Phase transitions for scaling of structural correlations in directed networks.

Pim van der Hoorn,[*] Nelly Litvak[†]

April 8, 2015

Abstract

Analysis of degree-degree dependencies in complex networks, and their impact on processes on networks requires null models, i.e. models that generate uncorrelated scale-free networks. Most models to date however show structural negative dependencies, caused by finite size effects. We analyze the behavior of these structural negative degree-degree dependencies, using rank based correlation measures, in the directed Erased Configuration Model. We obtain expressions for the scaling as a function of the exponents of the distributions. Moreover, we show that this scaling undergoes a phase transition, where one region exhibits scaling related to the natural cut-off of the network while another region has scaling similar to the structural cut-off for uncorrelated networks. By establishing the speed of convergence of these structural dependencies we are able to asses statistical significance of degree-degree dependencies on finite complex networks when compared to networks generated by the directed Erased Configuration Model.

1 Introduction

The tendency of nodes in a network to be connected to nodes of similar large or small degree, called network assortativity, degree mixing or degree-degree dependency, is an important characterization of the topology of the network, influencing many processes on the network. It has received significant attention in the literature, for instance in the field of network stability [29], attacks on P2P networks [25] and epidemics [2, 3].

An important method to analyze these degree-degree dependencies or their influence on other network properties or processes on the network, is to compare results to an average over several instances of similar networks with neutral mixing. These null models often come in two flavors. The first approach is to sample from graphs with the same degree sequence but neutral mixing. A widely accepted methodology for such sampling is through the local rewiring model, [17], which takes the original network and randomly swaps edges until a randomized version is attained. The disadvantage of these methods is that they have no theoretical performance guarantees. The second approach is to generate a random graph with neutral mixing, which preserves basic features, such as the degree distribution. A well known model of this type is the Configuration Model [6, 19, 21]. Here the degrees of vertices are drawn independently from the given distribution, under the restriction that the total sum of degrees is even. Then the stubs are paired uniformly at random to form edges. If we want to obtain a simple graph in this way, we can either rewire till a simple graph is generated (Repeated Configuration Model), or we remove the excess edges and self loops (Erased Configuration Model).

We note that there are many other methods, that generate simple random graphs and have theoretically established performance guarantees, for example, the sequential algorithm in [1] that creates a random graph with given degrees, or a grand-canonical model in [24] that generates a

---

[*]University of Twente, w.l.f.vanderhoorn@utwente.nl

[†]University of Twente, n.litvak@utwente.nl

Wikipedia	N	$N^{1/2}$	γ_+	γ_-	$\max D^+$	$\max D^-$
DE	1,532,978	1,238	1.80	1.05	5,032	118,064
EN	4,212,493	2,052	2.14	1.20	8,104	432,629
IT	1,017,953	1,009	1.96	1.05	5,212	91,588
NL	1,144,615	1,070	1.82	1.10	10,175	102,450
PL	949,153	974	1.90	1.04	4,100	112,537

Table 1: Basic degree characteristics of Wikipedia networks. The exponents of the degree distributions are estimated using the implementation of the techniques from [9] by Peter Bloem, http://github.com/Data2Semantics/powerlaws.

graph with given average degrees using a maximum-entropy method. However, to the best of our knowledge, none of these methods has an efficient implementation, that is feasible for a truly large network, such as Wikipedia or Twitter.

Although for both local rewiring and the Configuration Model neutral mixing is expected, since there is no preference in connecting two vertices, negative correlations are observed, [7, 18, 22], for scale-free networks with infinite variance of degrees, i.e. where the degree distribution satisfies

$$P(k) \sim k^{-(\gamma+1)}, \quad 1 < \gamma \le 2. \tag{1}$$

In [18] this phenomenon is explained by observing that if one allows at most one edge between two vertices, nodes with large degree must connect to nodes of small degree because there are simply not enough distinct large nodes to connect to. A similar explanation is given in [7]. Here, however, this is then related to the difference in scaling between the *natural* and *structural cut-off* of the network. The former is defined [10] as the degree value k_c, of which, on average, only one instance is observed:

$$N \int_{k_c}^{\infty} P(k)dk \sim 1. \tag{2}$$

The structural cut-off is defined as the value k_s for which the ratio between the average number of edges that connect any two vertices of degree k_s, and the maximum possible number of such edges in a simple graph, is 1. For networks with degree distribution (1) it follows from (2) that the natural cut-off scales as $N^{1/\gamma}$, while the structural cut-off for uncorrelated networks scales as, see [4], $N^{1/2}$. Therefore, when $\gamma < 2$, the natural cut-off scales at a slower rate which in turn gives rise to structural negative correlations.

To remedy these finite size effects the authors of [7] propose an Uncorrelated Configuration Model. This model follows the same procedure as the regular Configuration Model, with the addition that the sampled degrees are bounded, $m \le k_i \le N^{1/2}$. Experiments in [7] indeed show that these networks are uncorrelated. However, many scale-free networks, for instance Twitter, have nodes who's degree is of larger order than $N^{1/2}$, which is a characteristic property of scale-free graphs. For example, Table 1 displays the characteristics of Wikipedia networks for different languages. Here we see that the maximum out-degree could be considered to be of order $N^{1/2}$, while the maximum in-degree is definitely of a much larger scale. Therefore, randomized versions of these networks, generated by the Uncorrelated Configuration Model, do not have the same basic degree characteristics as the original network, since the maximum degree is restricted. Hence, they are less suitable for comparison of the degree-degree dependencies.

In this paper we consider the directed Erased Configuration Model, [8], where after the pairing self loops are removed and multiple edges are merged. This model was shown to have neutral mixing in the infinite network size limit, see [27] Section 5. Therefore, from a purely mathematical point of view, it is a null model for degree-degree dependencies in the limit. Moreover, asymptotically, the degree distributions are preserved and hence, all basic degree characteristics. Still, for finite sizes, structural dependencies are present.

Rather than trying to control these correlations, our goal is to evaluate their magnitude and investigate their size dependence. We obtain the scaling for the structural correlations in the
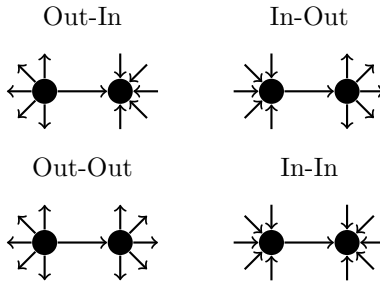
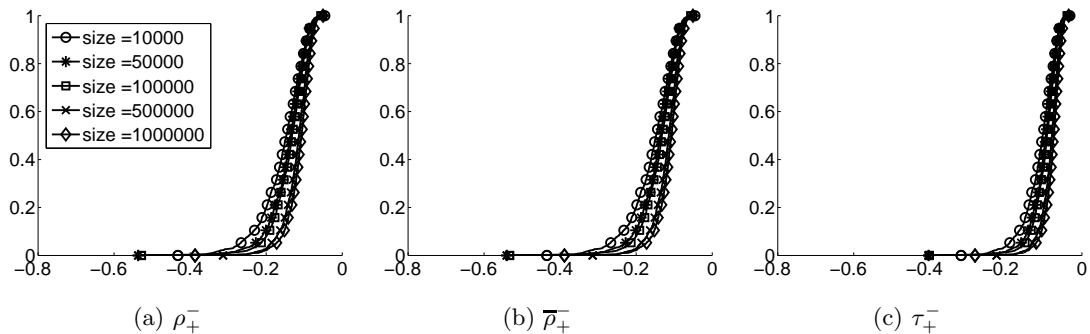Figure 1: The four different degree-degree dependency types in directed networks.

(a) ρ_+^- (b) $\overline{\rho}_+^-$ (c) τ_+^-

Figure 2: Plots of the empirical cumulative distribution of ρ_+^-, $\overline{\rho}_+^-$ and τ_+^- for ECM graphs of different sizes with $\gamma_\pm = 1.2$. Each plot is based on 10^3 realizations of the model.

