metrics, show in horizontal panel. In each subfigure, the network size grows from 2^{10} to 2^{20} and we generate 20 network instances for network size at each exponent integer

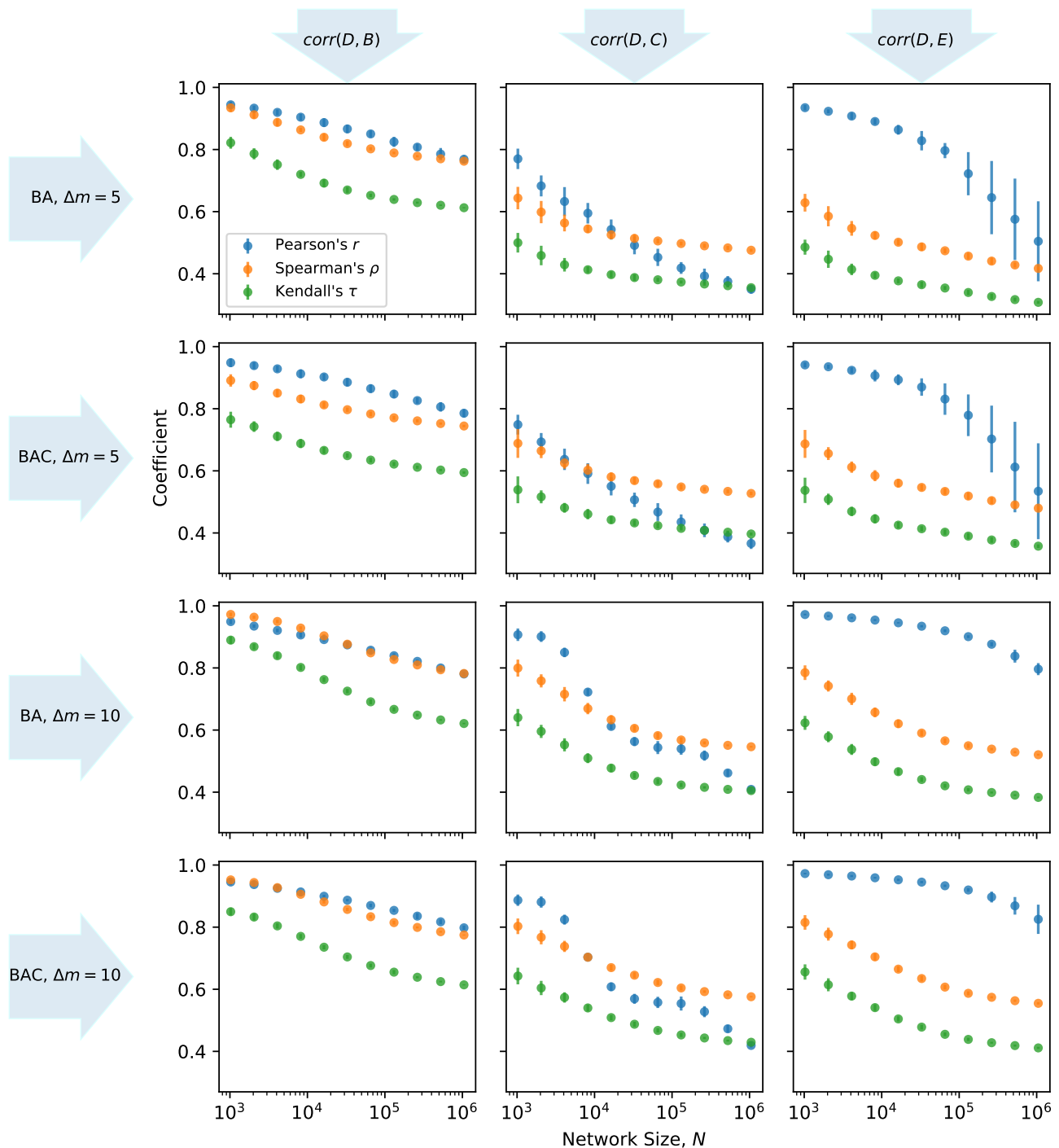


Figure 2: Correlation between centrality metrics in scale-free networks. The BA model uses two parameter settings: $\Delta m = 5$ and $\Delta m = 10$. Please refer Figure 2 for more information

$\text{corr}(D, E)$ panel column and (0, 0). In fact, in some cases, e.g. Panel (0, 3) and (1, 0), we could see small deviations of rank correlation coefficients when the size of networks are rather small. However, when networks grow big enough, e.g. more than 10^5 , no

deviations of rank correlation coefficients could not be visible anymore.

