Uncovering disassortativity in large scale-free networks

Nelly Litvak[*]

University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Sciences, P.O. Box 217,
7500 AE, Enschede, The Netherlands

Remco van der Hofstad[†]

Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513,
5600 MB, Eindhoven, The Netherlands

Mixing patterns in large self-organizing networks, such as the Internet, the World Wide Web, and social and biological networks, are often characterized by degree-degree dependencies between neighboring nodes. In this paper, we propose a new way of measuring degree-degree dependencies. One of the problems with the commonly used assortativity coefficient is that in disassortative networks its magnitude decreases with the network size. We mathematically explain this phenomenon and validate the results on synthetic graphs and real-world network data. As an alternative, we suggest to use rank correlation measures such as Spearman's ρ. Our experiments convincingly show that Spearman's ρ produces consistent values in graphs of different sizes but similar structure, and it is able to reveal strong (positive or negative) dependencies in large graphs. In particular, we discover much stronger negative degree-degree dependencies in Web graphs than was previously thought. Rank correlations allow us to compare the assortativity of networks of different sizes, which is impossible with the assortativity coefficient due to its genuine dependence on the network size. We conclude that rank correlations provide a suitable and informative method for uncovering network mixing patterns.

I. INTRODUCTION

This paper proposes a new way of measuring mixing patterns in large self-organizing networks, such as the Internet, the World Wide Web, and social and biological networks. Most of these real-world networks are scale-free; i.e., their degree distribution has huge variability and closely follows a power law (the fraction of nodes with degree k is roughly proportional to $k^{-\gamma-1}$, $\gamma > 0$). We study correlations between degrees of two nodes connected by an edge. This problem, first posed in Refs. [1,2], has received vast attention in the networks literature, in particular in physics, sociology, biology, and computer science. We show, however, analytically and on the data, that the presence of power laws makes currently used measures inadequate for comparison of mixing patterns in networks of different sizes and we provide an alternative that is free from this disadvantage.

Adequate measuring and comparison of degree-degree correlations is important because mixing patterns define many of the network's properties. For instance, the Internet topology is not sufficiently specified by the degree distribution; the negative degree-degree correlations in the Internet graph have a great influence on the robustness to failures [3], efficiency of Internet protocols [4], as well as distances and betweenness [5]. The Internet topology is totally different from the mixing patterns in networks of bank transactions [6], where the core of 25 most important banks is entirely connected. The correlation between in- and out-degree of tasks plays an important role in the dynamics of production and development systems [7].

Mixing patterns affect epidemic spread [8,9] and Web ranking [10].

In his seminal papers, Newman [1,2] proposed to measure degree-degree correlations using the *assortativity* coefficient, which is, in fact, an empirical estimate of the Pearson's correlation coefficient between the degrees at either ends of a random edge. A network is *assortative* when neighboring nodes are likely to have a similar number of connections. In *disassortative* networks, high-degree nodes mostly have neighbors with a small number of connections. The empirical data in Table I of Ref. [1] suggest that social networks tend to be assortative (which is indicated by the positive assortativity coefficient), while technological and biological networks tend to be disassortative.

In Table I of Ref. [1], it is striking that larger disassortative networks typically have an assortativity coefficient that is closer to 0 and, therefore, appear to have approximately *uncorrelated* degrees across edges. Similar conclusions can be drawn from Table II of Ref. [2]. In recent literature [11,12], the issue was raised that the Pearson's correlation coefficient in scale-free networks decreases with the network size. In this paper, we demonstrate analytically and on the data that in *all* scale-free disassortative networks with a realistic value of the power-law exponent, the assortativity coefficient decreases in magnitude with the size of the graph. In assortative networks, on the other hand, the assortativity coefficient can show two types of pathological behavior. It either decreases with graph size or it shows a considerable dispersion in values, even if large networks are constructed by the same mechanism.

We suggest an alternative solution based on the classical Spearman's ρ measure [13], which is the correlation coefficient computed on the *ranks* of degrees. The huge advantage of such dependency measures is that they work well *independently*

---

[*]n.litvak@utwente.nl

[†]r.w.v.d.hofstad@TUE.nl

of the degree distribution, while the assortativity coefficient, despite the fact that it is always in $[-1,1]$, suffers from a strong dependence on the extreme values of the degrees. The usefullness of the rank correlation approach to discover dependencies in skewed distributions has already been postulated in the 1936 paper by H. Hotelling and M. R. Pabst [14]: *"Certainly where there is complete absence of knowledge of the form of the bivariate distribution, and especially if it is believed not to be normal, the rank correlation coefficient is to be strongly recommended as a means of testing the existence of relationship."*

We compute Spearman's ρ on artificially generated random graphs and on real data from web and social networks. Our results agree with Ref. [1] concerning the presence of positive or negative correlations, but Spearman's ρ has two important advantages: (1) it is able to reveal strong disassortativity in large networks; and (2) it produces consistent values on the graphs created by the same mechanism, e.g., on preferential attachment graphs [15] of different sizes. Thus, Spearman's ρ correctly and consistently captures the underlying connection patterns and tendencies. We conclude that when networks are large, or two networks of difference sizes must be compared (e.g., in web crawls or social networks from different countries), Spearman's ρ is a preferred method for measuring and comparing degree-degree correlations.