ECM, in terms of the power law exponents of the in- and out-degrees. In particular, we show that this scaling undergoes an interesting phase transition, and can be dominated by terms related to either the structural or the natural cut-off of the network. To the best of our knowledge, this is the first study that provides a systematic mathematical characterization for the magnitude of negative correlations in a simple graph with neutral mixing.

By determining the scaling of the structural correlations we can asses the significance of measured correlations as well as their influence on network processes, on real world networks of finite size, by comparing them to the directed Erased Configuration Model. This approach has the advantage of preserving the degree characteristics of the original network, it can be easily implemented and applied to all networks with scale free-degree distributions and finite expectation.

2   Degree-degree dependencies in random directed networks

We analyze degree-degree dependencies in random directed networks of size N, where the distribution of the out- and in-degree (D^+, D^-) follow, respectively,

$$P^+(k) \sim k^{-(1+\gamma_+)} \text{ and } P^-(\ell) \sim \ell^{-(1+\gamma_-)}, \ \gamma_\pm > 1. \tag{3}$$

In directed networks one can consider four types of degree-degree dependencies, depending on the choice of the degree type on both sides of an edge, see Figure 1. For the remainder of this paper we denote by E the number of edges and adopt the notation style from [12, 28] to index the degree types by $\alpha, \beta \in \{+, -\}$.

A common measure for degree-degree dependencies, introduced in [20], computes Pearson's correlation coefficients on the joint data $(D_i^\alpha, D_j^\beta)_{i \to j}$, where the indices run over all i, j for which there is an edge $i \to j$.

However, Pearson's correlation coefficients are unable to measure strong negative degree-degree dependencies in large networks where the variance of the degrees is infinite, as was shown for

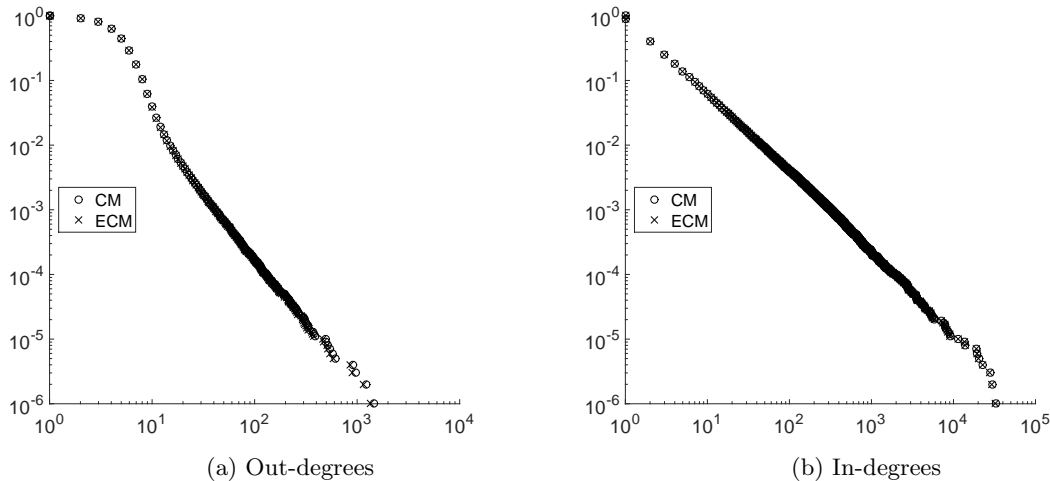(a) Out-degrees   (b) In-degrees

Figure 3: Plots of the out- and in-degree distribution, on log-log scale, for a graph generated by the ECM, of size 10^6 with $\gamma_+ = 1.9$ and $\gamma_- = 1.2$, before (CM) and after (ECM) the removing of edges.

undirected networks in [16, 11] and for directed networks in [28]. Since our interest is mainly in networks in the infinite variance domain, i.e. $1 < \gamma_\pm \leq 2$, we need different measures. In [28] it was suggested to use rank correlations, related to Spearman's rho [23] and Kendall's tau [14], to measure degree-degree dependencies.