3. From the view of each panel column, most of-
ten $\text{corr}(D, B)$ is the highest (at least 0.75 in
terms of Spearman's ρ), followed by $\text{corr}(D, E)$ and

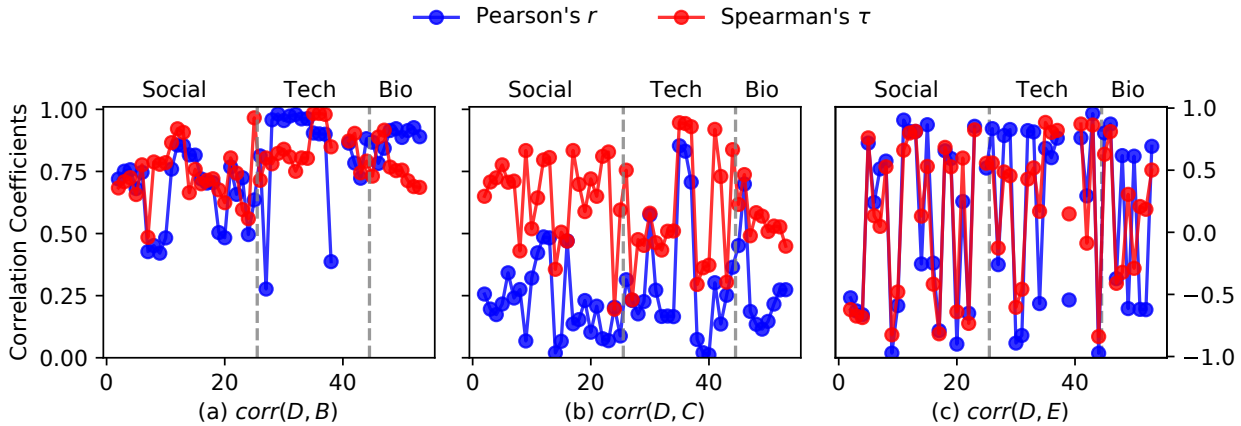


Figure 3: Correlation between centrality metrics in real networks. 52 real networks are used in this experiments, and are roughly divided into three categories: social, technological and biological. Two correlation coefficients are used, the Pearson's r (blue) and Spearman's ρ (red). Three pairs of centrality metrics are measured: the degree with betweenness ((a) $\text{corr}(D, B)$), closeness ((b) $\text{corr}(D, C)$), and eigenvector ((c) $\text{corr}(D, E)$)

$\text{corr}(D, C)$ (both are around or less than 0.50 in terms of ρ). This suggests that the degree metric is likely to have the same ranking order as the betweenness metric.

4. Among all these subfigures, we find no distribution difference between the two rank correlation coefficients (ρ and τ), except that ρ is always larger than the τ .

These results demonstrate that rank correlation coefficients perform better than the Pearson's in scale-free networks. The high correlation between degree and betweenness makes it possible to approximate the betweenness metric by using degree metric.

3.2 Correlation between centrality metrics in real networks

There are 52 different real networks, categorized as social, technological and biological networks. Figure 3 shows the correlation coefficients for these real networks (see Appendix Table A2). Again, we measure three pairs of centrality metrics. Since there is no much difference of the distribution between ρ and τ , to avoid the figure is too dense to read, we show the line of ρ only. From the figure, we have the following observations.

1. In general, the coefficients of $\text{corr}(D, B)$ is the highest, which is consistent with the results in scale-free networks. However, the coefficient of $\text{corr}(D, E)$ varies so much that it almost reaches the two ends (1 and -1) of the range. This suggests that our met-

ric approximation idea could be applied between degree and betweenness, while it could be infeasible for degree and eigenvector.

2. In (b), ρ (mean value is 0.61) is much higher than r (mean value is 0.26). This implies that Spearman correlation is capable of capturing the underlying ranking correlation between degree and closeness.
3. Among different network categories, we do not find significant signals.

These results suggest that real networks are much more complex than the model generated ones. Though we find that the coefficients of $\text{corr}(D, B)$ are the highest, which is consistent with the model generated ones, the $\text{corr}(D, E)$ in real networks varies so much that it is infeasible to approximate eigenvector by degree metric. In addition, in $\text{corr}(D, C)$, we find that ρ is much higher than r , which could be a signal indicative of a better performance of rank correlation.

3.3 Approximate high-complexity metrics by degree

The results of real networks show a high correlation between degree and betweenness, suggesting that we could use degree as the preliminary metric to approximate betweenness. But we never know whether this approximation is really good in applications. Again, we do find that ρ is much higher than r in $\text{corr}(D, C)$, but we don't have the evidence to prove ρ does better than r in real networks. Furthermore, the correlation coefficients evaluate the sim-

ilarity between metrics in an overall way. However, we are often only interested in the most important nodes, which are the top ranking nodes of a certain metric. Here, we use the top ranking nodes to evaluate the performance of approximating high-complexity metrics by degree.