The closing section discusses further challenges in the evaluation of network mixing patterns.

II. NO DISASSORTATIVE SCALE-FREE RANDOM GRAPH SEQUENCES

In this section, we present a simple analytical argument that in disassortative networks the assortativity coefficient always decreases in magnitude with the size of the graph. Formal proofs can be found in Ref. [16].

Assortativity in networks is usually measured using the assortativity coefficient, which is in fact a statistical estimator of Pearson's correlation coefficient for the degrees on the two ends of an arbitrary edge in a graph. Let $G = (V, E)$ be a graph with vertex set V, where $|V| = n$ denotes the size of the network, and edge set E. The assortativity coefficient of G is equal to (see, e.g., Eq. (4) in [1])

$$\rho_n = \frac{\frac{1}{|E|} \sum_{ij \in E} d_i d_j - \left[\frac{1}{|E|} \sum_{ij \in E} \frac{1}{2}(d_i + d_j) \right]^2}{\frac{1}{|E|} \sum_{ij \in E} \frac{1}{2}(d_i^2 + d_j^2) - \left[\frac{1}{|E|} \sum_{ij \in E} \frac{1}{2}(d_i + d_j) \right]^2}, \tag{2.1}$$

where the sum is over directed edges of G, i.e., ij and ji are two distinct edges, and d_i is the degree of vertex i. We compute that

$$\frac{1}{|E|} \sum_{ij \in E} \frac{1}{2}(d_i + d_j) = \frac{1}{|E|} \sum_{i \in V} d_i^2,$$

$$\frac{1}{|E|} \sum_{ij \in E} \frac{1}{2}(d_i^2 + d_j^2) = \frac{1}{|E|} \sum_{i \in V} d_i^3.$$

Thus, ρ_n can be written as

$$\rho_n = \frac{\sum_{ij \in E} d_i d_j - \frac{1}{|E|} \left(\sum_{i \in V} d_i^2 \right)^2}{\sum_{i \in V} d_i^3 - \frac{1}{|E|} \left(\sum_{i \in V} d_i^2 \right)^2}. \tag{2.2}$$

In practice, all quantities in (2.2) are finite, and ρ_n can always be computed. However, since many real-life networks are very large, a relevant question is how ρ_n behaves when n becomes large.

In the literature, many examples are reported of real-world networks where the degree distribution obeys a power law [17,18]. In particular, for scale-free networks, the observed proportion of vertices of degree k is close to $f(k) = c_0 k^{-\gamma-1}$, and most values of γ found in real-world networks are in $(1,3)$, see, e.g., Table I in Ref. [17] or Table I in Ref. [18]. For $p < \gamma$, let $\mu_p = \sum_k k^p f(k)$, and note that the series diverges if $p \geqslant \gamma$; let $a \sim b$ denote that $a/b \to 1$. Then, we can expect that, as n grows large,

$$|E| = \sum_{i \in V} d_i \sim \mu_1 n, \quad \sum_{i \in V} d_i^p \sim \mu_p n, \quad p < \gamma,$$

while $\max_{i \in V} d_i$ is of the order $n^{1/\gamma}$. As a direct consequence,

$$cn \leqslant |E| \leqslant Cn, \tag{2.3}$$

$$cn^{1/\gamma} \leqslant \max_{i \in V} d_i \leqslant Cn^{1/\gamma}, \tag{2.4}$$

$$cn^{\max\{p/\gamma,1\}} \leqslant \sum_{i \in V} d_i^p \leqslant Cn^{\max\{p/\gamma,1\}}, \quad p = 2,3, \tag{2.5}$$

for $\gamma \in (1,3)$ and some constants $0 < c < C < \infty$. We emphasize that conditions Eqs. (2.3)–(2.5) are very general and hold for any scale-free network of growing size, independently of its mixing patterns. From Eq. (2.2), we simply write

$$\rho_n \geqslant \rho_n^- \equiv -\frac{\frac{1}{|E|} \left(\sum_{i \in V} d_i^2 \right)^2}{\sum_{i \in V} d_i^3 - \frac{1}{|E|} \left(\sum_{i \in V} d_i^2 \right)^2},$$

and notice that

$$\sum_{i \in V} d_i^3 \geqslant (\max_{i \in V} d_i)^3 \geqslant c^3 n^{3/\gamma},$$

whereas

$$\frac{1}{|E|} \left(\sum_{i \in V} d_i^2 \right)^2 \leqslant (C^2/c) n^{2\max\{2/\gamma,1\}-1} = (C^2/c) n^{\max\{4/\gamma-1,1\}}.$$

