Degree-degree correlations in directed networks with heavy-tailed degrees

Pim van der Hoorn, Nelly Litvak
University of Twente

October 25, 2013

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, Pearson's correlation coefficients converge to a non-negative number in the infinite network size limit. We propose alternative measures for degree-degree correlations 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.

Keywords degree assortativity, degree-degree correlations, scale free directed networks, power laws, rank correlations.

1 Introduction

In the analysis of the topology of complex networks a feature that is often studied is the degree-degree correlation, also called degree assortativity of the network. A network has positive degree-degree correlation, 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 have negative degree-degree correlation, it is disassortative. A measure for degree assortativity was first given for undirected networks by Newman [15], 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 [16] and later adopted for analysis of directed complex networks in [18] and [8]. Analysis of the degree-degree correlation has been applied to networks in a variety of scientific fields such as neuroscience, molecular biology, information theory and social network sciences. In [10, 12] degree-degree correlations are used to investigate the structure of collaboration networks of a social news sharing website and Wikipedia discussion pages, respectively. Another

example is [9], where the influence of the phenotopic viability of a family of plants on the degree-degree correlations of their genetic network is investigated. Degree assortativity has also been found to influence several properties of networks. For instance, neural networks with high assortativity seem to behave more efficiently under the influence of noise [7]. Information content has been shown to depend on the absolute value of the degree assortativity [19] and networks with high degree assortativity have been shown to be less stable [4].

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 non-negative 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, 17, Table II], this could then explain why large networks are rarely classified as disassortative. In the same paper a new measure, corresponding to Spearman's rho [20], has been proposed as an alternative.

In this paper we will extend the analysis in [13] to the setting of directed networks. Here we have to consider four types of degree-degree correlations, 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 correlations. Although these results give some insights into the workings of these correlations we still do not fully understand the differences between the four correlation types or what they mean for the structural properties of the network.

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 [18] and [8] converge to a non-negative 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 correlations in such directed networks. To strengthen this argument we also give examples which clearly show that the values given by Pearson's correlation coefficients do not represent the correlation between the degrees, which it is suppose to measure. As an alternative we propose correlation measures based on Spearman's rho [20] and Kendall's tau [11]. These measures are based on the ranking of the degrees rather then their value and hence do not exhibit the size bias observed in Pearson's correlation coefficients. We will give several examples where the difference between these three measures is shown. 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 paper.

This paper 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 measures Spearman's rho and Kendall's tau for degree-degree correlations in Section 4. Example graphs that illustrate the differ-

ence between the three measures are presented in Section 5. Finally the degree-degree correlations for the Wikipedia graphs are presented in Section 6.

2 Definitions and notations

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

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 will 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 correlations

In this paper we are interested in measuring the correlation between the degrees at both sides of an edge. That is, we measure the correlation between two vectors X and Y as 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-degree correlation. In the directed setting however, we can consider any combination of the two degree types resulting in four types of degree-degree correlations, illustrated in Figure 1.

From Figure 1 one can already observe some interesting features of these correlations. For instance, in the Out/In correlation the edge that we consider contributes to the degrees on both sides. We will later see that the Out/In correlation actually generalizes the degree-degree correlation in the undirected case. To be more precise, our result for this correlation type generalizes the result obtained in [14] when we transform from the undirected case by making every edge bi-directional.

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

3 Pearson's correlation coefficient

Among all correlation measures, the measure proposed by Newman [15, 16] 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

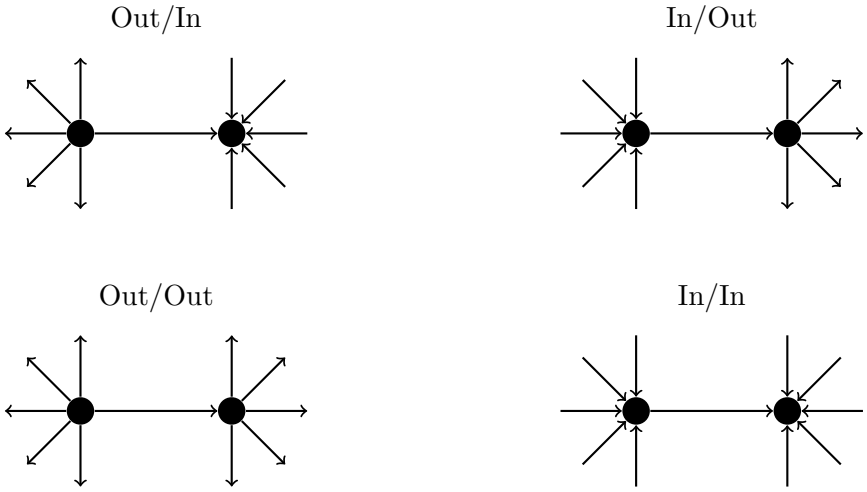

Figure 1: Four degree-degree correlation types

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 correlations. 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 correlations.