To conduct the evaluation, we define two metrics. One is to measure how many nodes with the highest degree are also the top ranking nodes in another centrality metric. The other one is to measure how bad it could be if we treat the largest degree nodes as another centrality metric. These two metrics are defined for selected nodes, either the top ranking nodes of a specified centrality or the largest degree nodes. We would evaluate the two metrics for all instances of real networks.

1. We denote the top ranking nodes in a centrality metric c as V_c^{top} , where c is one of degree (D), betweenness(B), closeness (C) and eigenvector (E). The first experiment question is that to approximate the same amount of V_c^{top} from other three metrics, how well does V_D^{top} perform. To measure it, we define the ratio of intersection as:

$$r_{\cap}(c) = |V_D^{top} \cap V_c^{top}| / |V_D^{top}|. \quad (1)$$

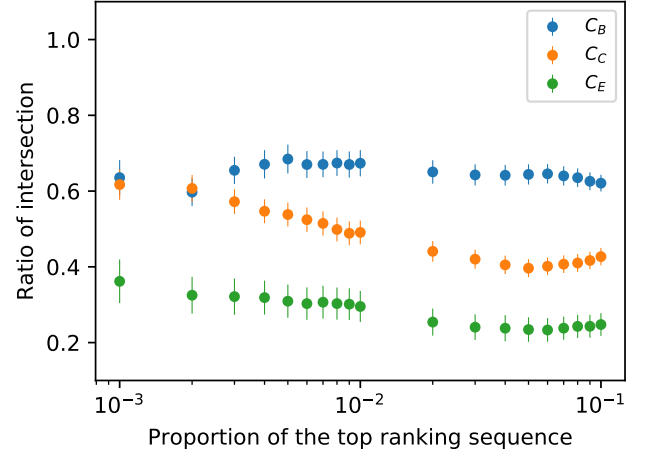
2. We denote the vertices with the largest degree as Δ . Our second experiment question is that how bad could Δ performs in the other three metrics. Supposing the index of a ranking metric is from 0 to $|V| - 1$, we define the ranking ratio of nodes v in a centrality metric c as

$$r_{rank}(v, c) = idx(v, c) / (|V| - 1), \quad (2)$$

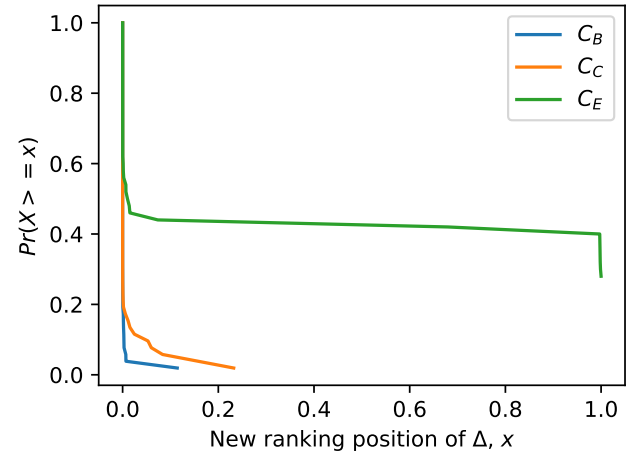
where the function idx returns the ranking index of node v in the centrality metric c . For nodes with same metric values, the idx returns the mean ranking index.

The results of these two experiments are shown in Figure 4. Figure 4 (a) shows the distribution of $r_{\cap}$ for these three centrality metrics in real networks. Each point represents the mean value of $r_{\cap}$ and the error bar indicates 95% confidence interval. From Figure 4 (a), firstly, we could see that $r_{\cap}(B)$ is the highest (more than 0.6) in these three metrics. This is not a surprising result, considering the high correlation coefficients of $corr(D, B)$ in real networks. Secondly, we find that $r_{\cap}(B)$ is also very reasonable (around 0.5). Remember that we have $\bar{\rho} = 0.61$ and $\bar{r} = 0.26$, here we believe this result provides evidence that Spearman correlation performs better than the Pearson one in real networks.

Figure 4 (b) shows the Complementary Cumulative Distribution Function (CCDF) of r_{rank} . The horizontal axis



(a) Distribution of $r_{\cap}$



(b) CCDF of r_{rank}

Figure 4: The performance of approximating high-complexity metrics by degree in real networks. (a) Distribution of $r_{\cap}$. $r_{\cap}$ is the ratio of intersected vertices between top ranking degree and another metric, see equation 2. The proportion of the top ranking sequence is from 10^{-3} to 10^{-1} . Each point represents the mean value and the error bar represents 95% confidence interval. (b) Complementary Cumulative Distribution Function (CCDF) of r_{rank} . The horizontal axis is the ranking position of vertices with largest degree (Δ) in another metric, denoted as x . The vertical axis is the probability of real networks, whose r_{rank} is equal or larger than x , denoted as $Pr(X \geq x)$.