Since $\gamma \in (1,3)$, we have $\max\{4/\gamma - 1, 1\} < 3/\gamma$, so that

$$\frac{\sum_{i \in V} d_i^3}{\frac{1}{|E|} \left(\sum_{i \in V} d_i^2 \right)^2} \to \infty \quad \text{as } t \to \infty.$$

Hence, the lower bound ρ_n^- is of the order $n^{\max\{1/\gamma-1, 1-3/\gamma\}}$. It is now easy to check that if $\gamma \in (1,3)$, then ρ_n^- converges to zero when the graph size increases. This means that any limit point of the assortativity coefficients ρ_n is nonnegative. Note also that ρ_n^- is defined by the degree sequence, and it does not depend on the mixing pattern at all. We conclude that by looking only at the value of ρ_n, one cannot discover even very strong disassortativity in large scale-free graphs. We will confirm this finding in Sec. IV, on artificially generated random graphs, and in Sec. V, on real-world networks.

We note that if $\gamma > 3$, then all terms in Eq. (2.1) converge to a number, and ρ_n does not scale with the network size. In practice this means that the dependence of ρ_n on the graph

size is observed when node degrees have a broad distribution, and this range increases when the network gets bigger. This is the case in most real-life networks and models for them, as is, e.g., obviously the case for preferential attachment models.

We further notice that Eqs. (2.3)–(2.5) imply that

$$\sum_{ij \in E} d_i d_j \leqslant \left(\max_{i \in V} d_i \right) \sum_{ij \in E} d_i$$
$$= \left(\max_{i \in V} d_i \right) \sum_{i \in V} d_i^2 \leqslant C^2 n^{1/\gamma + \max\{2/\gamma, 1\}}. \quad (2.6)$$

Mathematically, an interesting case is when $\sum_{ij \in E} d_i d_j$ and $\sum_{i \in V} d_i^3$ are of the same order of magnitude. Then the network is assortative, but, formally, ρ_n converges to a random variable. In practice this means that ρ_n can result in very different values on two very large graphs constructed by the same mechanism. We will give such an example in Sec. IV.

III. RANK CORRELATIONS

We propose an alternative measure for the degree-degree dependencies, based on rank correlations. For two-dimensional data $[(X_i, Y_i)]_{i=1}^n$, let r_i^X and r_i^Y be the rank of an observation X_i and Y_i, respectively, when the sample values $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ are arranged in a descending order. The rank correlation measures evaluate statistical dependencies on the data $[(r_i^X, r_i^Y)]_{i=1}^n$, rather than on the original data $[(X_i, Y_i)]_{i=1}^n$. Rank transformation is convenient, in particular because (r_i^X) and (r_i^Y) are samples from the same uniform distribution, which implies many nice mathematical properties.

The statistical correlation coefficient for the rank is known as Spearman's ρ [13]:

$$\rho^{\text{rank}} = \frac{\sum_{i=1}^n \left[r_i^X - (n+1)/2 \right] \left[r_i^Y - (n+1)/2 \right]}{\sqrt{\sum_{i=1}^n \left[r_i^X - (n+1)/2 \right]^2 \sum_i^n \left[r_i^Y - (n+1)/2 \right]^2}}. \quad (3.1)$$

The mathematical properties of the Spearman's ρ have been extensively investigated. In particular, if $[(X_i, Y_i)]_{i=1}^n$ consists of independent realizations of (X, Y), and the joint distribution function of X and Y is differentiable, then ρ_n^{rank} is a consistent statistical estimator, and its standard deviation is of the order $1/\sqrt{n}$ independently of the exact form of the underlying distributions, see, e.g., Ref. [19].

For a graph G of size n, we propose to compute ρ_n^{rank} using Eq. (3.1) as follows. We define the random variables X and Y as the degrees on two ends of a random *undirected* edge in a graph (that is, when rank correlations are computed, ij and ji represent the same edge). For each edge, when the observed degrees are a and b, we assign $[X = a, Y = b]$ or $[X = b, Y = a]$ with probability $1/2$. Many values of X and Y will be the same, making their rank ambiguous. We resolve this by adding independent uniformly distributed random variables on $[0, 1]$ to each value of X and Y. In the setting when the realizations (X_i, Y_i) are independent, this way of resolving ties preserves the original value of the Spearman's ρ on the population; see, e.g., Ref. [20]. We refer to Ref. [21] for a general treatment of rank correlations for noncontinuous distributions.

