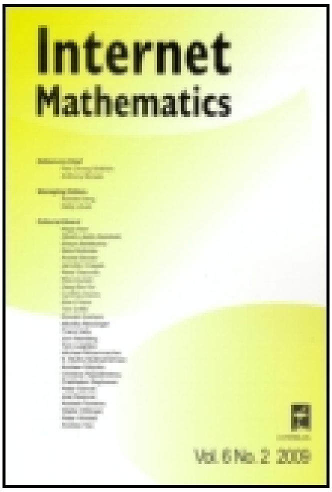

Internet Mathematics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uinm20

Degree-Degree Dependencies in Directed Networks with Heavy-Tailed Degrees

Pim van der Hoorn[a] & Nelly Litvak[a]
[a] Faculty of Electrical Engineering, Mathematics and Computer Sciences, University of Twente, Enschede, The Netherlands
Accepted author version posted online: 07 Jul 2014.

To cite this article: Pim van der Hoorn & Nelly Litvak (2015) Degree-Degree Dependencies in Directed Networks with Heavy-Tailed Degrees, Internet Mathematics, 11:2, 155-179, DOI: 10.1080/15427951.2014.927038

To link to this article: http://dx.doi.org/10.1080/15427951.2014.927038

Internet Mathematics, 11:155–179, 2015

ISSN: 1542-7951 print/1944-9488 online
DOI: 10.1080/15427951.2014.927038

DEGREE-DEGREE DEPENDENCIES IN DIRECTED NETWORKS WITH HEAVY-TAILED DEGREES

Pim van der Hoorn and Nelly Litvak
Faculty of Electrical Engineering, Mathematics and Computer Sciences, University of Twente, Enschede, The Netherlands

Abstract *In network theory, Pearson's correlation coefficients are most commonly used to measure the degree assortativity of a network. We investigate the behavior of these coefficients in the setting of directed networks with heavy-tailed degree sequences. We prove that for graphs where the in- and out-degree sequences satisfy a power law with realistic parameters, Pearson's correlation coefficients converge to a nonnegative number in the infinite network size limit. We propose alternative measures for degree-degree dependencies in directed networks based on Spearman's rho and Kendall's tau. Using examples and calculations on the Wikipedia graphs for nine different languages, we show why these rank correlation measures are more suited for measuring degree assortativity in directed graphs with heavy-tailed degrees.*

1. INTRODUCTION

In the analysis of the topology of complex networks, a feature that is often studied is the *degree-degree dependency*, also called the *degree assortativity* of the network. A network is called assortative, when nodes with high degree have a preference to be connected to nodes of similar large degree. When nodes with large degree have a connection preference for nodes with low degree, the network is said to be *disassortative*. A measure for degree assortativity was first given for undirected networks by Newman [16], which corresponds to Pearson's correlation coefficient of the degrees at the ends of a random edge in the network. A similar definition for directed networks was introduced in [17] and later adopted for analysis of directed complex networks in [9] and [20].

Degree assortativity in networks has been analyzed in a variety of scientific fields such as neuroscience, molecular biology, information theory, and social network sciences and has been found to influence several properties of a network. In [10] and [12] degree-degree correlations are used to investigate the structure of collaboration networks of a social news-sharing website and Wikipedia discussion pages, respectively. Neural networks with high assortativity seem to behave more efficiently under the influence of noise [8] and information content has been shown to depend on the absolute value of the degree assortativity [19]. The effects of degree-degree dependencies on epidemic spreading have been studied in percolation theory [2, 26], and it has been shown, for instance, that the epidemic threshold

Address correspondence to Pim van der Hoorn, Faculty of Electrical Engineering, Mathematics and Computer Sciences, University of Twente, P.O. Box 217, Enschede 7500AE, The Netherlands. E-mail: w.l.f.vanderhoorn@utwente.nl

depends on these correlations. Degree assortativity is also used in the analysis of networks under attack, e.g., P2P networks [23, 24]. Networks with high degree assortativity seem to be less stable under attack [5]. In the case of directed networks, recent research [15] has shown that degree-degree dependencies can influence the rate of consensus in directed social networks such as Twitter.

Recently it has been shown [13, 14] that for undirected networks of which the degree sequence satisfies a power-law distribution with exponent $\gamma \in (1, 3)$, Pearson's correlation coefficient scales with the network size, converging to a nonnegative number in the infinite network size limit. Because most real-world networks have been reported to be scale free with exponent in $(1, 3)$, c.f. [1, 18, Table II], this could then explain why large networks are rarely classified as disassortative. In [13, 14] a new measure, corresponding to Spearman's rho [22], has been proposed as an alternative.

In this article we will extend the analysis in [13] to the setting of directed networks. Here we have to consider four types of degree-degree dependencies, depending on the choice for in- or out-degree on either side of an edge. Our message is similar to that of [13]; that Pearson's correlation coefficients are size biased and produce undesirable results, hence we should look for other means to measure degree-degree dependencies.

We consider networks where the in- and out-degree sequences have a power-law distribution. We will give conditions on the exponents of the in- and out-degree sequences for which the assortativity measures defined in [9] and [20] converge to a nonnegative number in the infinite network size limit. This result is a strong argument against the use of Pearson's correlation coefficients for measuring degree-degree dependencies in such directed networks. To strengthen this argument, we also give examples that clearly show that the values given by Pearson's correlation coefficients do not represent the true dependency between the degrees, which it is supposed to measure. As an alternative, we propose correlation measures based on Spearman's rho [22] and Kendall's tau [11]. These measures are based on the ranking of the degrees rather than their value and, hence, do not exhibit the size bias observed in Pearson's correlation coefficients. We give several examples that show the difference between these three measures. We also include an example for which one of the four Pearson's correlation coefficients converges to a random variable in the infinite network size limit and therefore will obviously produce uninformative results. Finally, we calculate all four degree-degree correlations on the Wikipedia network for nine different languages, using all the assortativity measures proposed in this study.

This article is structured as follows. In Section 2 we introduce notations. Pearson's correlation coefficients are introduced in Section 3, and a convergence theorem is given for these measures. We introduce the rank correlations, Spearman's rho, and Kendall's tau for degree-degree dependencies in Section 4. Example graphs that illustrate the difference between the three measures are presented in Section 5, and the degree-degree correlations for the Wikipedia graphs are presented in Section 6. Finally, in Section 7 we briefly discuss the results and their interpretations.

2. DEFINITIONS AND NOTATIONS

We start with the formal definition of the problem and introduce the notations that will be used throughout the article.

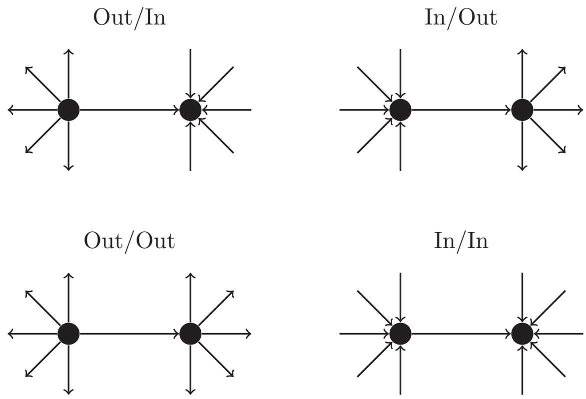

Figure 1 Four degree-degree dependency types.

2.1. Graphs, Vertices, and Degrees

We will denote by $G = (V, E)$ a directed graph with vertex set V and edge set $E \subseteq V \times V$. For an edge $e \in E$, we denote its source by e_* and its target by e^*. With each directed graph, we associate two functions $D^+, D^- : V \to \mathbb{N}$ where $D^+(v) := |\{e \in E | e_* = v\}|$ is the out-degree of the vertex v and $D^-(v) := |\{e \in E | e^* = v\}|$ the in-degree. When considering sequences of graphs, we denote by $G_n = (V_n, E_n)$ an element of the sequence $(G_n)_{n \in \mathbb{N}}$. We further use subscripts to distinguish between the different graphs in the sequence. For instance, D_n^+ and D_n^- will denote the out- and in-degree functions of the graph G_n, respectively.

2.2. Four Types of Degree-Degree Dependencies

In this work we are interested in measuring dependencies between the degrees at both sides of an edge. That is, we measure the relation between two vectors, X and Y, as a function of the edges $e \in E$, corresponding to the degrees of e_* and e^*, respectively. In the undirected case, this is called the degree assortativity. In the directed setting, however, we can consider any combination of the two degree types resulting in four types of degree-degree dependencies, illustrated in Figure 1.

From Figure 1 one can already observe some interesting features of these dependencies. For instance, in the Out/In case, the edge that we consider contributes to the degrees on both sides. We will later see that, for this reason, the Out/In dependency in fact generalizes the undirected case. More precisely, our result for the Out/In dependencies generalizes the result from [14] when we transform from the undirected to the directed case by making every edge bidirectional.

For the other three dependency types, we observe that there is always at least one side where the considered edge does not contribute toward the degree on that side. We will later see that, for these dependency types, the dependency of the in- and out-degree of a vertex will play a role.

3. PEARSON'S CORRELATION COEFFICIENT

Among degree-degree dependency measures, the measure proposed by Newman [16, 17] has been widely used. This measure is the statistical estimator for the Pearson correlation coefficient of the degrees on both sides of a random edge. However, for undirected networks with heavy-tailed degrees with exponent $\gamma \in (1,3)$ it was proved [14] that this measure converges, in the infinite size network limit, to a non-negative number. Therefore, in these cases, Pearson's correlation coefficient is not able to correctly measure negative degree-degree dependencies. In this section we will extend this result to directed networks proving that also here Pearson's correlation coefficients are not the right tool to measure degree-degree dependencies.

Let us consider Pearson's correlation coefficients as in [16,17], adjusted to the setting of directed graphs as in [9,20]. This will constitute four formulas that we combine into one. Take $\alpha, \beta \in \{+,-\}$, that is, we let α and β index the type of degree (out- or in-degree). Then we get the following expression for the four Pearson's correlation coefficients:

$$r_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)}\left(\frac{1}{|E|}\sum_{e\in E} D^\alpha(e_*)D^\beta(e^*) - \frac{1}{|E|^2}\sum_{e\in E} D^\alpha(e_*)\sum_{e\in E} D^\beta(e^*)\right), \quad (3.1)$$

where

$$\sigma_\alpha(G) = \sqrt{\frac{1}{|E|}\sum_{e\in E} D^\alpha(e_*)^2 - \frac{1}{|E|^2}\left(\sum_{e\in E} D^\alpha(e_*)\right)^2} \text{ and} \quad (3.2)$$

$$\sigma^\beta(G) = \sqrt{\frac{1}{|E|}\sum_{e\in E} D^\beta(e^*)^2 - \frac{1}{|E|^2}\left(\sum_{e\in E} D^\beta(e^*)\right)^2}. \quad (3.3)$$

Here we utilize the notations for the source and target of an edge by letting the superscript index denote the specific degree type of the target e^* and the subscript index the degree type of the source e_*. For instance r_+^- denotes the Pearson correlation coefficient for the Out/In relation.