is the ranking position of vertices with the largest degree (Δ) in other metrics, denoted as x . The vertical axis is the probability of real networks, whose r_{rank} is equal or larger than x , denoted as $Pr(X \geq x)$. In this CCDF figure, we find that Δ performs perfectly in betweenness and centrality metrics. In these real networks, more than 95% of them, their Δ would also be ranked in the top 0.1% betweenness metric; 80% of them could archive that for closeness metric; the worst case of closeness metric is that Δ is ranked around the top 2.5%. For eigenvector metric, the perfor-

mance is not good enough, even though in about 50% of real networks, it could perform as well as betweenness, Δ becomes the smallest eigenvector vertices in the worst cases. When investigating these worst cases, we could see that the correlation coefficients tend to be -1 .

In one word, there are many overlapping nodes between top ranking degree set and another top ranking centrality metric set. And the largest degree nodes are also the top first or near the top first nodes in other centrality metrics for most real network instances.

4 Conclusion

Over the decades, a number of centrality metrics have been proposed with different computation complexity. Researchers have studied the correlation between them. However, the Pearson correlation is commonly used in these studies, which we argue that it is not suitable for scale-free networks. At the same time, many real networks are reported as scale-free networks. Therefore, in this paper, we study the rank correlation between centrality metrics. We first demonstrate that rank correlation performs better than the Pearson one in scale-free networks. Then we study the correlation between centrality metrics in real networks. And lastly, we evaluate the performance of using top degree nodes to approximate three other metrics in the real networks.

Actually, we did apply the idea of this paper to other works. For example, in our previous work [31], when we tried to find out who are the most important accounts in the spread of misinformation. We first used the k -core metric to narrow the network to the most density part – the main core network part. Then we can use different metrics to measure the importance of nodes from different aspects, e.g. retweeted most often, retweeting most often and so on.

Of course, our analysis has unavoidable limitations. Firstly, we only focused on four types of centrality metrics, even though they are the most popular ones. There are many other metrics that are also widely used, such as k -core, PageRank and so on. The future analysis could take more metrics into consideration.

Secondly, to demonstrate that rank correlation is better than Pearson's, we conducted experiments on scale-free network instances. Neither all models are tested nor different parameters are tested. Moreover, for real networks, we did not test their scale-free property. In fact, it is not an easy task to check the scale-free property of the empirical data. Few real networks follow power-law for the whole set of nodes, thus the evaluation is often con-

ducted for nodes with degree larger than a specified value. Thanks to Clauset et al., who proposed a very good statistical framework to test the scale-free property in empirical data [32]. Furthermore, Broido et al. [33] presented the scale-free property results for nearly 1000 real networks. In our work, we acknowledged that only a portion of real networks are scale-free. But we would like to demonstrate that rank correlation is at least as good as the Pearson one and most often better than it. Theoretically, Pearson correlation assumes a normal distribution of two variables, whilst rank correlation is open for any distribution. Here we use scale-free network models to demonstrate this idea. It does not necessarily mean that the network should be scale-free. In fact, if the two centrality values do not follow the normal distribution, theoretically, rank correlation should be used to test their correlation. Thus we do not emphasize that real networks must be scale-free.

Finally, though we applied our analysis on many real networks, we still cannot say that our coverage is enough. Nevertheless, our analysis could provide some highlights in the study of network centrality.

Acknowledgement: C.S. was supported by the China Scholarship Council. X.J. was supported in part by the National Natural Science Foundation of China (No. 61272010). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This research was supported in part by Lilly Endowment Inc., through its support for the Indiana University Pervasive Technology Institute.

References

- [1] Newman M., *Networks: an introduction*. Oxford University Press, 2010.
- [2] Bavelas A., A mathematical model for group structures, *Applied anthropology*, vol. 7, no. 3, pp. 16–30, 1948.
- [3] Bavelas A., Communication patterns in task-oriented groups, *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.
- [4] Coleman J. S., Katz E., Menzel H., et al., *Medical innovation: A diffusion study*. Bobbs-Merrill Indianapolis, 1966.
- [5] Dong G., Gao J., Tian L., Du R., He Y., Percolation of partially interdependent networks under targeted attack, *Physical Review E*, vol. 85, no. 1, p. 016112, 2012.
- [6] Dong G., Gao J., Du R., Tian L., Stanley H. E., Havlin S., Robustness of network of networks under targeted attack, *Physical Review E*, vol. 87, no. 5, p. 052804, 2013.
- [7] Fitzgerald H. E., Bruns K., Sonka S. T., Furco A., Swanson L., The centrality of engagement in higher education, *Journal of Higher Education Outreach and Engagement*, vol. 20, no. 1, pp. 223–244, 2016.