In the remainder of the paper we will demonstrate that the measure ρ_n^{rank} gives consistent results for different n, and it is able to reveal strong negative degree-degree correlations in large networks.

IV. RANDOM GRAPH DATA

We consider four random graph models to highlight our results.

The configuration model. The *configuration model* was invented by Bollobás in Ref. [22], inspired by Ref. [23]. It was popularized by Newman, Strogatz, and Watts [24], who realized that it is a useful and simple model for real-world networks. In the configurations model, a node i has a given number d_i of half-edges, with $\ell_n = \sum_{i \in V} d_i$ assumed to be even. Each half-edge is connected to a randomly chosen other half-edge to form an edge in the graph. We chose $\gamma = 2$; thus, the maximum degree is of the order $n^{1/2}$, which corresponds to the case of uncorrelated random networks, such that the probability that two vertices are directly connected is close to $d_i d_j / \ell_n$ [25,26]. Although self-loops and multiple edges can occur, these become rare as $n \to \infty$, see, e.g., Ref. [27] or [28]. In simulations, we collapse multiple edges to a single edge and remove self-loops. This changes the degree distribution slightly and intuitively should yield negative dependencies. In Fig. 1(a) we observe that, on average, ρ_n and ρ_n^{rank} are indeed negative in smaller networks but then they converge to zero showing that the degrees on two ends of a random edge are uncorrelated.

Configuration model with intermediate vertices. In order to construct a strongly disassortative graph, we first generate a configuration model as described above, and then we replace every edge by two edges that meet at a middle vertex. In this model, there are $n + \ell_n/2$ vertices and $2\ell_n$ edges (recall that ij and ji are two different edges). Now, if E, V, and d_i, $i = 1, \ldots, n$ denote, respectively, the edge set, the vertex set, and the degrees of the original configuration model, then in the model with intermediate edges the assortativity coefficient is as follows:

$$\rho_n = \frac{2 \sum_{i \in V} 2d_i - \frac{1}{2\ell_n} \left(\sum_{i \in V} d_i^2 + 2\ell_n \right)^2}{\sum_{i \in V} d_i^3 + 4\ell_n - \frac{1}{2\ell_n} \left(\sum_{i \in V} d_i^2 + 2\ell_n \right)^2}.$$

When $\gamma < 3$, we have $\mu_3 = \infty$, and thus $\rho_n \to 0$ as $n \to \infty$. Furthermore, the lower bound ρ_n^- also converges to zero as n grows. It is clear that this particular random graph, of any size, is equally and strongly disassortative; however, ρ_n fails to capture this. In Fig. 1(b) it is clearly seen that both ρ_n and ρ_n^- quickly decrease in magnitude as n grows. It is striking that ρ_n^{rank} shows a totally different and very appropriate behavior. Its values remain around -0.75, identifying the strong negative dependencies, and the dispersion across different realizations of the graph decreases as $n \to \infty$.