Let us consider Pearson's correlation coefficients as in [15, 16], adjusted to the setting of directed graphs as in [8, 18]. This will constitute four formula's which 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:

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

where

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

$$\sigma^\beta(G) = \sqrt{\frac{1}{|E|} \sum_{e \in E} D^\beta(e^*)^2 - \frac{1}{|E|^2} \left(\sum_{e \in E} D^\beta(e^*) \right)^2}. \quad (3)$$

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 correlation.

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 (1)-(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 on G is defined by*

$$r_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \frac{1}{|E|} \sum_{e\in E} D^\alpha(e_*)D^\beta(e^*) - \hat{r}_\alpha^\beta(G), \tag{4}$$

where

$$\hat{r}_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \frac{1}{|E|^2} \sum_{v\in V} D^+(v)D^\alpha(v) \sum_{v\in V} D^-(v)D^\beta(v), \tag{5}$$

$$\sigma_\alpha(G) = \sqrt{\frac{1}{|E|}\sum_{v\in V} D^+(v)D^\alpha(v)^2 - \frac{1}{|E|^2}\left(\sum_{v\in V} D^+(v)D^\alpha(v)\right)^2}, \tag{6}$$

$$\sigma^\beta(G) = \sqrt{\frac{1}{|E|}\sum_{v\in V} D^-(v)D^\beta(v)^2 - \frac{1}{|E|^2}\left(\sum_{v\in V} D^-(v)D^\beta(v)\right)^2}. \tag{7}$$

Just as in the undirected case, c.f. [13, 14], the wiring of the network only contributes to the positive part of (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 equations (2) and (3) it follows that $\sigma_\alpha(G)$ and $\sigma^\beta(G)$ are the variance 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 non-zero 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 graph and take $\alpha, \beta \in \{+, -\}$. Then the following holds:*

$$\frac{1}{|E|}\left(\sum_{v\in V} D^\alpha(v)D^\beta(v)\right)^2 \leq \sum_{v\in V} D^\alpha(v)D^\beta(v)^2 \tag{8}$$

and strict inequality holds if and only if there exits 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:

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

which proves (8). From the last line one easily sees that strict inequality holds if and only if there exits 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 under rather general conditions Pearson's correlation coefficients (4) converges to a non-negative 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 and prove the result. Then we will motivate the chosen conditions. 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 and $N \in \mathbb{N}$ 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,

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

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 non-negative.

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 non-negative. Therefore we need only to prove 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 are 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

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

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:

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

□

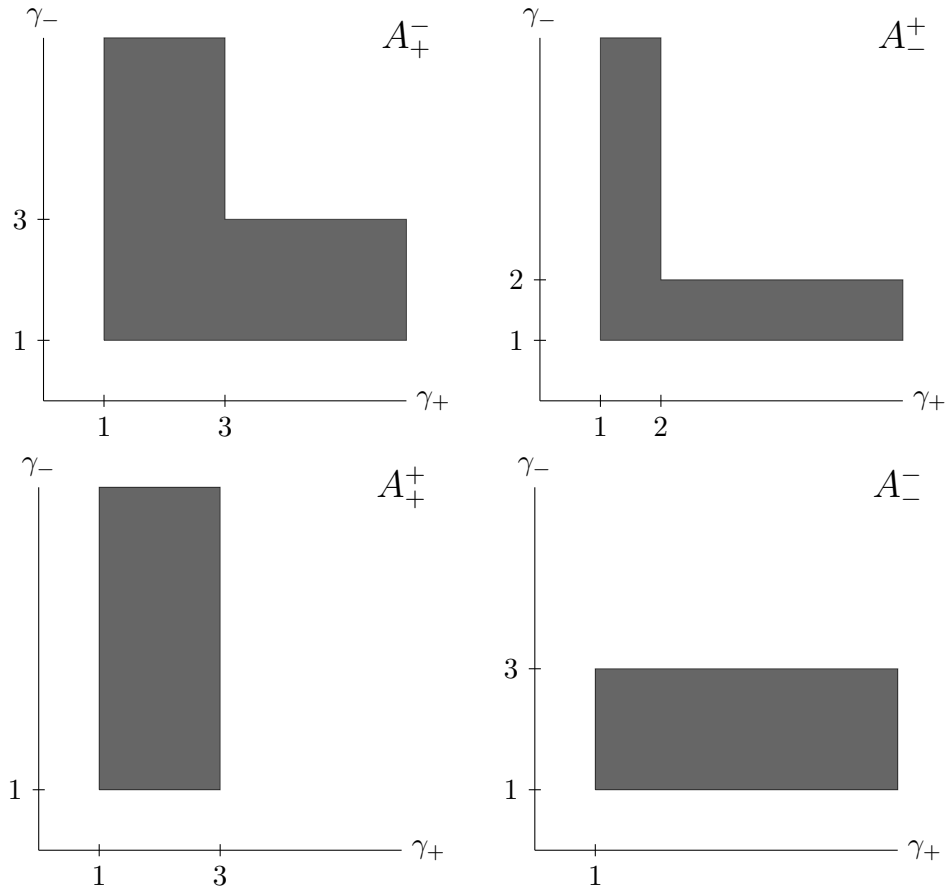

Figure 2

Let us now provide an intuitive explanation for the areas A_α^β, as depicted in Figure 2. The key observation is that due to 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^- only has terms of order one while 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^-.

For the Out/In correlation, i.e. $\alpha = +$ and $\beta = -$, we see from equations (5)-(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 a undirected network and make it directed by replacing each undirected edge with a bi-directional edge. Then $D^+(v) = D^-(v)$ for all $v \in V$ and hence $r_+^-(G)$ equals the expression of equation (3.4) in [13] when we replace D by either D^+ or D^-.

Theorem 3.5 has several consequences. First of all, 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 non-negative. This could explain why most large networks are said not to have disassortative degree-degree correlations. In Section 5 we will give examples where this behavior can be observed. 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 N, $r_\alpha^\beta(G_n)$ will always be 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 $\mathbb{P}(X > t) = L(t)t^{-\gamma}$ for some slowly varying function L, that is $\lim_{t\to\infty} L(tx)/L(t) = 1$ for all x. We write $\mathcal{R}_{-\gamma}$ for the space of all regularly varying random variables with exponent γ. For a regularly varying random variable $X \in \mathcal{R}_{-\gamma}$ 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 distribution closely resembles a power law distribution, c.f. Table II of [1] and [17]. 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 non-degenerate, we have $\mathbb{E}\left[(\mathcal{D}^\pm)^k\right] > \mathbb{E}[\mathcal{D}^\pm]^k$, and thus by taking $d(p,q) = \mathbb{E}[(\mathcal{D}^+)^p(\mathcal{D}^-)^q]$, we get G4 where the second part follows from Hölder's inequality. Although i.i.d. sequences for the in- and out-degrees 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 [5], where a directed version of the configuration model is introduced and it is proven 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. [6]. This means that

$$\frac{1}{a_n} \sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q \xrightarrow{d} S(\gamma_+/p \wedge \gamma_-/q), \quad \text{as } n \to \infty \tag{9}$$

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 (9) implies that G3 holds with high probability. We remark that although this motivation is based on results where the regularly varying random variables are assumed to be independent the dependent case can be included. For this one then needs to adjust the scaling parameters in G3 for the specified dependence.

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 correlations, Spearman's rho [20] and Kendall's tau [11], which are based on the rankings of the degrees rather then their actual value. We will define these correlation measures and argue that they do not have unwanted behavior as we observed for Pearson's correlation coefficients. We will later use examples to enforce this argument and show that Spearman's rho and Kendall's tau are better candidates for measuring degree-degree correlations.