It is convenient to rewrite the summations over edges to summations over vertices by observing that

$$\sum_{e\in E} D^\alpha(e_*)^k = \sum_{v\in V} D^+D^\alpha(v)^k,$$

and similarly

$$\sum_{e\in E} D^\alpha(e^*)^k = \sum_{v\in V} D^-D^\alpha(v)^k,$$

for all $k > 0$. Plugging this into (3.1)–(3.3) we arrive at the following definition.

Definition 3.1 *Let $G = (V, E)$ be a directed graph and let $\alpha, \beta \in \{+, -\}$. Then the Pearson's α-β correlation coefficient is defined by*

$$r_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \frac{1}{|E|} \sum_{e \in E} D^\alpha(e_*) D^\beta(e^*) - \hat{r}_\alpha^\beta(G), \tag{3.4}$$

where

$$\hat{r}_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \frac{1}{|E|^2} \sum_{v \in V} D^+(v) D^\alpha(v) \sum_{v \in V} D^-(v) D^\beta(v), \tag{3.5}$$

$$\sigma_\alpha(G) = \sqrt{\frac{1}{|E|} \sum_{v \in V} D^+(v) D^\alpha(v)^2 - \frac{1}{|E|^2} \left(\sum_{v \in V} D^+(v) D^\alpha(v) \right)^2}, \tag{3.6}$$

$$\sigma^\beta(G) = \sqrt{\frac{1}{|E|} \sum_{v \in V} D^-(v) D^\beta(v)^2 - \frac{1}{|E|^2} \left(\sum_{v \in V} D^-(v) D^\beta(v) \right)^2}. \tag{3.7}$$

Just as in the undirected case, c.f. [13, 14], the wiring of the network contributes only to the positive part of (3.4). All other terms are completely determined by the in- and out-degree sequences. This fact enables us to analyze the behavior of $r_\alpha^\beta(G)$, see Section 3.1. Observe also that in contrast to undirected graphs, in the directed case, the correlation between the in- and out-degrees of a vertex can play a role, take, for instance, $\alpha = -$ and $\beta = +$.

Note that, in general, $r_\alpha^\beta(G)$ might not be well defined, for either $\sigma_\alpha(G)$ or $\sigma^\beta(G)$ might be zero, for example, when G is a directed cyclic graph of arbitrary size. From (3.2) and (3.3) it follows that $\sigma_\alpha(G)$ and $\sigma^\beta(G)$ are the variances of X and Y, where $X = D^\alpha(e_*)$ and $Y = D^\beta(e^*)$, $e \in E$, with probability $1/|E|$. Thus, $\sigma_\alpha(G) \neq 0$ is only possible if $D^\alpha(v) \neq D^\alpha(w)$ for some $v, w \in V$. Moreover, v and w must have nonzero out-degree for at least one such pair v, w, so that $D^\alpha(v)$ and $D^\alpha(w)$ are counted when we traverse over edges. This argument is formalized in the next lemma, which provides necessary and sufficient conditions so that $\sigma_\alpha(G), \sigma^\beta(G) \neq 0$.

Lemma 3.2 *Let $G = (V, E)$ be a directed graph and take $\alpha, \beta \in \{+, -\}$. Then the following holds:*

$$\frac{1}{|E|} \left(\sum_{v \in V} D^\alpha(v) D^\beta(v) \right)^2 \leq \sum_{v \in V} D^\alpha(v) D^\beta(v)^2, \tag{3.8}$$

and strict inequality holds if and only if there exists distinct $v, w \in V$ such that $D^\alpha(v)$, $D^\alpha(w) > 0$ and $D^\beta(v) \neq D^\beta(w)$.

Proof. Recall that $|E| = \sum_{v \in V} D^\alpha(v)$ for any $\alpha \in \{+, -\}$. Then we have:

$$\begin{aligned} &|E| \sum_{v \in V} D^\alpha(v) D^\beta(v)^2 - \left(\sum_{v \in V} D^\alpha(v) D^\beta(v) \right)^2 \\ &= \sum_{w \in V} \sum_{v \in V \setminus w} D^\alpha(w) D^\alpha(v) D^\beta(v)^2 - D^\alpha(w) D^\beta(w) D^\alpha(v) D^\beta(v) \\ &= \frac{1}{2} \sum_{w \in V} \sum_{v \in V \setminus w} D^\alpha(w) D^\alpha(v) \left(D^\beta(w)^2 - 2 D^\beta(w) D^\beta(v) + D^\beta(v)^2 \right) \\ &= \frac{1}{2} \sum_{w \in V} \sum_{v \in V \setminus w} D^\alpha(w) D^\alpha(v) \left(D^\beta(w) - D^\beta(v) \right)^2 \geq 0, \end{aligned}$$

which proves (3.8). From the last line, one easily sees that strict inequality holds if and only if there exists distinct $v, w \in V$ such that $D^\alpha(v)$, $D^\alpha(w) > 0$, and $D^\beta(v) \neq D^\beta(w)$. □

3.1. Convergence of Pearson's Correlation Coefficients

In this section we will prove that Pearson's correlation coefficients (3.4), calculated on sequences of growing graphs satisfying rather general conditions, converge to a nonnegative value. We start by recalling the definition of big theta.

Definition 3.3 *Let $f, g : \mathbb{N} \to \mathbb{R}_{>0}$ be positive functions. Then $f = \Theta(g)$ if there exist $k_1, k_2 \in \mathbb{R}_{>0}$, and an $N \in \mathbb{N}$ such that for all $n \geq N$*

$$k_1 g(n) \leq f(n) \leq k_2 g(n).$$

When we have two sequences $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$, we write $a_n = \Theta(b_n)$ for $(a_n)_{n \in \mathbb{N}} = \Theta((b_n)_{n \in \mathbb{N}})$.

Next, we will provide the conditions that our sequence of graphs needs to satisfy to prove the result. These conditions are based on properties of i.i.d. sequences of regularly varying random variables, which are often used to model scale-free distributions. We will provide a more thorough motivation of the chosen conditions in Section 3.2. From here on, we denote by $x \vee y$ and $x \wedge y$ the maximum and minimum of x and y, respectively.

Definition 3.4 *For $\gamma_-, \gamma_+ \in \mathbb{R}_{>0}$ we denote by $\mathcal{G}_{\gamma_- \gamma_+}$ the space of all sequences of graphs $(G_n)_{n \in \mathbb{N}}$ with the following properties:*

G1 $|V_n| = n$.

G2 *There exists a $N \in \mathbb{N}$ such that for all $n \geq N$ there exist $v, w \in V_n$ with $D_n^\alpha(v)$, $D_n^\alpha(w) > 0$, and $D_n^\alpha(v) \neq D_n^\alpha(w)$, for all $\alpha \in \{+, -\}$.*

G3 *For all $p, q \in \mathbb{R}_{>0}$,*

$$\sum_{v \in V_n} D_n^+(v)^p D_n^-(v)^q = \Theta(n^{p/\gamma_+ \vee q/\gamma_- \vee 1}).$$

G4 *For all $p, q \in \mathbb{R}_{>0}$, if $p < \gamma_+$ and $q < \gamma_-$ then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{v \in V_n} D_n^+(v)^p D_n^-(v)^q := d(p, q) \in (0, \infty).$$

Where the limits are such that for all $a, b \in \mathbb{N}$, $k, m > 1$ with $1/k + 1/m = 1$, $a + p < \gamma_+$ and $b + q < \gamma_-$, we have

$$d(a, b)^{\frac{1}{m}} d(p, q)^{\frac{1}{k}} > d\left(\frac{a}{m} + \frac{p}{k}, \frac{b}{m} + \frac{q}{k}\right).$$

Now we are ready to give the convergence theorem for Pearson's correlation coefficients, Definition 3.1.

Theorem 3.5 *Let $\alpha, \beta \in \{+, -\}$. Then there exists an area $A_\alpha^\beta \subseteq \mathbb{R}^2$ such that for $(\gamma_+, \gamma_-) \in A_\alpha^\beta$ and $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}_{\gamma_-\gamma_+}$,*

$$\lim_{n\to\infty} \hat{r}_\alpha^\beta(G_n) = 0,$$

and, hence, any limit point of $r_\alpha^\beta(G_n)$ is nonnegative.

Proof. Let $(G_n)_{n\in\mathbb{N}}$ be an arbitrary sequence of graphs. It is clear that if $\hat{r}_\alpha^\beta(G_n) \to 0$ then any limit point of $r_\alpha^\beta(G_n)$ is nonnegative. Therefore, we need to prove only the first statement. To this end we define the following sequences,

$$a_n = \frac{1}{|E_n|}\left(\sum_{v\in V_n} D_n^+(v) D_n^\alpha(v)\right)^2, \qquad b_n = \frac{1}{|E_n|}\left(\sum_{v\in V_n} D_n^-(v) D_n^\beta(v)\right)^2,$$
$$c_n = \sum_{v\in V_n} D_n^+(v) D_n^\alpha(v)^2, \qquad d_n = \sum_{v\in V_n} D_n^-(v) D_n^\beta(v)^2,$$

and observe that $\hat{r}_\alpha^\beta(G_n)^2 = a_n b_n/(c_n - a_n)(d_n - b_n)$. Now if $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}_{\gamma_-\gamma_+}$, then because of G2 and Lemma 3.2 there exists an $N \in \mathbb{N}$ such that for all $n \geq N$ we have $c_n > a_n$ and $d_n > b_n$, so $\hat{r}_\alpha^\beta(G_n)$ is well defined for all $n \geq N$. Next, using G3, we get that $a_n = \Theta(n^a)$, $b_n = \Theta(n^b)$, $c_n = \Theta(n^c)$, and $d_n = \Theta(n^d)$ for certain constants a, b, c, and d, which depend on γ_-, γ_+ and the degree-degree correlation type chosen. Because $\hat{r}_\alpha^\beta(G_n) \to 0$ if and only if $\hat{r}_\alpha^\beta(G_n)^2 \to 0$, we need to find sufficient conditions for which $a_n b_n/(c_n - a_n)(d_n - b_n) \to 0$. It is clear that either $a < c$ and $b_n/(d_n - b_n)$ is bounded or $b < d$ and $a_n/(c_n - a_n)$ is bounded is sufficient. It turns out that this is exactly the case when either $a < c$ and $b \leq d$ or $a \leq c$ and $b < d$. We will do the analysis for the In/Out degree-degree correlation. The analysis for the other three correlation types is similar. Figure 2 shows all four areas A_α^β.