Preferential attachment model. We consider the basic version of the undirected preferential attachment model (PAM), where each new vertex adds only one edge to the network, connecting to the existing nodes with probability proportional to their degrees [15]. In this case, it is well known that $\gamma = 2$ (see, e.g., Ref. [29]). Newman [1] noticed the counterintuitive fact that the preferential attachment graph has asymptotically
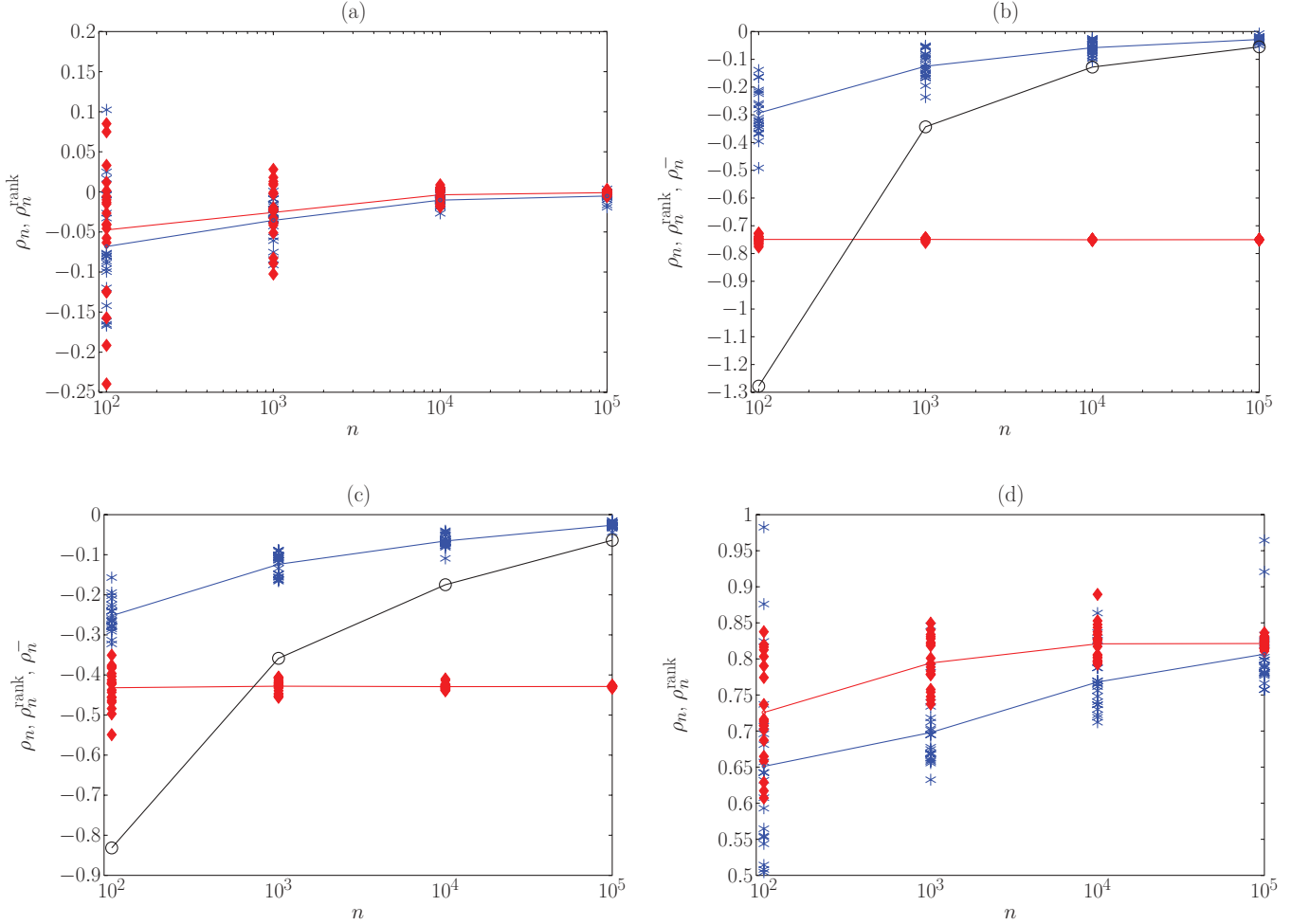
FIG. 1. (Color online) Scatter plots for samples of 20 graphs. For each size we plot the 20 realizations of ρ_n (blue asterisks) and ρ_n^{rank} (red diamonds) in random graphs. Solid lines connect the averages of the samples. In (c), (d) the circles connected by the solid line are the averages of ρ_n^- in the samples. (a) Configuration model, $\mathbb{P}(d \geqslant x) = x^{-2}$, $x \geqslant 1$. (b) Configuration model with intermediate vertices. (c) Preferential attachment model. (d) A collection of bipartite graphs, where $b = 1/2$, $a = 2$, and U has a generalized Pareto distribution $\mathbb{P}(U > x) = [(2.1 + x)/3.1]^{-3.1}$, $x > 1$.

neutral mixing, $\rho_n \to 0$ as $n \to \infty$. This phenomenon has been studied in detail by Dorogovtsev *et al.* [11], and it can be clearly observed in Fig. 1(c). The reason for this behavior is not the genuine neutral mixing in the PAM but rather the unnatural dependence of ρ_n on the graph size. Indeed, we see that PAMs of small sizes have $\rho_n < 0$, and then the magnitude of ρ_n decreases with the graph size. Again, Spearman's ρ consistently shows that the degrees are negatively dependent. This can be understood by noting that the majority of edges of vertices with high degrees, which are old vertices, come from vertices that are added late in the graph growth process and thus have small degree. On the other hand, by the growth mechanism of the PAM, vertices with low degree are more likely to be connected to vertices having high degree, which indeed suggests negative degree-degree dependencies.

A collection of complete bipartite graphs. We next present an example where the assortativity coefficient has a non-vanishing dispersion. Take $[(X_i, Y_i)]_{i=1}^n$ to be a sample of independent realizations of the vector (X, Y). We assume that $X = bU_1 + bU_2$ and $Y = bU_1 + aU_2$, where $b > 0$, $a > 1$,

and U_1, U_2 are independent identically distributed (i.i.d.) random variables with power law tail and tail exponent γ. Then, for $i = 1, \ldots, n$, we create a complete bipartite graph of X_i and Y_i vertices, respectively. These n complete bipartite graphs are not connected to one another. We denote such a collection of n bipartite graphs by G_n. This is an extreme scenario of a network consisting of highly connected clusters of different size. Such networks can serve as models for physical human contacts and are used in epidemic modeling [9].