4.1 Spearman's rho

Spearman's rho [20] 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*

$$\rho_\alpha^\beta(G) = \frac{12\sum_{e\in E} R^\alpha(e_*)R^\beta(e^*) - 3|E|(|E|+1)^2}{|E|^3 - |E|}, \tag{10}$$

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

From (10) we see that the negative part of $\rho_\alpha^\beta(G)$ depends only on the number of edges

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

while 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 non-negative number while $r_\alpha^\beta(G_n)$ 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 potentionally 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 $(x_i)_{1\leq i\leq n}$ 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 above 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 average 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 average α-β Spearman's rho of the graph G by*

$$\overline{\rho}_\alpha^\beta(G) = \frac{4\sum_{e\in E} \overline{R}^\alpha(e_*)\overline{R}^\beta(e^*) - |E|(|E|+1)^2}{\overline{\sigma}_\alpha(G)\overline{\sigma}^\beta(G)}, \tag{11}$$

where

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

and

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

4.2 Kendall's Tau

Another common rank correlation measure is Kendall's Tau [11], which measures the weighted difference between the number of concordant and disconcordant pairs of the

joint observations $(x_i, y_i)_{1\le i\le n}$. 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 disconcordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$.

Definition 4.4. *Let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{-, +\}$ and denote by $\mathcal{N}_c$ and $\mathcal{N}_d$, respectively, the number of concordant and disconcordant 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*

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

It might seem at first that τ does not suffer from ties. However, note that the numerator of τ includes only strictly (dis)concordant pairs, while the denominator is equal to the number of all possible pairs, irregardless 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 example in Section 5. Since, as discussed above, 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.

5 Bridge graph example

In this section we will provide a sequences of graphs to illustrate the difference between the four correlation measures in directed networks. We start with a deterministic sequence and will 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:

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

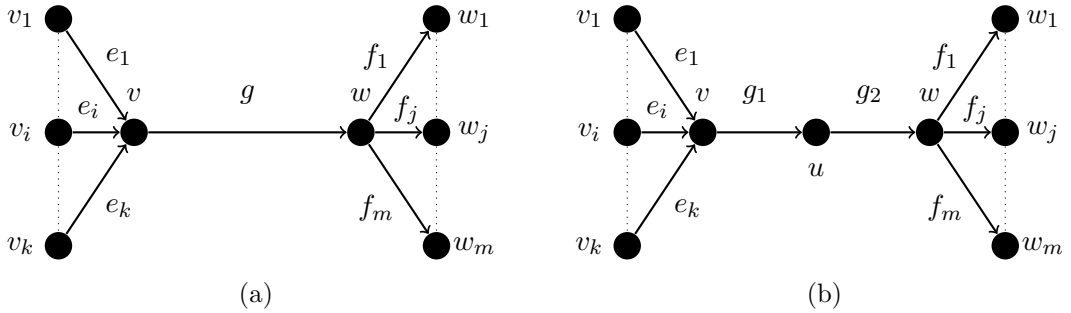

Figure 3: A graphical representation of the graphs $G(k,m)$ (a) and $\hat{G}(k,m)$ (b).

Looking at the scatter plot of $(D^-(e_*), D^+(e^*))_{e\in E(k,m)}$, Figure 4a, we see that the point (k,m) contributes towards a positive correlations while the points $(0,1)$ and $(1,0)$ contribute towards a negative correlation. Hence, depending on how much weight we put on each of these points we could argue equally well that this graph has positive or negative In/Out correlation. We can however extend the in-out bridge graph to a graph for which we do have a clear expectation for the In/Out correlation.

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 In/Out correlation. This intuition is supported by the scatter plot of $(D^+(e^*), D^-(e_*))_{e\in\hat{E}(k,m)}$, 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 above reasoning we would expect that any In/Out correlation measure of $\hat{G}_n^a$ would converge to -1.