- [8] Borgatti S. P., Everett M. G., A graph-theoretic perspective on centrality, *Social networks*, vol. 28, no. 4, pp. 466–484, 2006.
- [9] Freeman L. C., Centrality in social networks conceptual clarification, *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.
- [10] Borgatti S. P., Centrality and network flow, *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.
- [11] Laumann E. O., Pappi F. U., New directions in the study of community elites, *American Sociological Review*, pp. 212–230, 1973.
- [12] Granovetter M., Getting a job: a study of careers and contacts, 1995.
- [13] Burt R. S., Toward a structural theory of action, 1982.
- [14] Weng L., Menczer F., Topicality and social impact: Diverse messages but focused messengers, *arXiv preprint arXiv:1402.5443*, 2014.
- [15] Kleinberg J. M., Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.
- [16] Brin S., Page L., Reprint of: The anatomy of a large-scale hypertextual web search engine, *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.
- [17] Šikić M., Lancić A., Antulov-Fantulin N., Štefanić H., Epidemic centrality is there an underestimated epidemic impact of network peripheral nodes?, *The European Physical Journal B*, vol. 86, no. 10, pp. 1–13, 2013.
- [18] Kuhnert M.-T., Geier C., Elger C. E., Lehnertz K., Identifying important nodes in weighted functional brain networks: a comparison of different centrality approaches, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 2, p. 023142, 2012.
- [19] Brandes U., A faster algorithm for betweenness centrality*, *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.
- [20] Yang J., Chen Y., Fast computing betweenness centrality with virtual nodes on large sparse networks, *PloS one*, vol. 6, no. 7, p. e22557, 2011.
- [21] Low Y., Bickson D., Gonzalez J., Guestrin C., Kyrola A., Hellerstein J. M., Distributed graphlab: a framework for machine learning and data mining in the cloud, *Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.
- [22] Green O., Bader D. A., Faster betweenness centrality based on data structure experimentation, *Procedia Computer Science*, vol. 18, pp. 399–408, 2013.
- [23] Hounkaew C., Suzumura T., X10-based distributed and parallel betweenness centrality and its application to social analytics, in *High Performance Computing (HiPC)*, 2013 20th International Conference on, pp. 109–118, IEEE, 2013.
- [24] Li C., Li Q., Van Mieghem P., Stanley H. E., Wang H., Correlation between centrality metrics and their application to the opinion model, *The European Physical Journal B*, vol. 88, no. 3, pp. 1–13, 2015.
- [25] Lee C.-Y., Correlations among centrality measures in complex networks, *arXiv preprint physics/0605220*, 2006.
- [26] Newman M. E., Assortative mixing in networks, *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.
- [27] Litvak N., Van Der Hofstad R., Uncovering disassortativity in large scale-free networks, *Physical Review E*, vol. 87, no. 2, p. 022801, 2013.
- [28] Barabási A.-L., Albert R., Emergence of scaling in random networks, *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [29] Newman M. E., Mixing patterns in networks, *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.
- [30] Nej N., The structure and function of complex networks, *Siam Review*, vol. 45, no. 2, p. 167, 2003.
- [31] Shao C., Hui P.-M., Wang L., Jiang X., Flammini A., Menczer F., Ciampaglia G. L., Anatomy of an online misinformation network, *PloS one*, vol. 13, no. 4, p. e0196087, 2018.
- [32] Clauset A., Shalizi C. R., Newman M. E., Power-law distributions in empirical data, *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.
- [33] Broido A. D., Clauset A., Scale-free networks are rare, *arXiv preprint arXiv:1801.03400*, 2018.
- [34] Newman M. E., The structure of scientific collaboration networks, *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.
- [35] Newman M. E., Finding community structure in networks using the eigenvectors of matrices, *Physical review E*, vol. 74, no. 3, p. 036104, 2006.
- [36] Leskovec J., McAuley J. J., Learning to discover social circles in ego networks, in *Advances in neural information processing systems*, pp. 539–547, 2012.
- [37] Yang J., Leskovec J., Defining and evaluating network communities based on ground-truth, *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.
- [38] Leskovec J., Kleinberg J., Faloutsos C., Graph evolution: Densification and shrinking diameters, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 2, 2007.
- [39] Leskovec J., Kleinberg J., Faloutsos C., Graphs over time: densification laws, shrinking diameters and possible explanations, in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, ACM, 2005.
- [40] Jones B., Computational geometry database, february 2002, FTP/HTTP.
- [41] Gehrke J., Ginsparg P., Kleinberg J., Overview of the 2003 kdd cup, *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 149–151, 2003.
- [42] Leskovec J., Adamic L. A., Huberman B. A., The dynamics of viral marketing, *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.
- [43] Albert R., Jeong H., Barabási A.-L., Internet: Diameter of the world-wide web, *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.
- [44] Corman S. R., Kuhn T., McPhee R. D., Dooley K. J., Studying complex discursive systems., *Human communication research*, vol. 28, no. 2, pp. 157–206, 2002.
- [45] Langville A. N., Meyer C. D., A reordering for the pagerank problem, *SIAM Journal on Scientific Computing*, vol. 27, no. 6, pp. 2112–2120, 2006.
- [46] Kamvar S., Haveliwala T., Manning C., Golub G., Exploiting the block structure of the web for computing pagerank, *Stanford University Technical Report*, 2003.
- [47] Watts D. J., Strogatz S. H., Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [48] Ripeanu M., Foster I., Iamnitchi A., Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design, *arXiv preprint cs/0209028*, 2002.
- [49] Jeong H., Mason S.P., Barabási A.-L., Oltvai Z.N., Lethality and centrality in protein networks, *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
- [50] Jeong H., Tombor B., Albert R., Oltvai Z.N., Barabási A.-L., The large-scale organization of metabolic networks, *Nature*, vol.