The graph G_n has $|V| = \sum_{i=1}^n (X_i + Y_i)$ vertices and $|E| = 2 \sum_{i=1}^n X_i Y_i$ edges. Further,

$$\sum_{i \in V} d_i^p = \sum_{i=1}^n \left(X_i^p Y_i + Y_i^p X_i \right), \qquad \sum_{ij \in E} d_i d_j = 2 \sum_{i=1}^n (X_i Y_i)^2.$$

Assume that $\mathbb{P}(U_j > x) = c_0 x^{-\gamma}$, where $c_0 > 0$, $x \geqslant x_0$, and $\gamma \in (3,4)$, so that $\mathbb{E}[U^3] < \infty$, but $\mathbb{E}[U^4] = \infty$. As a result, $|E|/n \xrightarrow{\mathbb{P}} 2\mathbb{E}[XY] < \infty$ and $\frac{1}{n} \sum_{i \in V} d_i^2 \xrightarrow{\mathbb{P}} \mathbb{E}[XY(X +$

TABLE I. (i)–(iv) Web crawls: nodes are Web pages, and an (undirected) edge means that there is a hyperlink from one of the two pages to another; (iii) and (iv) are breadth-first crawls around one page. (v) E-mail exchange by Enron employees (mostly part of the senior management): nodes are employees, and an edge means that an e-mail message was sent from one of the two employees to another. (vi) and (vii) are scientific collaboration networks extracted from the DBLP bibliography service; each vertex represents a scientist and an edge means a co-authorship of at least one article. (viii) Vertices are actors, and two actors are connected by an edge if they appeared in the same movie.

nr	Dataset	Description	No. nodes	No. edges	Max. degree	ρ_n	ρ_n^{rank}	ρ_n^-
(i)	Stanford-cs	web domain	9 914	54 854	340	−0.1656	−0.1627	−0.4648
(ii)	eu-2005	.eu web domain	862 664	5 477 938	68 963	−0.0562	−0.2525	−0.0670
(iii)	uk@100,000	.uk web crawl	100 000	5 559 150	55 252	−0.6536	−0.5676	−1.117
(iv)	uk@1,000,000	.uk web crawl	1 000 000	77 123 940	403 441	−0.0831	−0.5620	−0.0854
(v)	enron	e-mail exchange	69 244	506 898	1 634	−0.1599	−0.6827	−0.1932
(vi)	dblp-2010	co-authorship	326 186	1 615 400	238	0.3018	0.2604	−0.7736
(vii)	dblp-2011	co-authorship	986 324	6 707 236	979	0.0842	0.1351	−0.2963
(viii)	Hollywood-2009	co-starring	1 139 905	113 891 327	11 468	0.3446	0.4689	−0.6737

$Y)] < \infty$. Further,

$$n^{-4/\gamma} b^{-4} \sum_{i=1}^{n} \left(X_i^3 Y_i + Y_i^3 X_i \right) \xrightarrow{d} (a^3 + a) Z_1 + 2 Z_2,$$

$$n^{-4/\gamma} b^{-4} \sum_{i=1}^{N} (X_i Y_i)^2 \xrightarrow{d} a^2 Z_1 + Z_2,$$

where Z_1 and Z_2 and two independent stable distributions with parameter $\gamma/4$. As a result,

$$\rho_n \xrightarrow{d} \frac{2a^2 Z_1 + 2 Z_2}{(a + a^3) Z_1 + 2 Z_2}, \qquad \text{as } n \to \infty,$$

which is a proper random variable taking values in $[2a/(1 + a^2), 1]$; see Ref. [16] for detailed proof.

Note that in this model there is a genuine dependence between the correlation measure and the graph size. Indeed, if $n = 1$ then the assortativity coefficient equals -1 because nodes with larger degrees are connected to nodes with smaller degrees. However, when the graph size grows, the positive linear dependence between X and Y starts dominating; thus, larger graphs of this structure are strongly assortative. While the example we present is quite special, we believe that the effect described is rather general.

In Fig. 1(d) we again see that ρ_n^{rank} captures the relation faster and gives consistent results with decreasing dispersion. On the contrary, ρ_n has a persistent dispersion in its values, and we know from the result above that this dispersion will not vanish as $n \to \infty$. In the limit, ρ_n has a nonzero density on $(0.8, 1)$. However, the convergence is too slow to observe it at $n = 100\,000$, because the vanishing terms are of the order $n^{-1/\gamma}$, which is only $n^{-1/3.1}$ in our example.