When $\alpha = -$ and $\beta = +$ we get the following constants:

$$a, b = 2\left(\frac{1}{\gamma_+} \vee \frac{1}{\gamma_-} \vee 1\right) - 1.$$
$$c = \left(\frac{1}{\gamma_+} \vee \frac{2}{\gamma_-} \vee 1\right).$$
$$d = \left(\frac{2}{\gamma_+} \vee \frac{1}{\gamma_-} \vee 1\right).$$

It is clear that when $1 < \gamma_-, \gamma_+ < 2$, then $a < c$ and $b < d$, and hence, $\hat{r}_\alpha^\beta \to 0$. Now, if $1 < \gamma_- < 2$ and $\gamma_+ \geq 2$, then $a = b = d = 1 < c$. Using G4 we get that $\lim_{n\to\infty} d_n/n = d(2, 1)$ and

$$\begin{aligned}
\lim_{n\to\infty} \frac{b_n}{n} &= \lim_{n\to\infty} \frac{\left(\sum_{v\in V_n} D_n^-(v)D_n^+(v)\right)^2}{n^2} \frac{n}{|E_n|} \\
&= \lim_{n\to\infty} \left(\frac{\sum_{v\in V_n} D_n^-(v)D_n^+(v)}{n}\right)^2 \left(\frac{\sum_{v\in V_n} D_n^-(v)}{n}\right)^{-1} \\
&= \frac{d(1, 1)^2}{d(0, 1)} < d(2, 1) = \lim_{n\to\infty} \frac{d_n}{n},
\end{aligned}$$

where, for the last part, we again used G4. From this it follows that $b_n/(d_n - b_n)$ is bounded and so $\hat{r}_\alpha^\beta \to 0$. A similar argument applies to the case $\gamma_- \geq 2$ and $1 < \gamma_+ < 2$, where the only difference is that $a = b = c = 1 < d$, hence,

$$A_-^+ = \{(x, y) \in \mathbb{R} | 1 < x < 2, \quad y > 1\} \cup \{(x, y) \in \mathbb{R} | 1 < y < 2, \quad x > 1\}.$$

Using similar arguments, we obtain

$$\begin{aligned}
A_+^- &= \{(x, y) \in \mathbb{R}^2 | 1 < x < 3, \quad y > 1\} \cup \{(x, y) \in \mathbb{R}^2 | 1 < y < 3, \quad x > 1\}, \\
A_+^+ &= \{(x, y) \in \mathbb{R}^2 | 1 < x < 3, \quad y > 1\}, \text{ and} \\
A_-^- &= \{(x, y) \in \mathbb{R}^2 | 1 < y < 3, \quad x > 1\}.
\end{aligned}$$

□

Let us now provide an intuitive explanation for the areas A_α^β, as depicted in Figure 2. The key observation is that because of G3, the terms with the highest power of either D_n^+ or D_n^- will dominate in $\hat{r}_\alpha^\beta(G_n)$. Therefore, if these moments do not exist, then the denominator will grow at a larger rate then the numerator, hence, $\hat{r}_\alpha^\beta \to 0$.

Taking $\alpha = + = \beta$, we see that D^- has terms only of order one whereas D^+ has terms up to order three. This explains why $A_+^+ = \{(x, y) \in \mathbb{R} | 1 < x \leq 3, y > 1\}$. Area A_-^- is then easily explained by observing that the expression for $r_-^-(G)$ is obtained from $r_+^+(G)$ by interchanging D^+ and D^-.

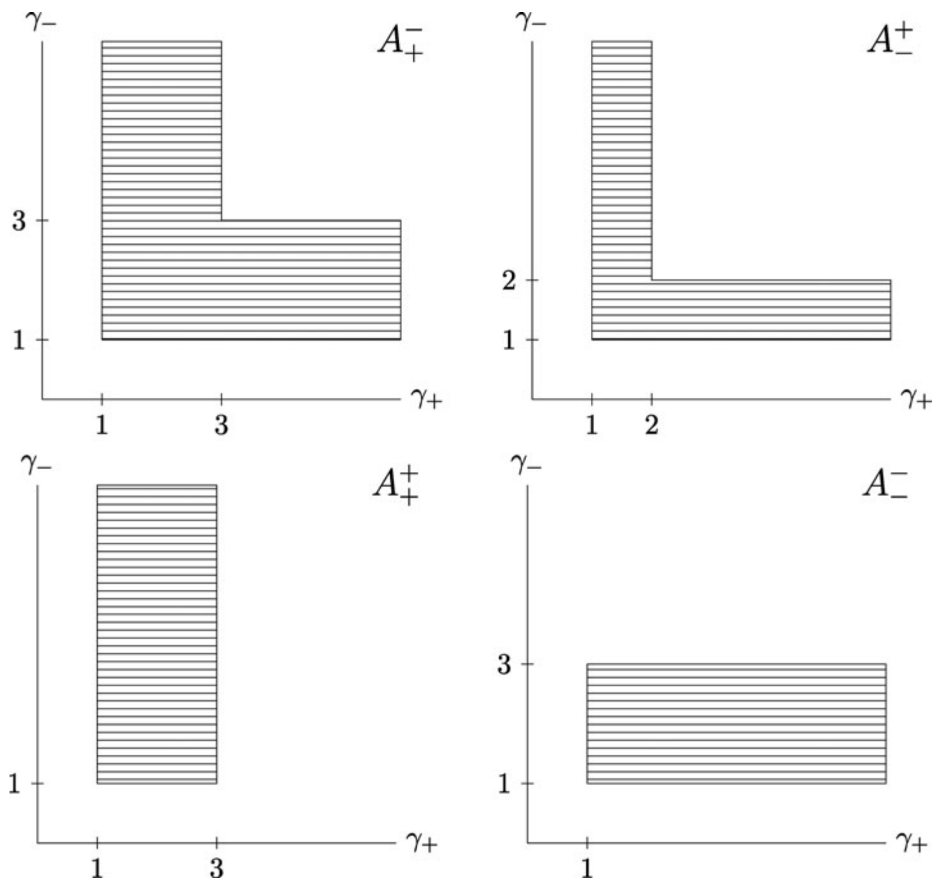

Figure 2 Four areas A_α^β, where r_α^β converges to a nonnegative number.

For the Out/In correlation, i.e., $\alpha = +$ and $\beta = -$, we see from (3.5)–(3.7) that $\hat{r}_+^-(G)$ splits into a product of two terms, each completely determined by either in- or out-degrees,

$$\frac{\frac{1}{|E|}\sum_{v\in V} D^\alpha(v)^2}{\sqrt{\frac{1}{|E|}\sum_{v\in V} D^\alpha(v)^3 - \frac{1}{|E|^2}\left(\sum_{v\in V} D^\alpha(v)^2\right)^2}},$$

with $\alpha \in \{+, -\}$. These terms are of the exact same form as the expression in [13] for the undirected degree-degree correlation. Because both D^+ and D^- have terms of order three, one sees that

$$A_+^- = \{(x, y) \in \mathbb{R}^2 | 1 < x < 3, \quad y > 1\} \cup \{(x, y) \in \mathbb{R}^2 | 1 < y < 3, \quad x > 1\}.$$

Now take an undirected network and make it directed by replacing each undirected edge with a bidirectional edge. Then $D^+(v) = D^-(v)$ for all $v \in V$ and hence, $r_+^-(G)$ equals the expression of (3.4) in [13] when we replace D by either D^+ or D^-.

Theorem 3.5 has several consequences. First, no matter what mechanism is used for generating networks, if the conditions of the theorem are satisfied, then for large enough networks the degree-degree correlations will always be nonnegative. This could explain why in most large networks strong disassortativity has not been registered. We will present such examples in Section 5. Second, if the underlying model that governs the topology of the network is in line with the conditions of the theorem, then one cannot compare networks of different sizes that arise from this model. For in this case, the degree-degree correlation coefficients r_α^β will decrease with the network size.

3.2. Motivation for $\mathcal{G}_{\gamma_- \gamma_+}$

In this section we will motivate Definition 3.4. G1 is easily motivated, for we want to consider infinite network size limits. G2 combined with Lemma 3.2 ensures that from a certain graph size N, $r_\alpha^\beta(G_n)$ is always well defined. Conditions G3 and G4 are related to heavy-tailed degree sequences that are modeled using regularly varying random variables.

A random variable X is called regularly varying with exponent γ if for all $t > 0$, $\mathbb{P}(X > t) = L(t)t^{-\gamma}$ for some slowly varying function L, that is $\lim_{t\to\infty} L(tx)/L(t) = 1$ for all $x > 0$. We write $\mathcal{R}_{-\gamma}$ for the class of all such distribution functions and write $X \in \mathcal{R}_{-\gamma}$ to denote a regularly varying random variable with exponent γ. For such a random variable X, we have that $\mathbb{E}[X^p] < \infty$ for all $0 < p < \gamma$.

Through experiments it has been shown that many real-world networks, both directed and undirected, have degree sequences whose distributions closely resemble a power-law distribution, c.f. Table II of [1] and [18]. Suppose we take two random variables $\mathcal{D}^+ \in \mathcal{R}_{\gamma_+}$, $\mathcal{D}^- \in \mathcal{R}_{\gamma_-}$ and consider, for each n, the degree sequences $(D_n^\pm(v))_{v\in V_n}$ as i.i.d. copies of these random variables. Then for all $0 < p < \gamma_+$ and $0 < q < \gamma_-$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q = \mathbb{E}\left[(\mathcal{D}^+)^p(\mathcal{D}^-)^q\right].$$

Moreover, since $\mathcal{D}^\pm$ is nondegenerate, we have $\mathbb{E}\left[(\mathcal{D}^\pm)^k\right] > \mathbb{E}\left[\mathcal{D}^\pm\right]^k$, and thus, by taking $d(p,q) = \mathbb{E}\left[(\mathcal{D}^+)^p(\mathcal{D}^-)^q\right]$, we get G4 where the second part follows from Hölder's inequality. Although i.i.d. sequences generated by sampling from in- and out-degree distributions do not in general constitute a graphical sequence, it is often the case that one can modify this sequence into a graphical sequence preserving i.i.d. properties asymptotically. Consider, for example, [6], where a directed version of the configuration model is introduced and it is proven (Theorem 2.4) that the degree sequences are asymptotically independent.