Spearman's rho computes Pearson's correlation coefficient on the ranks of $(D_i^\alpha, D_j^\beta)_{i \to j}$ rather then their actual values. Since this data will contain many ties, one needs to use ranking schemes that deal with these ties. In [28] two such schemes are considered, resolving ties at random and assigning an average rank to tied values, which give two correlation measures denoted by ρ_α^β and $\overline{\rho}_\alpha^\beta$, respectively. Here, the subscript index denotes the degree type of the source, while the superscript index denotes the degree type of the target of a directed edge. For instance, ρ_+^- denotes Spearman's rho for the Out-In dependency. The second rank correlation measure, Kendall's tau τ_α^β, calculates the normalized number of swaps needed to match the ranks of the joint data.

Exact formulas for these three measures, in terms of the degrees, are given in [28]. In [27] formulas are given in terms of the empirical distributions of D^α and D^β and their joint distribution, evaluated at (D_i^α, D_j^β) for an edge $i \to j$ selected uniformly at random. From these it follows that if the network has neutral mixing, then ρ_α^β and τ_α^β are similar, while ρ_α^β and $\overline{\rho}_\alpha^\beta$ differ by a term of $O(1)$, which does not influence the scaling. To illustrate this we plotted the empirical cdf's of ρ_+^-, $\overline{\rho}_+^-$ and τ_+^- for a collection of ECM graphs in Figure 2; where we clearly observe the similar behavior of the three measures. Therefore, for the analysis of degree-degree dependencies, we will only consider ρ_α^β, which corresponds to Spearman's rho where ties are resolved uniformly at random.
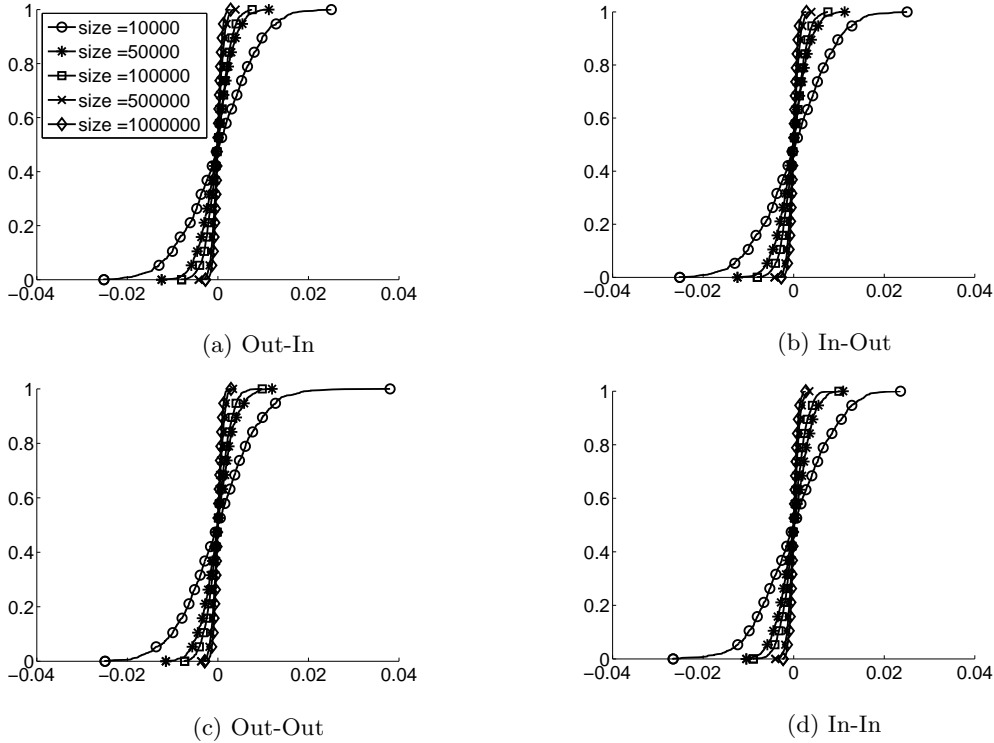
Figure 4: Plots of the empirical cumulative distribution of ρ_α^β for all four degree-degree dependency types for ECM graphs of different sizes with $\gamma_\pm = 2.1$. Each plot is based on 10^3 realizations of the model.