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

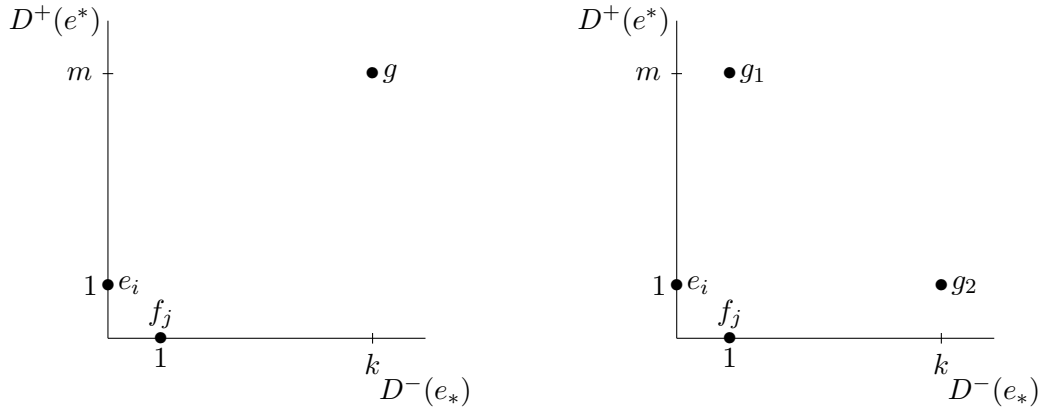

Figure 4: The scatter plots for the degrees of (a) $G(k,m)$ and (b) $\hat{G}(k,m)$.

5.1.1 Pearson In/Out correlation

We start with the graph G_n^a. Basic calculations yield that

$$\sum_{e\in E_n^a} D^-(e_*)D^+(e^*) = an^2, \tag{12}$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v) = (1+a)n, \tag{13}$$

$$\sum_{v\in V_n^a} D^-(v)^2D^+(v) = n^2+an, \tag{14}$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v)^2 = n+a^2n^2, \tag{15}$$

hence, using (6) and (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 (4) with $\alpha=-$ and $\beta=+$ we get

$$\begin{aligned}r_-^+(G_n^a) &= \frac{|E_n^a|an^2-(1+a)^2n^2}{|E_n^a|\sigma_\alpha(G_n^a)|E_n^a|\sigma^\beta(G_n^a)}\\ &= \frac{a(1+a)n^3-(a^2+a+1)n^2}{a\sqrt{(1+a)n^3-(n-1)an}\sqrt{(1+a)n^3-(an-1)n}}.\end{aligned} \tag{16}$$

From (16) 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 equations (13)-(15), while equation (12) 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 correlation measure on $\hat{G}_n^a$ to converge to -1.

5.1.2 Spearman In/Out correlation

We start by calculation $\overline{\rho}_-^+(G_n^a)$. For this observe that by (11) 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 to the one for $r_-^+(G_n^a)$. 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 equation (11) 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

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

which equals $\lim_{n\to\infty}\rho(G_n^a)$. 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 will show that the choice of ranking of the tied values can have a great effect on the outcome of the In/Out correlation. In this example we will pick two rankings, one will yield a positive correlation while the other will give a negative correlation.

It is clear from the definition of G_n^a that the in- and out-degrees of all e_i are the same and similar 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

$$\rho_-^+(G_n^a) = \frac{(a^3-3a^2-3a+1)n^3+3(a+1)^2n^2+2(a+1)n}{(a^3+3a^2+3a+1)n^3+3(a+1)^2n^2+2(a+1)n}. \tag{17}$$

Now 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 one 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,

$$\rho_-^+(G_n^a) = \frac{-(a+1)^3n^3+3(a+1)^2n^2+2(a+1)n}{(a+1)^3n^3+3(a+1)^2n^2+2(a+1)n} \tag{18}$$

When we compare (18) with (17) we see that for the former $\lim_{n\to\infty}\rho_-^+(G_n^a) = -1$ for all $a\in\mathbb{N}$ while 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 (17), which becomes positive when $a\ge 4$. This indicates what was already mentioned in Section 4.1.1, that changing the order of the ties can have a large impact on the value of $\rho_\alpha^\beta(G)$.

5.1.3 Kendall's Tau In/Out correlation

The last correlation measure we compute is Kendall's Tau. In order to do this we need to determine the number of concordant and disconcordant pairs. Starting with G_n^a, we observe that we have three kinds of joint observations, namely

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

The combinations I and III, and II and III are concordant while I and II are disconcordant. From this it follows that $\mathcal{N}_c = (a+1)n$ while $\mathcal{N}_d = an^2$. Hence we get, see Definition 4.4.

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

which gives $\lim_{n\to\infty} \tau_-^+(G_n^a) = -\frac{2a}{(a+1)^2}$.

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 disconcordant, while 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 $\lim_{n\to\infty} \tau_-^+(G_n^a) = -\frac{2a}{(a+1)^2}$ which equals the limit for $\tau_-^+(G_n^a)$.