407, no. 6804, pp. 651–654, 2000.

Appendix A: Real networks

In appendix 4, we show the data collection of 52 real networks. They are classified into three loose categories: social, technological and biological. Table A1 gives a summary of the datasets. Table A2 shows the results of real networks. We have 9 centrality correlations: 3 correlation methods combined with 3 centrality pairs. To make the table tight so that we can list them all, we use (D, B) to represent centrality correlation between degree and betweenness, which was noted as (C_D, C_B) before, and (D, C) for (C_D, C_C) , (D, E) for (C_D, C_E) .

In Table A2, you may notice the ‘NA’ values, most of them are related to eigenvector centrality (C_e or E) correlation. These NAs are caused by the failure of eigenvector calculation. In the eigenvector calculation, power iteration is a commonly used algorithm. R program also takes this algorithm and the default maximum number of iteration is 1000. If R is running to the maximum iteration, but fails to get to the convergence, an error occurs. In this case, we treat the eigenvector as ‘NA’, an unknown value. Therefore, the following correlation calculation related to the eigenvector is also set to be ‘NA’.

Table A1: Description of real networks^a

Idx	Name	Source	V	E	Description
1	ca-CondMat-99	Newman ^b [34, 35]	16726	47594	Coauthor networks , constructed from the e-print arXiv archive(www.arxiv.org) on Condensed Matter, Astrophysics and High-Energy Theory. The file tail indicates the year of the data collection. The last one is a coauthor network of scientists working on network theory and experiment.
2	ca-CondMat-03		31163	120029	
3	ca-CondMat-05		40421	175692	
4	ca-AstroPh-99		16706	121251	
5	ca-HepTh-99		8361	15751	
6	ca-netscience-06		1589	2742	
7	ego-Facebook	SNAP ^c [36–39]	4, 039	88, 234	Combined Ego networks , sampled from large online social networks Facebook, Google+ and Twitter
8	ego-Gplus		107614	13673453	
9	ego-Twitter		81306	1768149	
10	soc-Epinions1		75879	508837	sampled from online social networks Epinions and Slash
11	soc-Slash0811		77360	905468	
12	soc-Slash0922		82168	948464	
13	email-EuAll		265214	420045	Email networks
14	email-Enron		36692	183831	
15	ca-DBLP		317080	1049866	
16	ca-AstroPh		18772	198110	Co-author networks , constructed from DBLP community and e-print arXiv archive
17	ca-CondMat		23133	93497	
18	ca-GrQc		5242	14496	
19	ca-HepPh		12008	118521	
20	ca-HepTh		9877	25998	
21	actors	Pajek ^d [28, 40]	520223	1470418	Co-actors networks , based on www.imdb.com
22	ca-Geom		7343	11898	Co-author network in Computational Geometry.
23	cit-HepPh	SNAP ^e [39, 41, 42]	34546	421578	Co-citation networks from the e-print arXiv archive
24	cit-HepTh		27770	352807	
25	com-Amazon		334863	925872	Co-purchasing network of the Amazon website.
26	web-NotreDame		325729	1497134	
27	web-Stanford	Pajek [43–46]	281903	2312497	World Wide Web networks for domains at Notre Dome, Stanford and Berkeley-Stanford, and for search engine queries like ‘California’ and ‘epa’
28	web-BerkStan		685230	7600595	
29	web-qry-ca		9664	15969	
30	web-query-epa		4772	8909	Reuters terror news network FOLDOC dictionary network.
31	terror news		13314	243447	
32	foldoc		13356	120238	