3   The directed Erased Configuration Model

The directed Configuration Model (CM) starts with degree sequences $(D_i^+, D_i^-)_{1 \le i \le N}$ that satisfy,

$$E = \sum_{i=1}^N D_i^\pm \quad \sim \quad \mu N \tag{4a}$$

$$\sum_{i=1}^N D_i^+ D_i^- \quad \sim \quad \mu^2 N \tag{4b}$$

$$\sum_{i=1}^N (D_i^\pm)^p \quad \sim \quad N^{p/\gamma_\pm}, \quad p > \gamma_\pm, \tag{4c}$$

for some $\mu > 0$. The stubs are then paired at random to form edges. This will in general constitute a graph with self-loops and multiple edges between nodes. If the degree variance is finite, then the probability of generating a simple graph is bounded away from zero and thus, by repeating the pairing step until such a graph is generated, we get a network randomly sampled from all networks of given size and degree sequences. This is called the Repeated Configuration Model (RCM).

When the variance of the degrees is infinite, the probability of generating a simple graph converges to zero as the graph size increases, and therefore we need to enforce that the resulting graph is simple. For this we use the Erased Configuration Model (ECM), where, during the pairing, a new edge is removed if it already exists or if it is a self loop. Although this seems to be a strong alteration of the initial degree sequence, asymptotically, the degrees of the resulting
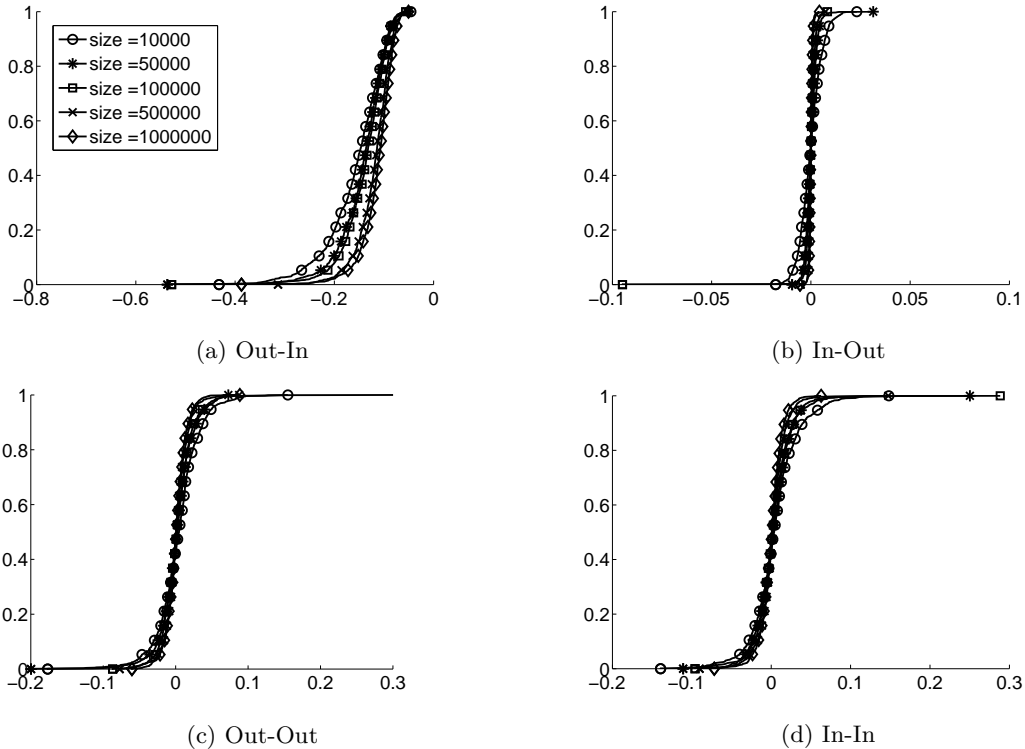
Figure 5: Plots of the empirical cumulative distribution of ρ_α^β for all four degree-degree dependency types for ECM graphs of different sizes with $\gamma_\pm = 1.2$. Each plot is based on 10^3 realizations of the model.