The property G3 is associated with the scaling of the sums $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ and is related to the central limit theorem for regularly varying random variables. When we model the degrees as i.i.d. copies of independent, regularly varying, random variables $\mathcal{D}^+ \in \mathcal{R}_{-\gamma_+}$, $\mathcal{D}^- \in \mathcal{R}_{-\gamma_-}$ and take $p \geq \gamma_+$ or $q \geq \gamma_-$, then $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ is in the domain of attraction of a γ-stable random variable $S(\gamma)$, where $\gamma = (\gamma_+/p \wedge \gamma_-/q)$, c.f. [7]. This means that

$$\frac{1}{a_n} \sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q \xrightarrow{d} S(\gamma_+/p \wedge \gamma_-/q), \quad \text{as } n \to \infty \tag{3.9}$$

for some sequence $a_n = \Theta(n^{q/\gamma_- \vee p/\gamma_+})$, where $\xrightarrow{d}$ denotes convergence in distribution. Informally, one could say that $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ scales as $n^{q/\gamma_- \vee p/\gamma_+}$ when either the p or q moment does not exist and as n when both moments exist, hence, $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$

scales as $n^{q/\gamma_- \vee p/\gamma_+ \vee 1}$, which is what G3 states. For completeness we include the next lemma, which shows that (3.9) implies that G3 holds with high probability.

We remark that, although the motivation for G3 is based on results where in the regularly varying random variables are assumed to be independent, the dependent case can be included. For this, one needs to adjust the scaling parameters in G3 for the specified dependence. In our numerical experiments, the in- and out- degrees in the Wikipedia graphs show strong independence, hence G3 holds for networks such as Wikipedia.

Lemma 3.6 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of positive random variables such that*

$$\frac{X_n}{a_n} \xrightarrow{d} X, \quad \text{as } n \to \infty,$$

for some sequence $(a_n)_{n\in\mathbb{N}}$ and positive random variable X. Then for each $0 < \varepsilon < 1$, there exists an $N_\varepsilon \in \mathbb{N}$ and $\kappa_\varepsilon \geq \ell_\varepsilon > 0$, such that for all $n \geq N_\varepsilon$,

$$\mathbb{P}(\ell_\varepsilon a_n \leq X_n \leq \kappa_\varepsilon a_n) \geq 1 - \varepsilon.$$

Proof. Let $0 < \varepsilon < 1$ and take $\delta > 0$, $0 < \ell \leq \kappa$ such that

$$\mathbb{P}(\ell \leq X \leq \kappa) \geq 1 - \varepsilon + \delta.$$

Then, because $X_n/a_n \xrightarrow{d} X$ as $n \to \infty$, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$,

$$|\mathbb{P}(\ell \leq X \leq \kappa) - \mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n)| < \delta.$$

Now we get for all $n \geq N$,

$$1 - \varepsilon + \delta - \mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n) \leq \mathbb{P}(\ell \leq X \leq \kappa) - \mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n) \leq \delta,$$

hence, $\mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n) \geq 1 - \varepsilon$. □

4. RANK CORRELATIONS

In this section we consider two other measures for degree-degree dependencies, Spearman's rho [22] and Kendall's tau [11], which are based on the rankings of the degrees rather than their actual value. We will define these dependency measures and argue that they do not have unwanted behavior as we observed for Pearson's correlation coefficients. We later use examples to enforce this argument and show that Spearman's rho and Kendall's tau are better candidates for measuring degree-degree dependencies.

4.1. Spearman's Rho

Spearman's rho [22] is defined as the Pearson correlation coefficient of the vector of ranks. Let $G = (V, E)$ be a directed graph and $\alpha, \beta \in \{+, -\}$. In order to adjust the definition of Spearman's rho to the setting of directed graphs, we need to rank the vectors $(D^\alpha(e_*))_{e\in E}$ and $(D^\beta(e^*))_{e\in E}$. These will, however, in general have many tied values. For instance, suppose that $D^\alpha(v) = m$ for some $v \in V$, then edges $e \in E$ with $e_* = v$ satisfy $D^\alpha(e_*) = D^\alpha(v)$. Therefore, we will encounter the value $D^\alpha(v)$ at least m times in the vector $(D^{\alpha}(e_*))_{e\in E}$. We will consider two strategies for resolving ties: uniformly at random (Section 4.1.1) and using an average ranking scheme (Section 4.1.2).

4.1.1. Resolving Ties Uniformly at Random. Given a sequence $\{x_i\}_{1\le i\le n}$ of distinct elements in $\mathbb{R}$, we denote by $R(x_j)$ the rank of x_j, i.e., $R(x_j) = |\{i|x_i \ge x_j\}|$, $1 \le j \le n$. The definition of Spearman's rho in the setting of directed graphs is then as follows.

Definition 4.1 *Let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{+, -\}$ and let $(U_e)_{e\in E}, (W_e)_{e\in E}$ be i.i.d. copies of independent uniform random variables U and W on $(0, 1)$, respectively. Then we define the α-β Spearman's rho of the graph G as*

$$\rho_{\alpha}^{\beta}(G) = \frac{12\sum_{e\in E} R^{\alpha}(e_*)R^{\beta}(e^*) - 3|E|(|E|+1)^2}{|E|^3 - |E|}, \tag{4.1}$$

where $R^{\alpha}(e_*) = R(D^{\alpha}(e_*) + U_e)$, and $R^{\beta}(e^*) = R(D^{\beta}(e^*) + W_e)$.

From (4.1) we see that the negative part of $\rho_{\alpha}^{\beta}(G)$ depends only on the number of edges

$$\frac{3(|E|+1)^2}{(|E|^2-1)} = 3 + \frac{6|E|+4}{|E|^2-1},$$

whereas, for $r_{\alpha}^{\beta}(G)$ it depended on the values of the degrees; see Definition 3.1. When $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}_{\gamma_+,\gamma_-}$, with $\gamma_+, \gamma_- > 1$, then it follows that $|E_n| = \theta(n)$, hence, $3 + (6|E| + 4)/(|E|^2 - 1) \to 3$, as $n \to \infty$. Therefore, we see that the negative contribution will always be at least 3, and so $\rho_{\alpha}^{\beta}(G_n)$ does not in general converge to a nonnegative number although $r_{\alpha}^{\beta}(G_n)$ does.

When calculating $\rho_{\alpha}^{\beta}(G)$ on a graph G, one has to be careful, for each instance will give different ranks of the tied values. This could potentially give rise to very different results among several instances, see Section 5.1.2 for an example. Therefore, in experiments, we will take an average of $\rho_{\alpha}^{\beta}(G)$ over several instances of the uniform ranking.

4.1.2. Resolving Ties with Average Ranking. A different approach for resolving ties is to assign the same average rank to all tied values. Consider, for example, the sequence $(1, 2, 1, 3, 3)$. Here the two values of 3 have ranks 1 and 2, but instead we assign the rank $3/2$ to both of them. With this scheme, the sequence of ranks becomes $(9/2, 3, 9/2, 3/2, 3/2)$. This procedure can be formalized as follows.

Definition 4.2 *Let $(x_i)_{1\le i\le n}$ be a sequence in $\mathbb{R}$; then we define the average rank of an element x_i as*

$$\overline{R}(x_i) = |\{j|x_j > x_i\}| + \frac{|\{j|x_j = x_i\}| + 1}{2}.$$

Observe that in the definition the total average rank is preserved: $\sum_{i=1}^{n} \overline{R}(x_i) = n(n+1)/2$. The difference with resolving ties uniformly at random is that we in general do not know $\sum_{i=1}^{n} \overline{R}(x_i)^2$, for this depends on how many ties we have for each value. We now define the corresponding version of Spearman's rho of graphs as follows.

Definition 4.3 *Let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{+, -\}$ and denote by $\overline{R}^{\alpha}(e_*)$ and $\overline{R}^{\beta}(e^*)$ the average ranks of $D^{\alpha}(e_*)$ among $(D^{\alpha}(e_*))_{e\in E}$ and $D^{\beta}(e^*)$ among $(D^{\beta}(e^*))_{e\in E}$, respectively. Then we define the α-β Spearman's rho with average resolution of ties by*

$$\overline{\rho}_{\alpha}^{\beta}(G) = \frac{4\sum_{e\in E}\overline{R}^{\alpha}(e_*)\overline{R}^{\beta}(e^*) - |E|(|E|+1)^2}{\overline{\sigma}_{\alpha}(G)\overline{\sigma}^{\beta}(G)}, \tag{4.2}$$

where

$$\overline{\sigma}_{\alpha}(G) = \sqrt{4\sum_{e\in E}\overline{R}^{\alpha}(e_*)^2 - |E|(|E|+1)^2},$$

and

$$\overline{\sigma}^{\beta}(G) = \sqrt{4\sum_{e\in E}\overline{R}^{\beta}(e^*)^2 - |E|(|E|+1)^2}.$$

Note that $\overline{\rho}_{\alpha}^{\beta}(G)$ does not suffer from any randomness in the ranking of the degrees. Hence, in contrast to (4.1), here we do not need to take an average over multiple instances. The next lemma shows that taking the expectation over the uniform ranking is actually equal to applying the average ranking scheme.

Lemma 4.4 *Let $G = (V, E)$ be a graph, $e \in E$ and $\alpha, \beta \in \{+, -\}$. Then*

(i) $\mathbb{E}[R^{\alpha}(e_*)] = \overline{R}^{\alpha}(e_*)$, $\quad \mathbb{E}\left[R^{\beta}(e^*)\right] = \overline{R}^{\beta}(e^*)$, *and*
(ii) $\mathbb{E}\left[R^{\alpha}(e_*)R^{\beta}(e^*)\right] = \overline{R}^{\alpha}(e_*)\overline{R}^{\beta}(e^*)$

Proof. We will prove only the first statement of (i). The proof for the second is similar.