Table A1: ... continued

Idx	Name	Source	V	E	Description
33	power grid	Newman	4941	6594	Power grid network , data citation [47]
34	as-060722		22963	48436	
35	as-980101		3213	11710	
36	as-990101		531	2527	Internet AS networks , constructed by BGP (Border Gateway Protocol) logs from University of Oregon Route View Project.
37	as-000101	SNAP[39, 48]	3570	14783	
38	as-010331		10670	22003	
39	as-Caida-1015		26258	107202	Internet AS relationship networks , constructed by BGP logs and relationship inference algorithms, see caida
40	as-Caida-1105		26475	106762	
41	p2p-Gnutella04		10876	39994	
42	p2p-Gnutella06		8717	31525	p2p networks , Gnutella peer-to-peer file sharing networks
43	p2p-Gnutella08		6301	20777	
44	US air lines	Pajek	332	2126	US air lines
45	neural network	Newman	297	2151	Neural network of C. Elegans, data citation [47]
46	protein-protein		1870	4480	Protein interaction network data (for yeast).
47	celluar-AA		1485	3400	
48	celluar-AB		1416	3230	
49	celluar-AG	CCNR ^d [47, 49, 50]	1567	3631	Cellular networks and its metabolic networks , the file tail
50	metabolic-AA		1057	2527	indicates different organisms.
51	metabolic-AB		993	2368	
52	metabolic-AG		1268	3011	

a Idx: index number; Source: sites to find these collections; |V|: number of vertices; |E|: number of edges.
b Newman's personal site dataset collections: <http://www-personal.umich.edu/~mejn/netdata/>
c Stanford Large Network Dataset Collection: <http://snap.stanford.edu/data/index.html>
d Pajek dataset collections: <http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm>
e CCNR dataset collections: <http://www3.nd.edu/~networks/resources.htm>

Table A2: Network properties of real networks^a

Idx	l	C	r_{ac}	ρ_{ac}	$r(D, B)$	$r(D, C)$	$r(D, E)$	$\rho(D, B)$	$\rho(D, C)$	$\rho(D, E)$	$\tau(D, B)$	$\tau(D, C)$	$\tau(D, E)$
Social	1	6.63	0.36	0.18	0.28	0.72	0.26	-0.53	0.56	0.50	-0.46	0.65	-0.62
	2	5.77	0.28	0.18	0.28	0.75	0.20	-0.63	0.58	0.54	-0.50	0.71	-0.67
	3	5.50	0.25	0.19	0.28	0.76	0.17	-0.66	0.59	0.56	-0.51	0.72	-0.69
	4	4.80	0.43	0.24	0.43	0.68	0.22	0.72	0.52	0.61	0.58	0.78	0.76
	5	7.03	0.33	0.29	0.32	0.75	0.34	0.24	0.67	0.57	0.10	0.78	0.13
	6	5.82	0.69	0.46	0.58	0.43	0.24	0.51	0.42	0.60	0.04	0.71	0.05
	7	3.69	0.52	0.06	0.54	0.45	0.27	0.57	0.61	0.32	0.39	0.79	0.53
	8	2.86	0.22	0.12	0.27	0.42	0.07	-0.97	0.60	0.65	-0.65	0.78	-0.83
	9	3.89	0.17	0.22	0.38	0.48	0.32	-0.59	0.60	0.36	-0.34	0.79	-0.48
	10	4.31	0.07	-0.01	0.17	0.76	0.42	0.90	0.75	0.50	0.52	0.87	0.66
	11	4.02	0.02	-0.05	0.15	0.86	0.49	0.81	0.80	0.64	0.64	0.92	0.80
	12	4.07	0.02	-0.05	0.16	0.85	0.48	0.81	0.78	0.65	0.65	0.91	0.81
	13	4.12	0.00	-0.20	-0.61	0.82	0.02	-0.25	0.64	0.29	0.11	0.66	0.13
	14	4.03	0.09	-0.11	-0.01	0.82	0.07	0.87	0.64	0.38	0.40	0.76	0.53
	15	6.79	0.31	0.27	0.19	0.72	0.47	-0.25	0.59	0.34	-0.30	0.70	-0.42
	16	4.19	0.32	0.21	0.35	0.70	0.14	-0.79	0.56	0.66	-0.63	0.71	-0.82
	17	5.35	0.26	0.13	0.26	0.70	0.15	0.65	0.59	0.53	0.51	0.72	0.68
	18	6.05	0.63	0.66	0.59	0.50	0.23	0.59	0.57	0.45	0.39	0.68	0.53
	19	4.67	0.66	0.63	0.72	0.48	0.10	-0.90	0.51	0.56	-0.48	0.62	-0.64
	20	5.95	0.28	0.27	0.29	0.77	0.21	0.25	0.69	0.50	0.44	0.80	0.60
	21	7.17	0.00	-0.09	-0.17	0.64	0.09	0.52	0.87	0.46	0.42	0.97	0.56
Informational	22	5.31	0.24	0.24	0.39	0.81	0.31	0.83	0.61	0.63	0.42	0.71	0.56
	23	4.33	0.15	-0.01	0.13	0.66	0.08	-0.65	0.57	0.62	-0.55	0.74	-0.73
	24	4.28	0.12	-0.03	0.27	0.72	0.07	0.85	0.44	0.64	0.64	0.60	0.82
	25	11.95	0.21	-0.06	-0.19	0.50	0.20	NA	0.42	0.14	NA	0.56	NA
	26	7.17	0.09	-0.05	-0.11	0.43	0.07	-0.52	0.73	0.23	-0.29	0.87	-0.38
	27	6.82	0.01	-0.11	-0.16	0.55	0.02	-0.55	0.25	0.33	0.14	0.34	0.18
	28	7.11	0.01	-0.11	-0.17	0.46	0.01	NA	0.21	0.33	NA	0.29	NA
	29	5.03	0.02	-0.22	-0.35	0.86	0.30	0.76	0.77	0.80	0.72	0.87	0.87
	30	4.50	0.01	-0.30	-0.59	0.78	0.14	0.29	0.80	0.60	-0.08	0.90	-0.09