network still follow the same distribution, see [8]. For illustration, in Figure 3, we plotted the degree distributions of and ECM graph of size 10^6 before and after the removing of edges. Clearly there is hardly any difference between the two distributions. In particular the degree sequences of ECM graphs still satisfy (4). Unlike many other methods, random pairing of the stubs can be implemented very efficiently for even billions of nodes. Moreover, the ECM is computationally less expensive than RCM, since we do not need to repeat the pairing. Therefore we suggest to use the ECM as a standard null-model. In the rest of the paper we will characterize the structural dependencies in the ECM.

4 Degree-degree dependencies in the ECM

It is clear that when we use the CM, i.e. allow for multiple edges and self loops, then our graphs will have neutral mixing since all stubs are connected completely at random. For the ECM however, we remove edges to make the graph simple, which has been shown [18, 7] to give rise to negative correlations. Nevertheless, the ECM has asymptotically neutral mixing, which can be shown as follows.

Let E_{ij} be the matrix counting the number of edges between i and j after the pairing and let E_{ij}^c denote the matrix counting the number of removed edges between i and j by the ECM. Then for the CM it holds that $D_i^+ = \sum_{j=1}^N E_{ij}$ while for the ECM we have $D_i^{+\prime} = \sum_{j=1}^N (E_{ij} - E_{ij}^c)$. Therefore, the difference between the empirical distributions of D_i^α and D_j^β, for an edge $i \to j$ sampled at random, in the CM and ECM, will be of the order $\sum_{i,j=1}^N E_{ij}^c / E$, whose average, with respect to the degree sequences, converges to zero [27],

6

N	$\langle \rho_+^- \rangle$	$\langle \rho_-^+ \rangle$	$\langle \rho_+^+ \rangle$	$\langle \rho_-^- \rangle$
10000	-0.1568	-0.0001	0.0039	0.0048
50000	-0.1439	0.0001	0.0014	0.0029
100000	-0.1388	-0.0001	0.0026	0.0028
500000	-0.1198	0.0001	0.0011	0.0017
1000000	-0.1131	0.0000	0.0009	0.0002

Table 2: The average values for ρ for all four degree-degree dependencies types, for ECM graphs of different sizes, with $\gamma_\pm = 1.2$, based on 10^3 realizations of the model.

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i,j=1}^{N} \langle E_{ij}^c \rangle = 0. \tag{5}$$

This implies that the values of ρ_α^β for an ECM graph will converge to that of a CM graph, hence, asymptotically, $\rho_\alpha^\beta = 0$ and also $\overline{\rho}_\alpha^\beta = 0 = \tau_\alpha^\beta$, for the ECM.

However, for finite realizations in the infinite variance regime, negative correlations are still observed. To illustrate this we plotted the empirical cumulative distribution functions of ρ_α^β for graphs generated by the ECM with, both, finite and infinite degree variance, see Figure 4 and Figure 5, respectively. In addition, Table 2 contains the average values for all four correlation types in the infinite variance regime. One immediately observes that the Out-In dependency in ECM graphs with infinite variance, Figure 5a, displays strong structural negative correlations which decrease as the network grows, while for other three dependency types the values are concentrated around zero. Moreover, we see, Figure 4, that all four dependency types behave similar when the variance of the degrees is finite.