(i) Since $R^{\alpha}(e_*) = R(D^{\alpha}(e_*) + U_e)$ and $(U_e)_{e\in E}$ are i.i.d. copies of a uniform random variable U on $(0, 1)$ we have that

$$\begin{aligned}
&\sum_{f\in E} I\{D^{\alpha}(f_*) = D^{\alpha}(e_*)\}\,\mathbb{E}\left[I\left\{U_f \geq U_e\right\}\right] \\
&= \sum_{f\in E} I\{D^{\alpha}(f_*) = D^{\alpha}(e_*)\}\left(I\{f = e\} + \frac{1}{2}I\{f \neq e\}\right) \\
&= \frac{1}{2}\sum_{f\in E} I\{D^{\alpha}(f_*) = D^{\alpha}(e_*)\} + \frac{1}{2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\left[R^\alpha(e_*)\right] &= \mathbb{E}\left[\sum_{f\in E} I\left\{D^\alpha(f_*)+U_f \geq D^\alpha(e_*)+U_e\right\}\right]\\
&= \sum_{f\in E} I\left\{D^\alpha(f_*) > D^\alpha(e_*)\right\}\\
&+ \sum_{f\in E} I\left\{D^\alpha(f_*) = D^\alpha(e_*)\right\}\mathbb{E}\left[I\left\{U_f \geq U_e\right\}\right]\\
&= \sum_{f\in E} I\left\{D^\alpha(f_*) > D^\alpha(e_*)\right\} + \frac{1}{2}\sum_{f\in E} I\left\{D^\alpha(f_*) = D^\alpha(e_*)\right\} + \frac{1}{2}\\
&= \overline{R}^\alpha(e_*).
\end{aligned}$$

(ii) By definition we have that

$$\begin{aligned}
R^\alpha(e_*)R^\beta(e^*) &= \sum_{f,g\in E} I\left\{D^\alpha(f_*) > D^\alpha(e_*)\right\} I\left\{D^\beta(g^*) > D^\beta(e^*)\right\}\\
&+ \sum_{f,g\in E} I\left\{D^\alpha(f_*) > D^\alpha(e_*)\right\} I\left\{D^\beta(g^*) = D^\beta(e^*)\right\} I\left\{W_g \geq W_e\right\}\\
&+ \sum_{f,g\in E} I\left\{D^\alpha(f_*) = D^\alpha(e_*)\right\} I\left\{U_f \geq U_e\right\} I\left\{D^\beta(g^*) > D^\beta(e^*)\right\}\\
&+ \sum_{f,g\in E} I\left\{D^\alpha(f_*) = D^\alpha(e_*)\right\} I\left\{D^\beta(g^*) = D^\beta(e^*)\right\}\\
&\times I\left\{U_f \geq U_e\right\} I\left\{W_g \geq W_e\right\}.
\end{aligned}$$

Therefore, because $(U_f)_{f\in E}$ and $(W_g)_{g\in E}$ are i.i.d. copies of independent uniform random variables U and W on (0, 1), respectively, the result follows by applying (i). □

From Lemma 4.4 we conclude that instead of calculating ρ_α^β several times and then taking the average, we can immediately apply the average ranking, which limits the total calculations to just one. Moreover, we have that

$$\mathbb{E}\left[\rho_\alpha^\beta(G)\right] = \frac{3\overline{\sigma}_\alpha\overline{\sigma}^\beta}{|E|^3-|E|}\overline{\rho}_\alpha^\beta(G), \tag{4.3}$$

which emphasizes that the difference between the uniform at random and average ranking scheme is determined by the number of ties in the degrees.

4.2. Kendall's Tau

Another common rank correlation is Kendall's tau [11], which measures the weighted difference between the number of concordant and discordant pairs of the joint observations $(x_i, y_i)_{1\leq i\leq n}$. More precisely, a pair (x_i, y_i) and (x_j, y_j) of joint observations is concordant if $x_i < x_j$ and $y_i < y_j$ or if $x_i > x_j$ and $y_i > y_j$. They are called discordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$.

Definition 4.5 *Let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{-, +\}$ and denote by $\mathscr{N}_c$ and $\mathscr{N}_d$, respectively, the number of concordant and discordant pairs among $\left(D^\alpha(e_*), D^\beta(e^*)\right)_{e\in E}$. Then we define the α-β Kendall's tau of G by*

$$\tau_\alpha^\beta(G) = \frac{2(\mathscr{N}_c - \mathscr{N}_d)}{|E|(|E|-1)}.$$

It might seem at first that τ does not suffer from ties. However, note that the numerator of τ includes only strictly concordant and discordant pairs, whereas the denominator is equal to the number of all possible pairs, regardless of the presence of ties. Hence, when the number of ties is large, the denominator may become much larger than the numerator, resulting in small, even vanishing in the graph size limit, values of τ_α^β. We will provide such an example in Section 5. Because, as discussed previously, the sequences $(D^\alpha(e_*))_{e\in E}$ and $\left(D^\beta(e^*)\right)_{e\in E}$ naturally have a large number of ties, we cannot expect $\tau_\alpha^\beta(G)$ to take very large (positive or negative) values. To address this issue, a weighted extension of Kendall's tau was very recently introduced [27]. This new measure also puts more emphasis on nodes with large in- or out-degrees.

5. BRIDGE GRAPH EXAMPLE

In this section we will provide a sequence of graphs to illustrate the difference between the four correlation measures in directed networks. We start with a deterministic sequence and later adapt this to a randomized sequence using regularly varying random variables.

5.1. A Deterministic In-Out Bridge Graph

Let $k, m \in \mathbb{N}_{>0}$, then we define the bridge graph $G(k, m) = (V(k, m), E(k, m))$, displayed in Figure 3a, as follows:

$$V(k,m) = v \cup w \cup \bigcup_{i=1}^{k} v_i \cup \bigcup_{j=1}^{m} w_j, \quad E(k,m) = g \cup \bigcup_{i=1}^{k} e_i \cup \bigcup_{j=1}^{m} f_j, \text{ where}$$

$$e_i = (v_i, v),\ f_j = (w, w_j) \text{ and } g = (v, w).$$

It follows that $|E(k, m)| = m + k + 1$. For the degrees of $G(k, m)$ we have:

$$\begin{aligned}
&D^+(v_i) = 1, && D^-(v_i) = 0, && \text{for all } 1 \le i \le k;\\
&D^+(w_j) = 0, && D^-(w_j) = 1, && \text{for all } 1 \le j \le m;\\
&D^+(v) = 1, && D^-(v) = k, && \\
&D^+(w) = m, && D^-(w) = 1. &&
\end{aligned}$$

Looking at the scatter plot of $(D^-(e_*), D^+(e^*))_{e\in E(k,m)}$, in Figure 4a, we see that the point (k, m) contributes toward a positive dependency whereas the points $(0, 1)$ and $(1, 0)$

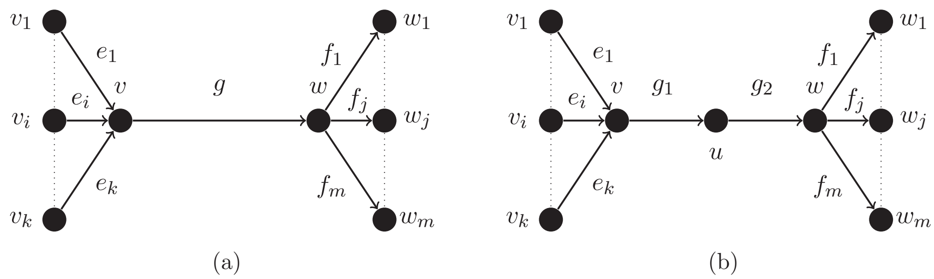

Figure 3 A graphical representation of the graphs (a) $G(k, m)$ and (b) $\hat{G}(k, m)$.

contribute toward a negative dependency. Hence, depending on how much weight we put on each of these points, we could argue equally well that this graph could have a positive or negative value for the In/Out dependency. We can, however, extend the in-out bridge graph to a graph for which we do have a clearly negative In/Out dependency.

We define the disconnected in-out bridge graph $\hat{G}(k, m) = (\hat{V}(k, m), \hat{E}(k, m))$ from $G(k, m)$ by adding a vertex u and replacing the edge $g = (v, w)$ by the edges $g_1 = (v, u)$ and $g_2 = (u, w)$; see Figure 3b. In this graph, the node with the largest in-degree, v, is connected to node u, of out-degree 1. Similarly, u, which has in-degree 1, is connected to the node with the highest out-degree, w. Therefore, we would expect a negative value of In/Out dependency measures. This intuition is supported by the scatter plot of $(D^+(e^*), D^-(e_*))_{e\in\hat{E}(k,m)}$ in Figure 4b.

Now consider for a fixed $a \in \mathbb{N}$, the sequence of graphs $G_n^a := G(n, an)$ and $\hat{G}_n^a := \hat{G}(n, an)$. Then, following the previous reasoning, we would expect any In/Out dependency measure of $\hat{G}_n^a$ to converge to -1.

In Sections 5.1.1–5.1.3 we will show that $\lim_{n\to\infty} r_-^+(\hat{G}_n^a) = 0$, whereas the other three measures indeed yield negative values. Furthermore, we show that $\lim_{n\to\infty} r_-^+(G_n^a) = 1$ although $\lim_{n\to\infty} \overline{\rho}_-^+(G_n^a) = -1$, reflecting the two possibilities for the In/Out correlation represented in the scatter plot in Figure 4a.

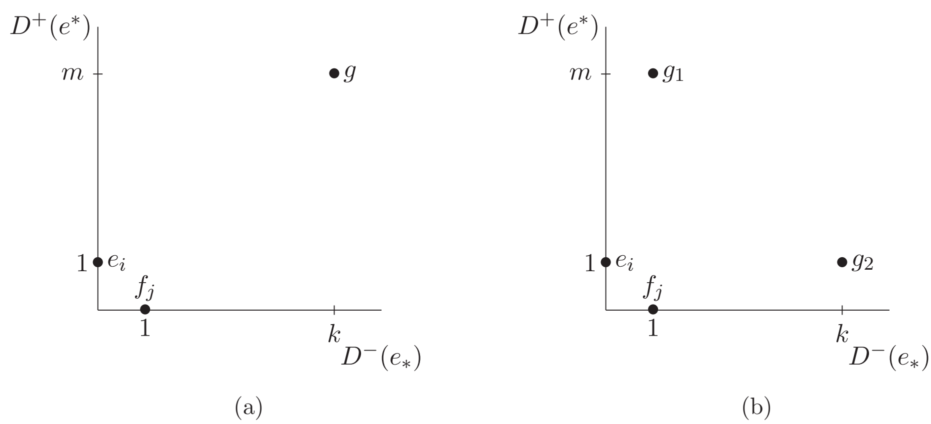

Figure 4 The scatter plots for the degrees of (a) $G(k, m)$ and (b) $\hat{G}(k, m)$.

5.1.1. Pearson In/Out Correlation. We start with the graph G_n^a. Basic calculations yield that

$$\sum_{e\in E_n^a} D^-(e_*)D^+(e^*) = an^2, \tag{5.1}$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v) = (1+a)n, \tag{5.2}$$

$$\sum_{v\in V_n^a} D^-(v)^2 D^+(v) = n^2 + an, \text{ and} \tag{5.3}$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v)^2 = n + a^2n^2, \tag{5.4}$$

hence, using (3.6) and (3.7), we obtain:

$$\begin{aligned}|E_n^a|\sigma_-(G_n^a) &= \sqrt{((1+a)n+1)(n^2+an) - (1+a)^2n^2}\\ &= \sqrt{(1+a)n^3 - (n-1)an},\end{aligned}$$

and

$$\begin{aligned}|E_n^a|\sigma^+(G_n^a) &= \sqrt{((1+a)n+1)(n+a^2n^2) - (1+a)^2n^2}\\ &= \sqrt{(1+a)n^3 - (an-1)n}.\end{aligned}$$

When we plug this into (3.4) with $\alpha = -$ and $\beta = +$, we get

$$\begin{aligned}r_-^+(G_n^a) &= \frac{|E_n^a|an^2 - (1+a)^2n^2}{|E_n^a|\sigma_\alpha(G_n^a)|E_n^a|\sigma^\beta(G_n^a)}\\ &= \frac{a(1+a)n^3 - (a^2+a+1)n^2}{a\sqrt{(1+a)n^3-(n-1)an}\sqrt{(1+a)n^3-(an-1)n}}.\end{aligned} \tag{5.5}$$

From (5.5) it follows that if $a \in \mathbb{N}$ is fixed, then $\lim_{n\to\infty} r_-^+(G_n^a) = 1$, thus $r_-^+(G_n^a)$ in fact reflects the connection between v and w where the point (n, an) in the scatter plot received the most mass. However, when we turn to $\hat{G}_n^a$ we get a less expected result. Splitting the edge g in two adds one to (5.2)–(5.4), while (5.1) becomes $(a+1)n$, which is linear in n instead of quadratic. Because all other terms keep their scale with respect to n, we easily deduce that for a fixed $a \in \mathbb{N}$, $\lim_{n\to\infty} r_-^+(\hat{G}_n^a) = 0$. This is undesirable for we would expect any In/Out correlation on $\hat{G}_n^a$ to converge to -1.