V. WEB SAMPLES AND SOCIAL NETWORKS

We computed ρ_n, ρ_n^{rank}, and ρ_n^- on several Web samples (disassortative networks) and social network samples (assortative networks). We used the compressed graph data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano [30,31]. We used the bvgraph MATLAB package [32]. The *stanford-cs* database [33] is a 2001 crawl that includes all pages in the cs.stanford.edu domain. In datasets (iv), (vii), and (viii) we evaluate ρ_n, ρ_n^{rank}, and ρ_n^-

over 1000 random edges and present the average over 10 such evaluations (in 10 samples of 1000 edges, the observed dispersion of the results was small).

The results are presented in Table I. We clearly see that the assortativity coefficient ρ_n and Spearman's ρ_n^{rank} always agree about whether dependencies are positive or negative. They also agree in magnitude of correlations when the graph size is small or the lower bound ρ_n^- is sufficiently far from zero. However, ρ_n is not consistent for graphs of similar structure but different sizes. This is especially apparent on the two .uk crawls (iii) and (iv). Here, ρ_n is significantly smaller in magnitude on a larger crawl. Intuitively, mixing patterns should not depend on the crawl size. This is indeed confirmed by the value of Spearman's ρ, which consistently shows strong negative correlations in both crawls. We could not observe a similar phenomenon so sharply in (vi) and (vii), probably because a larger coauthorship network incorporates articles from different areas of science, and the culture of scientific collaborations can vary greatly from one research field to another.

We also notice that, as predicted by our results, the assortativity coefficient tends to take smaller values than ρ_n^{rank} if ρ_n^- is small in magnitude. This is clearly seen in the data sets (ii), (iv), and (v). Again, (ii) and (iv) are the largest among the analyzed web crawls.

The observed behavior of the assortativity coefficient is explained by the above stated results that ρ_n is strongly influenced by the large dispersion in the degree values. The latter increases with graph size because of the scale-free phenomenon. As a result, ρ_n becomes smaller in magnitude, which makes it impossible to compare graphs of different sizes. In contrast, the *ranks* of the degrees are drawn from a uniform distribution on [0,1], scaled by the factor n. Clearly, when a correlation coefficient is computed, the scaling factor cancels, and therefore Spearman's ρ provides consistent results in the graphs of different sizes.

VI. DISCUSSION

The assortativity coefficient ρ_n proposed in [1,2] has been the first dependency measure introduced to describe degree-degree correlations in networks. The assortativity coefficient has provided many interesting insights. It has been successfully

used for comparison of dependencies in graphs with the same degree sequences [34,35] and to generate graphs with given degrees and desired mixing patterns [36]. An important drawback of ρ_n is its dependence on the network size n. It has been noticed by many authors, and shown in this paper for disassortative networks, that ρ_n converges to zero as n grows. In particular, the decay with network size of the assortativity coefficient ρ_n implies that it cannot be used for comparing dependencies in networks of different sizes. Therefore, it prohibits the investigation as to whether growing networks become more or less assortitative over time.

This paper suggests to use rank-correlation measures such as Spearman's ρ. Our experiments convincingly show that Spearman's ρ does not suffer from the size-dependence deficiency. In networks of different sizes but similar structure, Spearman's ρ yields consistent results, and it is able to reveal strong (positive or negative) correlations in large networks. We conclude that rank correlations are a suitable and informative method for uncovering network mixing patterns.

For the correct interpretation of degree-degree dependencies, it is important to realize that positive or negative correlations can be predefined by the degree sequence itself. For instance, there is only one simple graph with degrees $(3,1,1,1)$, and the result $\rho_4 = -1$ is not informative in this case. It has been discussed in the literature that, conditioned on not having self-loops and multiple edges, random networks with given degrees exhibit disassortative patterns [25,35,37], also called *structural* correlations. In order to filter out the structural correlations, one needs to compare the real-world networks to their null models—graphs with the same degree sequences but *random* connections. This null model is a uniform simple random graph with the same degree sequence. Here a network is called simple when it has no self-loops nor multiple edges. Such a graph can be obtained by randomly pairing half-edges, as described in Sec. IV, and taking the first realization that is simple. This is especially problematic when $(\max_i d_i)^2 > |E|$, which is the case in many examples, since then one needs a prohibitingly large number of attempts before a simple graph is generated [28,38].

A widely accepted method for constructing a null model is the random rewiring of the connections in a given graph [34,35]. The disadvantage is the unknown running time before a graph is produced that is close enough to being uniform. Recent work [39] presents a sequential algorithm, where, at each step, the remaining unconnected edges maintain the ability to generate a simple graph. This method always produces the desired outcome but its worst-case running time $O(n^2 \sum_i d_i)$ is infeasible for large networks. The recently introduced grand-canonical model [40] computes the probability of connection between two nodes in a maximum entropy graph with prescribed degree sequence and enables the evaluation of many characteristics of the graph. To the best of our knowledge, efficient implementation of this method for large networks has not been developed yet.