These negative Out-In correlations (ρ_+^-) can be explained by first observing that multiple edges are more likely to start in a node of large out-degree and end in a node of large in-degree, since these are more likely to be sampled. Now, consider the algorithm as first connecting all stubs at random and then removing self loops and merging multiple edges. By construction, immediately after the pairing the network will have neutral mixing. When merging multiple edges we will often delete connections from nodes of large out-degree to nodes of large in-degree. Such edges have contributed positively into ρ_+^-, thus, deleting them will shift ρ_+^- from zero in the CM to a negative value in the ECM. The other three dependency types are not effected since the out- and in-degree of a node in the ECM are independent.

Motivated by the analysis in this section, we will further focus on the behavior of ρ_+^- in the infinite-variance case, $1 < \gamma_+, \gamma_- \leq 2$, as the only scenario where we observe prominent structural correlations. We will discuss other scenarios in Section 6.

5 Scaling of the Out-In degree-degree dependency in the ECM

We will determine the scaling of ρ_+^- as a function of the exponents $\gamma_\pm$. That is, we will find coefficients $f(\gamma_+, \gamma_-)$ such that

$$\frac{\rho_+^- - \langle \rho_+^- \rangle}{N^{f(\gamma_+, \gamma_-)}}$$

converges to some limiting distribution. Here the expectation $\langle \rho_+^- \rangle$ is taken over all possible graphs of size N, generated by the ECM, with degree sequences satisfying (4). We note that although $\langle \rho_+^- \rangle$ is of similar order as the typical spreading of ρ_+^-, the latter, which we are going to evaluate, will define the magnitude of the structural negative correlations.

We obtain the scaling exponents $f(\gamma_+, \gamma_-)$ by establishing upper bounds on the scaling, and then show empirically that these bounds are tight. The scaling is an important quantity, characterizing the spread around the sample mean of ρ_+^- as a function of N. Roughly, this tells us how

much the measured values on a ECM graph of size N can deviate from the average and therefore enable us to asses the significance of the measured correlations of the corresponding real world networks.

5.1 Scaling of the erased number of edges

As we discussed in the previous section, the structural negative correlations appear after multiple edges and self-loops are erased. Hence, part of the scaling of ρ_+^- comes from the scaling of the average total number of erased edges. The latter scaling has a phase transition, which we will show by establishing two different upper bounds.

For the first upper bound, observe that

$$\sum_{i,j=1}^{N} E_{ij}^c = \sum_{i=1}^{N} S_{ii} + \sum_{i,j=1}^{N} M_{ij}, \tag{6}$$

where S is the diagonal matrix counting the number of self loops and M is the zero diagonal matrix that counts the excess edges, so $M_{ij} = k > 0$ means that $E_{ij} = k + 1$. For the self loops it holds that

$$\langle S_{ii} \rangle = \frac{D_i^+ D_i^-}{E}. \tag{7}$$

If we now take the total number of pairs of edges between i and j as an upper bound for the M_{ij}, then

$$\langle M_{ij} \rangle \leq \frac{(D_i^+)^2 (D_j^-)^2}{E^2}. \tag{8}$$

Applying this to (6) we get

$$\sum_{i,j=1}^{N} \frac{\langle E_{ij}^c \rangle}{E} \leq \frac{\sum_{i,j=1}^{N} (D_i^+)^2 (D_j^-)^2}{E^3} + \frac{\sum_{i=1}^{N} D_i^+ D_i^-}{E^2}. \tag{9}$$

We remark that if the second moment of both the out- and in-degree exists, then this upper bound scales as N^{-1}. When this is not the case, we get the scaling from (4) as

$$\frac{1}{E} \sum_{i,j=1}^{N} \langle E_{ij}^c \rangle = O\left(N^{(2/\gamma_+)+(2/\gamma_-)-3}\right). \tag{10}$$

The upper bound (10) is rather crude in the sense that for certain $1 < \gamma_\pm \leq 2$, we have $(2/\gamma_+) + (2/\gamma_-) > 3$ so that the right-hand side of (10) becomes infinite as $N \to \infty$.