5.1.2. Spearman In/Out Correlation. We start by calculating $\overline{\rho}_-^+(G_n^a)$. For this observe that by (4.2) and the definition of G_n^a we have that,

$$\begin{aligned}\overline{R}^+((e_i)^*) &= 1 + \frac{n+1}{2}, & \overline{R}^-((e_i)_*) &= an + 1 + \frac{n+1}{2};\\ \overline{R}^+((f_j)^*) &= n + 1 + \frac{an+1}{2}, & \overline{R}^-((f_j)_*) &= 1 + \frac{an+1}{2};\\ \overline{R}^+(g^*) &= 1, & \overline{R}^-(g_*) &= 1.\end{aligned}$$

After some basic calculations, we get

$$\overline{\rho}_-^+(G_n^a) = \frac{-(a^2+a)n^3 + (a+1)^2n^2 + (a+1)n}{(a^2+a)n^3 + (a+1)^2n^2 + (a+1)n} \to -1 \quad \text{as } n \to \infty.$$

This result is in striking contrast with $r_-^+(G_n^a)$. Indeed, $\overline{\rho}_-^+$ places all the weight on the points (0, 1) and (1, 0). However, based on the scatter plot, see Figure 4a, both results could be plausible.

Let us now compute $\overline{\rho}_-^+(\hat{G}_n^a)$. For the rankings we have

$$\begin{aligned}
\overline{R}^+((e_i)^*) &= 2 + \frac{n}{2}, & \overline{R}^-((e_i)_*) &= an + 2 + \frac{n+1}{2};\\
\overline{R}^+((f_j)^*) &= n + 2 + \frac{an+1}{2}, & \overline{R}^-((f_j)_*) &= 2 + \frac{an}{2};\\
\overline{R}^+((g_1)^*) &= 2 + \frac{n}{2}, & \overline{R}^-((g_1)_*) &= 1;\\
\overline{R}^+((g_2)^*) &= 1, & \overline{R}^-((g_2)_*) &= 2 + \frac{an}{2}.
\end{aligned}$$

Filling this into (4.2) we get

$$\overline{\rho}_-^+(\hat{G}_n^a) = \frac{-(a^2+a)n^3 - (a^2+a)n^2 + (a+1)n - 2}{\bar{\sigma}_-(\hat{G}_n^a)\bar{\sigma}^+(\hat{G}_n^a)},$$

where

$$\begin{aligned}
\bar{\sigma}_-(\hat{G}_n^a) &= \sqrt{(a^2+a)n^3 + (a^2+4a+2)n^2 + (3a+4)n - 2} \text{ and}\\
\bar{\sigma}^+(\hat{G}_n^a) &= \sqrt{(a^2+a)n^3 + (2a^2+4a+1)n^2 + (4a+3)n + 2}.
\end{aligned}$$

Because

$$\lim_{n\to\infty} \frac{1}{n^3}\bar{\sigma}_-(\hat{G}_n^a)\bar{\sigma}^+(\hat{G}_n^a) = (a^2+a),$$

it follows that

$$\lim_{n\to\infty} \overline{\rho}_-^+(\hat{G}_n^a) = \lim_{n\to\infty} \frac{1/n^3}{1/n^3}\left(\frac{-(a^2+a)n^3 - (a^2+a)n^2 + (a+1)n - 2}{\bar{\sigma}_-(\hat{G}_n^a)\bar{\sigma}^+(\hat{G}_n^a)}\right) = -1,$$

which equals $\lim_{n\to\infty} \overline{\rho}_-^+(G_n^a)$. We have already argued that based on the graph and the scatter plot we would expect negative In/Out correlation for the sequence $(\hat{G}_n^a)_{n\in\mathbb{N}}$. This result is in agreement with what we would expect, while $r_-^+(\hat{G}_n^a)$ converges to 0 as $n \to \infty$.

Now we turn to $\rho_-^+(G_n^a)$. We show that the choice of ranking of the tied values can have a great effect on the outcome of the Spearman's In/Out correlation. In this example we pick two rankings, one will yield $\rho_-^+(G_n^a) > 0$, and the other will give $\rho_-^+(G_n^a) < 0$.

It is clear from the definition of G_n^a that the in- and out-degrees of all e_i are the same, and this is also true for f_j. Let us now impose the following ranking of the vectors

$(D^+(e^*))_{e\in E_n^a}$ and $(D^-(e_*))_{e\in E_n^a}$:

$$\begin{aligned} &R^+((e_i)^*) = an + i, && R^-((e_i)_*) = i, && \text{for all } 1 \le i \le n;\\ &R^+((f_j)^*) = j, && R^-((f_j)_*) = n + j, && \text{for all } 1 \le j \le an;\\ &R^+(g^*) = 1 + (a+1)n, && R^-(g_*) = 1 + (a+1)n. && \end{aligned}$$

Here we ordered the ties by the order of their indices. We calculate that

$$\rho_-^+(G_n^a) = \frac{(a^3 - 3a^2 - 3a + 1)n^3 + 3(a+1)^2n^2 + 2(a+1)n}{(a^3 + 3a^2 + 3a + 1)n^3 + 3(a+1)^2n^2 + 2(a+1)n}. \tag{5.6}$$

Let us now order $(D^+(e^*))_{e\in E_n^a}$ and $(D^-(e_*))_{e\in E_n^a}$ as follows:

$$\begin{aligned} &R^+((e_i)^*) = (a+1)n + 1 - i, && R^-((e_i)_*) = i, && \text{for all } 1 \le i \le n;\\ &R^+((f_j)^*) = an + 1 - j, && R^-((f_j)_*) = n + j, && \text{for all } 1 \le j \le an;\\ &R^+(g^*) = 1 + (a+1)n, && R^-(g_*) = 1 + (a+1)n. && \end{aligned}$$

This order differs from the first only on the vector $(D^+(e^*))_{e\in E_n^a}$, where we now ordered the ties based on the reversed order of their indices. Here we get, after some calculations,

$$\rho_-^+(G_n^a) = \frac{-(a+1)^3n^3 + 3(a+1)^2n^2 + 2(a+1)n}{(a+1)^3n^3 + 3(a+1)^2n^2 + 2(a+1)n}. \tag{5.7}$$

When we compare (5.7) with (5.6), we see that for the former $\lim_{n\to\infty} \rho_-^+(G_n^a) = -1$ for all $a \in \mathbb{N}$, and for the latter we have $\lim_{n\to\infty} \rho_-^+(G_n^a) = (a^3 - 3a^2 - 3a + 1)/(a+1)^3$. This means that increasing a will actually increase the limit of (5.6), which becomes positive when $a \ge 4$. If we denote by d_n^a the absolute value of the difference between (5.6) and (5.7), we get that $\lim_{n\to\infty} d_n^a = 2(a^3 + 1)/(a+1)^3$, which converges to 2 as $a \to \infty$. This agrees with the fact that for (5.6) it holds that $\lim_{a\to\infty} \lim_{n\to\infty} \rho_-^+(G_n^a) = 1$, whereas $\lim_{a\to\infty} \lim_{n\to\infty} \rho_-^+(G_n^a) = -1$ for (5.7). We see that changing the order of the ties can have a large impact on the value of $\rho_\alpha^\beta(G)$, as mentioned in Section 4.1.1. Now, using (4.3), $\lim_{n\to\infty} \overline{\rho}_-^+(G_n^a) = -1$ and the fact that

$$\overline{\sigma}_\alpha(G_n^a)\overline{\sigma}^\beta(G_n^a) = (a^2 + a)n^3 + (a+1)^2n^2 + (a+1)n,$$

we get that $\lim_{n\to\infty} \mathbb{E}\left[\rho_-^+(G_n^a)\right] = -2a/(a+1)^2$. Notice that, unlike $\overline{\rho}_-^+(G_n^a)$, this result still depends on a and converges to 0 as $a \to \infty$. This is not unexpected because the majority of edges produce ties, hence, most of the ranks are defined by independent realizations of U and W. These results indicate that Spearman's rho with average resolution of ties is the most informative correlation for this graph.

5.1.3. Kendall's Tau In/Out Correlation. In order to compute Kendall's Tau, we need to determine the number of concordant and discordant pairs. Starting with G_n^a, we observe that we have three kinds of joint observations:

$$\begin{aligned} I &: \left(D^-(e_{i*}), D^+(e_i^*)\right),\\ II &: \left(D^-(f_{j*}), D^+(f_j^*)\right) \text{ and}\\ III &: \left(D^-(g_*), D^+(g^*)\right). \end{aligned}$$

The combinations I and III, and II and III are concordant whereas I and II are discordant. It follows that $\mathcal{N}_c = (a+1)n$ while $\mathcal{N}_d = an^2$. Hence we get (see Definition 4.5),

$$\tau_-^+(G_n^a) = \frac{2(a+1)n - 2an^2}{(a+1)^2n^2 + (a+1)n},$$

which gives $\lim_{n\to\infty} \tau_-^+(G_n^a) = -\frac{2a}{(a+1)^2}$. We observe that this equals $\lim_{n\to\infty} \mathbb{E}\left[\rho_-^+(G_n^a)\right]$, calculated in the previous section.

For the graph $\hat{G}_n^a$ we have four kinds of joint observations:

$$\begin{aligned} I &: \left(D^-(e_{i*}), D^+(e_i^*)\right), \\ II &: \left(D^-(f_{j*}), D^+(f_j^*)\right), \\ III &: \left(D^-(g_{1*}), D^+(g_1^*)\right) \text{ and} \\ IV &: \left(D^-(g_{2*}), D^+(g_2^*)\right). \end{aligned}$$

Again the combinations I and II are discordant, although now I and III, and II and IV are concordant. Therefore, we get $\mathcal{N}_c = (a+1)n$ and $\mathcal{N}_d = an^2$, hence $\lim_{n\to\infty} \tau_-^+(\hat{G}_n^a) = -\frac{2a}{(a+1)^2}$, which equals the limit for $\tau_-^+(G_n^a)$. This is because the tied values, which are the majority in this example, make the influence of the extra node on Kendall's tau negligible.