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 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 $(X_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$ 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 the disconnected collection of the graphs $(G(W_i, Z_i))_{1\leq i\leq n}$. 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 a specific choice of a.

Using the calculations in Section 5.1.1 we obtain:

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

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

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

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

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

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, due to Lemma 2.2 in [13] we have

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

Combining this we get

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

and hence

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

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

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

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 above $\frac{1+aT}{\sqrt{1+T}\sqrt{1+a^2T}}$ has support contained in $(\varepsilon,1)$.

This example shows that Pearson's correlation coefficients r_α^β can converge to a non-negative 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 [2, 3] and the fastutil package from The Laboratory for Web Algorithmics (LAW) at the Universit degli studi di Milano, `http://law.di.unimi.it`. The calculations where done on the Wikepedia graphs, `http://wikipedia.org`, 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 paper.

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 decribed in Section 3 of [5]. 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 selfloop we remove it. When we end up with multiple edges between two vertices we combine them into one edge. Proposition 3.7 of [5] 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 correlations where calculated using all four measures and then for each correlation type and measure we took the average. The results are presented in Table 1.

The first observation is that for each Wikipedia graph and correlation type, the measures ρ, $\overline{\rho}$ and τ have the same sign while r in many cases has a different sign. Furthermore, there are many cases where the absolute value of the three rank correlations is at least an order of magnitute 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 one could freely change between these two when calculating degree-degree correlations. Because for ρ both the average and the variance are known upfront, it is computationally easier than $\overline{\rho}$ while the latter is easier to analyze in a non-random setting.

Finally, we notice that in the synthetic configuration model, all correlation measures are close to zero, and the difference between different realizations of the model is remakarbly 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 behaviour 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 should be seen as significant and how to construct mathmatically sound statistical tests for establishing such significant dependencies.

		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

Table 1: Degree-degree correlations for Wikipedia graphs.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Paolo Boldi and Sebastiano Vigna. The webgraph framework i: compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, pages 595–602. ACM, 2004.

[3] Paolo Boldi and Sebastiano Vigna. The webgraph framework ii: Codes for the world-wide web. In *Data Compression Conference, 2004. Proceedings. DCC 2004*, page 528. IEEE, 2004.

[4] Markus Brede and Sitabhra Sinha. Assortative mixing by degree makes a network more unstable. *arXiv preprint cond-mat/0507710*, 2005.

[5] Ningyuan Chen and Mariana Olvera-Cravioto. Directed random graphs with given degree distributions. *arXiv preprint arXiv:1207.2475*, 2012.

[6] Daren B.H. Cline. Convolution tails, product tails and domains of attraction. *Probability Theory and Related Fields*, 72(4):529–557, 1986.

[7] Sebastiano de Franciscis, Samuel Johnson, and Joaquín J. Torres. Enhancing neural-network performance via assortativity. *Physical Review E*, 83(3):036114, 2011.

[8] Jacob G. Foster, David V. Foster, Peter Grassberger, and Maya Paczuski. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences*, 107(24):10815–10820, 2010.

[9] Adrien Henry, Françoise Monéger, Areejit Samal, and Olivier C. Martin. Network function shapes network structure: the case of the arabidopsis flower organ specification genetic network. *Mol. BioSyst.*, 2013.

[10] Andreas Kaltenbrunner, Gustavo Gonzalez, Ricard Ruiz De Querol, and Yana Volkovich. Comparative analysis of articulated and behavioural social networks in a social news sharing website. *New Review of Hypermedia and Multimedia*, 17(3):243–266, 2011.

[11] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[12] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *ICWSM*, 2011.

[13] Nelly Litvak and Remco 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] Nelly Litvak and Remco van der Hofstad. Uncovering disassortativity in large scale-free networks. *Physical Review E*, 87(2):022801, 2013.

[15] Mark E.J. Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[16] Mark E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[17] Mark E.J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[18] Mahendra Piraveenan, Mikhail Prokopenko, and Albert Zomaya. Assortative mixing in directed biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(1):66–78, 2012.

[19] Mahendra Piraveenan, Mikhail Prokopenko, and Albert Y. Zomaya. Assortativeness and information in scale-free networks. *The European Physical Journal B*, 67(3):291–300, 2009.

[20] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.