Table A2: ... continued

Idx	l	C	r_{ac}	ρ_{ac}	$r(D, B)$	$r(D, C)$	$r(D, E)$	$\rho(D, B)$	$\rho(D, C)$	$\rho(D, E)$	$\tau(D, B)$	$\tau(D, C)$	$\tau(D, E)$
Technological	31	3.12	0.11	0.00	0.09	0.88	0.36	-0.97	0.64	0.68	-0.68	0.79	0.84
	32	3.87	0.11	0.00	0.28	0.87	0.45	0.80	0.56	0.45	0.46	0.62	0.63
	33	18.99	0.10	0.00	-0.05	0.28	0.23	-0.26	0.67	0.18	-0.09	0.80	0.23
	34	3.84	0.01	-0.20	-0.51	0.96	0.18	0.78	0.69	0.37	0.38	0.78	0.48
	35	3.77	0.01	-0.21	-0.48	0.98	0.23	0.83	0.74	0.35	0.35	0.82	0.45
	36	3.39	0.12	-0.22	-0.16	0.95	0.57	-0.89	0.72	0.46	-0.48	0.84	0.58
	37	3.80	0.02	-0.16	-0.38	0.97	0.27	-0.83	0.71	0.36	-0.35	0.81	0.46
	38	3.64	0.01	-0.19	-0.55	0.98	0.17	0.82	0.68	0.33	0.33	0.75	0.43
	39	3.86	0.01	-0.20	-0.52	0.96	0.17	0.81	0.72	0.40	0.41	0.80	0.51
	40	3.88	0.01	-0.19	-0.53	0.96	0.17	-0.57	0.72	0.40	0.13	0.80	0.51
Biological	41	4.64	0.01	-0.01	0.00	0.90	0.85	0.68	0.92	0.82	0.73	0.98	0.95
	42	4.57	0.01	0.05	0.01	0.90	0.83	0.60	0.92	0.82	0.64	0.98	0.94
	43	4.64	0.02	0.04	0.03	0.90	0.71	0.76	0.91	0.80	0.65	0.98	0.93
	44	2.74	0.40	-0.21	-0.14	0.72	0.25	0.96	0.59	0.21	0.72	0.74	0.30
	45	2.46	0.18	-0.16	-0.11	0.78	0.70	0.87	0.74	0.55	0.63	0.89	0.74
	46	6.81	0.06	-0.16	-0.18	0.85	0.19	-0.35	0.83	0.40	-0.33	0.92	0.51
	47	4.49	0.00	-0.19	-0.07	0.91	0.22	-0.62	0.59	0.41	0.19	0.71	0.54
	48	4.40	0.00	-0.16	0.00	0.88	0.15	-0.62	0.72	0.49	0.18	0.84	0.63
	49	4.46	0.00	-0.19	-0.13	0.91	0.13	-0.65	0.57	0.41	-0.44	0.70	0.54
	50	4.49	0.00	-0.19	-0.07	0.91	0.22	-0.62	0.59	0.41	0.19	0.71	0.54
	51	4.40	0.00	-0.16	0.00	0.88	0.15	-0.62	0.72	0.49	0.18	0.84	0.63
	52	4.46	0.00	-0.19	-0.13	0.91	0.13	-0.65	0.57	0.41	-0.44	0.70	0.54

a This table shows network properties and correlation coefficients between centrality metrics of real networks. Idx: index number; l : average path length; C : cluster coefficient; r_{ac} and ρ_{ac} : degree assortativity, the ρ_{ac} is defined in work [27]. We use r , ρ and τ to represent Pearson, Spearman and Kendall coefficients, and D , B , C and E to represent degree, betweenness, closeness and eigenvector centrality metrics, therefore $r(D, B)$ represents the Pearson correlation coefficient between degree and betweenness centrality. Please see Table A1 for the description of datasets and section 3 for the analysis of results.