Note that $\lim_{n\to\infty} \tau_-^+(G_n^a)$ decreases when we increase a. This is because the number of tied values among the degrees increases with a. We already mentioned that τ_α^β gives smaller values when more ties are involved. Here this behavior is clearly present.

5.2. A Collection of Random In/Out Bridge Graphs

Let us now consider a collection of In/Out bridge graphs $G(W, Z)$ as defined in Section 5.1, where the values of W and Z are integer valued regularly varying random variables.

Let $X, Y \in \mathcal{R}_{-\gamma}$ be independent and integer valued and fix $a \in \mathbb{R}_{>0}$. For each $n \in \mathbb{N}$, take $(X_i)_{1\le i\le n}$ and $(Y_i)_{1\le i\le n}$ to be i.i.d. copies of X and Y, respectively, and define $W_i = X_i + Y_i$ and $Z_i = \lfloor X_i + aY_i \rfloor$. Then we define the graph $\mathcal{G}_n^a$ as a disconnected collection of the graphs $(G(W_i, Z_i))_{1\le i\le n}$. We will calculate $r_-^+(\mathcal{G}_n^a)$ and prove that it converges to a random variable, which can have support on $(\varepsilon, 1)$ for any $\varepsilon \in (0, 1]$ depending on a specific choice of a.

Using the calculations in Section 5.1.1 we obtain

$$\sum_{e\in E_n^a} D^-(e_*)D^+(e^*) = \sum_{i=1}^n \left(X_i^2 + aY_i^2 + (1+a)X_iY_i\right),$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v) = \sum_{i=1}^n (2X_i + (1+a)Y_i),$$

$$\sum_{v\in V_n^a} D^-(v)^2D^+(v) = \sum_{i=1}^n \left(X_i^2 + Y_i^2 + 2X_iY_i + X_i + aY_i\right),$$

$$\sum_{v \in V_n^a} D^-(v) D^+(v)^2 = \sum_{i=1}^{n} \left(X_i^2 + a^2 Y_i^2 + 2a X_i Y_i + X_i + Y_i \right) \text{ and}$$

$$|E_n^a| = \sum_{i=1}^{n} (2X_i + (1+a)Y_i + 1) .$$

By the stable limit law we have a sequence $(a_n)_{n \in \mathbb{N}}$ such that

$$\frac{1}{a_n} \sum_{i=1}^{n} X_i^2 \xrightarrow{d} S_X \quad \text{and} \quad \frac{1}{a_n} \sum_{i=1}^{n} Y_i^2 \xrightarrow{d} S_Y \quad \text{as } n \to \infty,$$

where S_X and S_Y are stable random variables. Further, from Lemma 2.2 in [13] we have

$$\frac{1}{a_n} \sum_{i=1}^{n} X_i Y_i \xrightarrow{d} 0, \quad \frac{1}{a_n} \sum_{i=1}^{n} X_i \xrightarrow{d} 0 \quad \text{and} \quad \frac{1}{a_n} \sum_{i=1}^{n} Y_i \xrightarrow{d} 0 \quad \text{as } n \to \infty.$$

Combining this we get

$$\frac{1}{\sqrt{a_n}} \sigma_-(\mathcal{G}_n^a) \xrightarrow{d} \sqrt{S_X + S_Y}, \quad \frac{1}{\sqrt{a_n}} \sigma^+(\mathcal{G}_n^a) \xrightarrow{d} \sqrt{S_X + a^2 S_Y} \quad \text{as } n \to \infty,$$

and hence,

$$r_-^+(\mathcal{G}_n^a) \xrightarrow{d} \frac{S_X + a S_Y}{\sqrt{S_X + S_Y}\sqrt{S_X + a^2 S_Y}} \quad \text{as } n \to \infty,$$

which has support on (0, 1). Now, take $0 < \varepsilon \leq 1$ and consider the function $f(x) : (0, \infty) \to \mathbb{R}$ defined as

$$f(x) = \frac{1 + ax}{\sqrt{1+x}\sqrt{1+a^2 x}}.$$

This function attains its minimum in $1/a$ and by solving $f(1/a) = \varepsilon$ for a, we get that for

$$a = \frac{2 - \varepsilon^2 \pm \sqrt{1 - \varepsilon}}{\varepsilon^2},$$

this minimum equals ε. If we now introduce the random variable $T = S_Y / S_X$ we see that for a defined as previously, $\frac{1+aT}{\sqrt{1+T}\sqrt{1+a^2 T}}$ has support contained in $(\varepsilon, 1)$.

This example shows that Pearson's correlation coefficients r_α^β can converge to a nonnegative random variable in the infinite size network limit. This behavior is undesirable, for if we consider two instances of the same model $\mathcal{G}_n^a$, then the values of r_-^+ will be random and, hence, could be very far apart. Therefore r_-^+ is not suitable for measuring the In/Out correlation if we would like to find one number (population value) that characterizes the In/Out correlation in this model.

6. EXPERIMENTS

In this section we present experimental results for the degree-degree correlations introduced in Sections 3 and 4. For the calculations we used the WebGraph framework [3,4] and the fastutil package from The Laboratory for Web Algorithmics (LAW) at the Universitá degli studi di Milano.[1] The calculations were executed on the Wikepedia graphs[2] of nine different languages, obtained from the LAW dataset database. For each Wikipedia graph we calculated all four degree-degree correlations using the four measures introduced in this article.

The in- and out-degree distributions of these networks satisfy conditions of scale-free distributions with parameters between 1 and 2.5. Moreover, we evaluated the dependency between in- and out-degrees of the vertices, using angular measure [21, p. 313], and found them to be independent. Therefore, one could consider the Wikipedia networks as being generated by a model satisfying the conditions of Definition 3.4.

In an attempt to quantify the results, we compared them to a randomized setting. For this we did 20 reconfigurations of the degree sequences of each graph, using the scheme described in Section 4.2 of [6]. More precisely, we used the *erased directed configuration model*. In this scheme we first assign to each vertex v, $D^+(v)$ outbound stubs and $D^-(v)$ inbound stubs. Then we randomly select an available outbound stub and combine it with a inbound stub, selected uniformly at random from all available inbound stubs, to make an edge. When this edge is a self-loop we remove it. When we end up with multiple edges between two vertices, we combine them into one edge. Proposition 4.2 of [6] now tells us that the distribution of the degrees of the resulting simple graph will, with high probability, be the the same as the original distribution. For each of these reconfigurations, all four types of degree-degree dependencies were evaluated using the four measures discussed above, and then for each dependency type and each measure, we took the average. The results are presented in Table I.

The first observation is that for each Wikipedia graph and dependency type, the measures ρ, $\overline{\rho}$, and τ have the same sign whereas r in many cases has a different sign. Furthermore, there are many cases in which the absolute value of the three rank correlations is at least an order of magnitude larger than that of Pearson's correlation coefficients. See, for instance, the Out/In correlations for DE, EN, FR, and NL or the In/Out correlation for KO and RU.

These examples illustrate the fact that Pearson's correlation coefficients are scaled down by the high variance in the degree sequences, which in turn gave rise to Theorem 3.5, while the rank correlations do not have this deficiency. Another interesting observation is that the values for ρ and $\overline{\rho}$ are almost in full agreement with each other. This would then suggest that, looking back at (4.3), $3\overline{\sigma}_\alpha\overline{\sigma}^\beta \approx |E|^3 - |E|$ for the Wikipedia networks. Therefore, one could freely change between these two when calculating degree-degree correlations. Note that ρ is somewhat computationally easier than $\overline{\rho}$ because there is no need to compute $\overline{\sigma}_\alpha\overline{\sigma}^\beta$.

Finally, we notice that, for the configuration model instances of the graphs, all correlation measures are close to zero, and the difference between different realizations of the model is remarkably small (see the values of σ). However, at this point very little can be said about statistical significance of these results because, as we proved above, r shows pathological behavior on large power-law graphs and the setting of directed graphs is very different from the setting of independent observations. This raises important and challenging questions for future research: which magnitude of degree-degree dependencies

[1]http://law.di.unimi.it
[2]http://wikipedia.org

Table I Degree-degree correlations for Wikipedia graphs. The data in the columns Randomized correspond to the results for the reconfigurations of the given Wikipedia network.