Constructing a null model and filtering out the structural correlations in large networks is an interesting and demanding computational task that is beyond the scope of this paper. We believe that structural correlations will affect ρ_n to a larger extent than the rank correlation ρ_n^{rank} because it is usually the nodes with largest degrees that produce self-loops and multiple edges, and thus the relative contribution of these edges in the cross products will be larger for ρ_n than for ρ_n^{rank}. This conjecture requires a further investigation.

We conclude by stating that rank correlation measures deserve to become a standard tool in the analysis of complex networks. The use of rank correlation measures has become common ground in the area of statistics for analyzing heavy-tailed data. We hope to have provided sufficient evidence that this method is preferred for analyzing network data with heavy-tailed degrees as well.

[1] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).

[2] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).

[3] J. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, Proc. Natl. Acad. Sci. USA **102**, 14497 (2005).

[4] L. Li, D. Alderson, J. Doyle, and W. Willinger, Internet Math. **2**, 431 (2005).

[5] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, ACM SIGCOMM Comput. Commun. Rev. **36**, 135 (2006).

[6] R. May, S. Levin, and G. Sugihara, Nature (London) **451**, 893 (2008).

[7] D. Braha and Y. Bar-Yam, Manag. Sci. **53**, 1127 (2007).

[8] V. M. Eguíluz and K. Klemm, Phys. Rev. Lett. **89**, 108701 (2002).

[9] S. Eubank, H. Guclu, V. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, Nature (London) **429**, 180 (2004).

[10] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, Internet Math. **4**, 245 (2007).

[11] S. N. Dorogovtsev, A. L. Ferreira, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. E **81**, 031135 (2010).

[12] M. Raschke, M. Schläpfer, and R. Nibali, Phys. Rev. E **82**, 037102 (2010).

[13] C. Spearman, Am. J. Psychol. **15**, 72 (1904).

[14] H. Hotelling and M. Pabst, Ann. Math. Stat. **7**, 29 (1936).

[15] R. Albert and A. Barabási, Science **286**, 509 (1999).

[16] N. Litvak and R. van der Hofstad (to be published).

[17] R. Albert and A. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[18] M. Newman, SIAM Rev. **45**, 167 (2003).

[19] C. Borkowf, Comput. Stat. Data Anal. **39**, 271 (2002).

[20] M. Mesfioui and A. Tajar, Nonparametric Stat. **17**, 541 (2005).

[21] J. Nevslehová, J. Multivariate Anal. **98**, 544 (2007).

[22] B. Bollobás, Eur. J. Combin. **1**, 311 (1980).

[23] E. Bender and E. Canfield, J. Comb. Theory, Ser. A **24**, 296 (1978).

[24] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[25] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 205 (2004).

[26] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).

[27] B. Bollobás, *Random Graphs*, Vol. 73 (Cambridge University Press, New York, 2001).

[28] S. Janson, Combinat. Prob. Comput. **18**, 205 (2009).

[29] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, Random Struct. Algor. **18**, 279 (2001).

[30] P. Boldi and S. Vigna, in *Proceedings of the 13th International World Wide Web Conference* (ACM Press, New York, 2004), pp. 595–601.

[31] P. Boldi, M. Rosa, M. Santini, and S. Vigna, in *Proceedings of the 20th International World Wide Web Conference* (ACM Press, New York, 2011).

[32] D. Gleich, A. Gray, C. Greif, and T. Lau, SIAM J. Sci. Comput. **32**, 349 (2010).

[33] P. Constantine and D. Gleich, in *Proceedings of the 5th Workshop on Algorithms and Models for the Web Graph*, Lecture Notes in Computer Science, Vol. 4863, edited by A. Bonato and F. C. Graham (Springer, Berlin, 2007), pp. 82–95.

[34] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[35] S. Maslov, K. Sneppen, and A. Zaliznyak, Physica A **333**, 529 (2004).

[36] P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. Kuipers, Eur. Phys. J. B **76**, 643 (2010).

[37] J. Park and M. E. J. Newman, Phys. Rev. E **68**, 026112 (2003).

[38] R. van der Hofstad, Random Graphs and Complex Networks (2012, in press) [www.win.tue.nl/~rhofstad/NotesRGCN.pdf].

[39] J. Blitzstein and P. Diaconis, Internet Math. **6**, 489 (2011).

[40] T. Squartini and D. Garlaschelli, New J. Phys. **13**, 083001 (2011).