		Pearson			Spearman Uniform			Spearman Average			Kendall		
			Randomized			Randomized			Randomized			Randomized	
Graph	α/β	**Data**	μ	σ	**Data**	μ	σ	**Data**	μ	σ	**Data**	μ	σ
DE wiki	+/−	−0.0552	−0.0178	0.0001	−0.1434	−0.0059	0.0002	−0.1435	−0.0059	0.0002	−0.0986	−0.0038	0.0008
	−/+	0.0154	−0.0030	0.0002	0.0481	−0.0008	0.0002	0.0484	−0.0008	0.0002	0.0.326	−0.0005	0.0001
	+/+	−0.0323	−0.0091	0.0002	−0.0640	−0.0048	0.0002	−0.0640	−0.0048	0.0002	−0.0446	−0.0006	0.0001
	−/−	−0.0123	−0.0060	0.0001	0.0119	−0.0009	0.0002	0.0120	−0.0009	0.0002	0.0074	−0.0032	0.0001
EN wiki	+/−	−0.0557	−0.0180	0	−0.1999	−0.0064	0.0001	−0.1999	−0.0064	0.0001	−0.1364	−0.0043	0.0001
	−/+	−0.0007	−0.0015	0.0001	0.0239	−0.0011	0.0001	0.0240	−0.0011	0.0001	0.0163	−0.0008	0.0001
	+/+	−0.0713	−0.0125	0.0001	−0.0855	−0.0053	0.0001	−0.0855	−0.0053	0.0001	−0.0581	−0.0035	0.0001
	−/−	−0.0074	−0.0024	0.0001	−0.0664	−0.0013	0.0001	−0.0666	−0.0013	0.0001	−0.0457	−0.0009	0.0001
ES wiki	+/−	−0.1031	−0.0336	0.0002	−0.1429	−0.0186	0.0003	−0.1429	−0.0186	0.0003	−0.0972	−0.0126	0.0002
	−/+	−0.0033	−0.0071	0.0002	−0.0407	−0.0047	0.0003	−0.0417	−00048	0.0003	−0.0294	−0.0034	0.0002
	+/+	−0.0272	−0.0201	0.0002	0.0178	−0.0125	0.0003	0.0178	−0.0125	0.0003	0.0119	−0.0084	0.0002
	−/−	−0.0262	−0.0116	0.0001	−0.1627	−0.0071	0.0003	−0.1669	−0.0072	0.0003	−0.1174	−0.0051	0.0002
FR wiki	+/−	−0.0536	−0.0252	0.0001	−0.1065	−0.0123	0.0002	−0.1065	−0.0123	0.0002	−0.0720	−0.0083	0.0002
	−/+	0.0048	−0.0031	0.0002	0.0119	−0.0016	0.0003	0.0121	−0.0016	0.0003	0.0085	−0.0011	0.0002
	+/+	−0.0512	−0.0173	0.0002	−0.0126	−0.0093	0.0002	−0.0126	−0.0090	0.0015	−0.0087	−0.0063	0.0001
	−/−	−0.0094	−0.0054	0.0001	−0.0262	−0.0021	0.0003	−0.0267	−0.0025	0.0015	−0.0186	−0.0015	0.0002
HU wiki	+/−	−0.1048	−0.0378	0.0003	−0.1280	−0.0220	0.0006	−0.1280	−0.0220	0.0006	−0.0877	−0.0148	0.0004
	−/+	0.0120	−0.0056	0.0005	0.0525	0.0002	0.0005	0.0595	0	0.0006	0.0442	0	0.0004
	+/+	−0.0579	−0.0261	0.0005	−0.0207	−0.0157	0.0005	−0.0207	−0.0157	0.0004	−0.0140	−0.0107	0.0003
	−/−	−0.0279	−0.0084	0.0004	0.0051	0.0004	0.0005	0.0060	0.0002	0.0006	0.0050	−0.0001	0.0005
IT wiki	+/−	−0.0711	−0.0319	0.0001	−0.0964	−0.0158	0.0002	−0.0964	−0.0158	0.0002	−0.0653	−0.0106	0.0002
	−/+	0.0048	−0.0031	0.0002	0.0468	−0.0013	0.0002	0.0469	−0.0013	0.0003	0.0319	−0.0009	0.0002
	+/+	−0.0704	−0.0204	0.0002	−0.0277	−0.0121	0.0002	−0.0277	−0.0122	0.0002	−0.0189	−0.0081	0.0001
	−/−	−0.0115	−0.0050	0.0001	−0.0428	−0.0016	0.0002	−0.0429	−0.0016	0.0002	−0.0296	−0.0011	0.0002
KO wiki	+/−	−0.0805	−0.0562	0.0004	−0.2696	−0.0476	0.0037	−0.2722	−0.0482	0.0038	−0.1985	−0.0328	0.0073
	−/+	0.0157	−0.0009	0.0030	0.1760	0.0019	0.0046	0.2323	0.0034	0.0046	0.1902	0.0031	0.0035
	+/+	−0.1697	−0.0357	0.0035	0.0016	−0.0267	0.0041	0.0191	−0.0272	0.0040	0.0170	0.0298	0.0415
	−/−	−0.0138	−0.0034	0.0015	−0.0493	0.0062	0.0045	−0.0618	0.0083	0.0042	−0.0463	0.0065	0.0032
NL wiki	+/−	−0.0585	−0.0346	0.0001	−0.3017	−0.0211	0.0002	−0.3018	−0.0211	0.0002	−0.2089	−0.0142	0.0002
	−/+	0.0100	−0.0025	0.0003	0.0727	−0.0007	0.0003	0.0730	−0.0007	0.0003	0.0504	−0.0004	0.0003
	+/+	−0.0628	−0.0194	0.0001	0.0016	−0.0104	0.0003	0.0016	−0.0104	0.0003	0.0015	−0.0070	0.0002
	−/−	−0.0233	−0.0091	0.0001	−0.1498	−0.0019	0.0003	−0.1505	−0.0019	0.0003	−0.1048	−0.0013	0.0002
RU wiki	+/−	−0.0911	−0.0225	0.0004	−0.1080	−0.0093	0.0015	−0.1084	−0.0093	0.0015	−0.0755	−0.0064	0.0010
	−/+	0.0398	−0.0006	0.0009	0.1977	0	0.0008	0.2200	0.0001	0.0009	0.1655	0.0001	0.0007
	+/+	0.0082	−0.0038	0.0010	0.2472	0.0002	0.0015	0.2480	0.0001	0.0015	0.1736	0.0001	0.0010
	−/−	−0.0242	−0.0030	0.0007	0.0236	0.0009	0.0011	0.0255	0.0007	0.0015	0.0187	0.0006	0.0007

should be seen as significant and how to construct mathematically sound statistical tests for establishing such significant dependencies.

7. DISCUSSION

From Theorem 3.5 and the examples in Section 5, it is clear that Pearson's correlation coefficients have undesirable properties, based on their limiting behavior when the graph size goes to infinity. The question of whether or not rank correlations converge to correct population values in infinite graph size limit, has not been addressed in this study, but it can be already answered affirmatively. For undirected graphs, it has been proved in [13], and the results for directed graphs are the subject of our current research and will be presented in our upcoming paper [25]. This provides sufficient motivation for using such rank correlation measures instead of Pearson's correlation coefficients for measuring degree-degree dependencies in directed networks with heavy-tailed degrees.

Nevertheless, we have also seen that, when using rank correlations, one needs to be careful when resolving the ties among the degrees. Furthermore, Spearman's rho and Kendall's tau make very skewed distributions uniform, thus they do not detect the influence of important hubs, as we saw in the example of the G_n^a graph in Section 5.1. Possibly, these measures should be considered in combination with measures for extremal dependencies, such as angular measure. Angular measure for two vectors $(X_i)_{i=1,\ldots,n}$ and $(Y_i)_{i=1,\ldots,n}$ is a rank correlation measure that characterizes whether X_i and Y_i tend to attain extremely large values simultaneously. We used this measure to verify the independence between in- and out- degrees of a node in Wikipedia graphs.

There is also an intriguing question of whether the four types of dependencies are related to one another. For instance, it is reasonable to think that if the Out/In and Out/Out correlations are highly positive, then the other two must also be (highly) positive. Indeed, if we take a node v with high in-degree, then it tends to have nodes of high out-degree connecting to it. Hence, out-degree of v tends to be high as well because of the high positive Out/Out dependency. Therefore, if v connects to another node w, then w tends to have large in- and out-degree, implying positive In/In and In/Out dependencies. It is very interesting to understand what the feasibility bounds are for possible combinations of the four dependency types in terms of different correlation measures.

Finally, although the results from percolation theory and the analysis of network stability under attack give some insights to the impact of degree assortativity, it remains an open question of what specific values of degree-degree correlation measures mean for the topology of directed networks in general. This shows that there are still many fundamental questions regarding degree-degree correlations in scale-free directed graphs.

FUNDING

This work is supported by the EU-FET Open grant NADINE (288956).

REFERENCES

[1] R. Albert and A.-L. Barabási. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74:1(2002), 47.

[2] M. Boguná and R. Pastor-Satorras. "Epidemic Spreading in Correlated Complex Networks." *Physical Review E* 66:4(2002), 047104.

[3] P. Boldi and S. Vigna. "The Webgraph Framework I: Compression Techniques." In *Proceedings of the 13th International Conference on the World Wide Web*, pp. 595–602. ACM, 2004.

[4] P. Boldi and S. Vigna. "The Webgraph Framework II: Codes For The World-Wide Web." In *Proceedings of the Data Compression Conference*, 2004, p. 528. IEEE, 2004.

[5] M. Brede and S. Sinha. "Assortative Mixing by Degree Makes a Network More Unstable." arXiv preprint cond-mat/0507710, 2005.

[6] N. Chen and M. Olvera-Cravioto. "Directed Random Graphs with Given Degree Distributions." *Stochastic Systems* 3:1(2013), 147–186.

[7] D. B. H. Cline. "Convolution Tails, Product Tails and Domains of Attraction." *Probability Theory and Related Fields* 72:4(1986), 529–557.

[8] S. de Franciscis, S. Johnson, and J. J. Torres. "Enhancing Neural Network Performance via Assortativity." *Physical Review E* 83:3(2011), 036114.

[9] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski. "Edge Direction and the Structure of Networks." In *Proceedings of the National Academy of Sciences* 107:24(2010), 10815–10820.

[10] A. Kaltenbrunner, G. Gonzalez, R. R. De Querol, and Y. Volkovich. "Comparative Analysis of Articulated and Behavioural Social Networks in a Social News Sharing Website." *New Review of Hypermedia and Multimedia* 17:3(2011), 243–266.

[11] M. G. Kendall. "A New Measure of Rank Correlation." *Biometrika* 30:1–2(1938), 81–93.

[12] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. "When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages." In *International Conference on Weblogs and Social Media*. AAAI, 2011.

[13] N. Litvak and R. van der Hofstad. "Degree-Degree Correlations in Random Graphs with Heavy-Tailed Degrees." arXiv preprint arXiv:1202.3071, 2012. To appear in *Internet Mathematics*.

[14] N. Litvak and R. van der Hofstad. "Uncovering Disassortativity in Large Scale Free Networks." *Physical Review E* 87:2(2013), 022801.

[15] X. F. Liu and C. K. Tse. "Impact of Degree Mixing Pattern on Consensus Formation in Social Networks." *Physica A: Statistical Mechanics and its Applications* 407(2014), 1–6.

[16] M. E. J. Newman. "Assortative Mixing in Networks." *Physical Review Letters* 89:20(2002), 208701.

[17] M. E. J. Newman. "Mixing Patterns in Networks." *Physical Review E* 67:2(2003), 026126.

[18] M. E. J. Newman. "The Structure and Function of Complex Networks." *SIAM Review* 45:2(2003), 167–256.

[19] M. Piraveenan, M. Prokopenko, and A. Y. Zomaya. "Assortativeness and Information in Scale-Free Networks." *The European Physical Journal B* 67:3(2009), 291–300.

[20] M. Piraveenan, M. Prokopenko, and A. Zomaya. "Assortative Mixing in Directed Biological Networks." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9:1(2012), 66–78.

[21] S. I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.

[22] C. Spearman. "The Proof and Measurement of Association Between Two Things." *The American Journal of Psychology* 15:1(1904),72–101.

[23] A. Srivastava, B. Mitra, N. Ganguly, and F. Peruani. "Correlations in Complex Networks Under Attack." *Physical Review E* 86:3(2012), 036106.

[24] A. Srivastava, B. Mitra, F. Peruani, and N. Ganguly. "Attacks on Correlated Peer-To-Peer Networks: An Analytical Study." In *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2011*, pp. 1076–1081. IEEE, 2011.

[25] P. van der Hoorn and N. Litvak. "Convergence of Rank Based Degree-Degree Correlations in Random Directed Networks." arXiv preprint arXiv:1407.7662, 2014.

[26] A.Vázquez and Y. Moreno. "Resilience to Damage of Graphs with Degree Correlations." *Physical Review E* 67(2003), 015101.

[27] S. Vigna. "A Weighted Correlation Index for Rankings with Ties." arXiv preprint arXiv:1404.3325, 2014.