To get a more precise upper bound let $p(n, m, L)$ denote the probability that none of the outbound stubs from a set of size n connect to an inbound stub from a set of size m, given that the total number of available stubs is L. We will establish a recursive relation for $p(D_i^+, D_j^-, E)$ by adopting the analysis from [26], Section 4. Similarly we get, by conditioning on whether we pick an inbound stub of i or not,

$$p(D_i^+, D_j^-, E) \leq \left(1 - \frac{D_j^-}{E}\right) p(D_i^+ - 1, D_j^-, E - 1),$$

where the upper bound comes from neglecting the event $D_i^+ + D_j^- > E$, in which case $p(D_i^+, D_j^-, E) = 0$. Continuing the recursion yields

$$p(D_i^+, D_j^-, E) \leq \prod_{k=0}^{D_i^+ - 1} \left(1 - \frac{D_j^-}{E - i}\right),$$

and a first order Taylor expansion then gives

$$p(D_i^+, D_j^-, E) \le e^{-D_i^+ D_j^- / E}. \tag{11}$$

Now, recall that E_{ij} denotes the total number of edges between i and j in the CM, before the removal step. Therefore,

$$\left\langle E_{ij}^c \right\rangle = \left\langle E_{ij} \right\rangle - (1 - p(D_i^+, D_j^-, E)).$$

Since $E = \sum_{i,j=1}^N \left\langle E_{ij} \right\rangle$ it follows that

$$\frac{1}{E} \sum_{i,j=1}^N \left\langle E_{ij}^c \right\rangle = 1 - \frac{N^2}{E} + \frac{1}{E} \sum_{i,j=1}^N p(D_i^+, D_j^-, E) \tag{12}$$

Hence, by plugging (11) into (12) we arrive at the following upper bound for the total average number of erased edges,

$$\frac{1}{E} \sum_{i,j=1}^N \left\langle E_{ij}^c \right\rangle \le 1 - \frac{N^2}{E} + \frac{1}{E} \sum_{i,j=1}^N e^{-D_i^+ D_j^- / E}. \tag{13}$$

The right hand side of (13) can be slightly rewritten to obtain the more informative expression

$$\frac{N^2}{E} \left(\frac{1}{E} \sum_{i,j=1}^N \frac{D_i^+ D_j^-}{N^2} - 1 + \sum_{i,j=1}^N \frac{e^{-(D_i^+ D_j^-)/E}}{N^2} \right). \tag{14}$$

Next, we note that (14) can be seen as an empirical form of

$$\frac{N}{\mu} \left(\frac{1}{N\mu} \left\langle \xi \right\rangle - 1 + \left\langle e^{-\xi/(N\mu)} \right\rangle \right), \tag{15}$$

where, using $a \wedge b$ to denote the minimum, ξ has distribution

$$P_\xi(k) \sim k^{-(\gamma_+ \wedge \gamma_-)-1}.$$

From a classical Tauberian Theorem for regularly varying random variables, see for instance [15], it follows that (15) scales as $N^{1-(\gamma_+ \wedge \gamma_-)}$. However, by the Central limit Theorem for stable random variables, see [30], we have

$$\left| \frac{1}{E} \sum_{i,j=1}^N \frac{D_i^+ D_j^-}{N^2} - \frac{\mu}{N} \right| = O\left(n^{-2+1/(\gamma_+ \wedge \gamma_-)} \right).$$

Therefore, since $E \sim \mu N$, the difference between (14) and (15) scales as $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$. This last term dominates (15) when $1 < \gamma_\pm \le 2$, hence it follows that

$$\frac{1}{E} \sum_{i,j=1}^N \left\langle E_{ij}^c \right\rangle = O(N^{-1+1/(\gamma_+ \wedge \gamma_-)}) \tag{16}$$

The scaling in (16) is related to that of the structural cut off described in [4], adjusted to the setting of directed networks with degree distributions (3). Moreover, comparing (16) to (10) we observe a phase transition, with respect to the tail exponents $\gamma_\pm$ of the degree distributions, in the scaling of the average total number of removed edges in the ECM, which will induce a phase transition in the scaling of the Out-In degree-degree dependency.
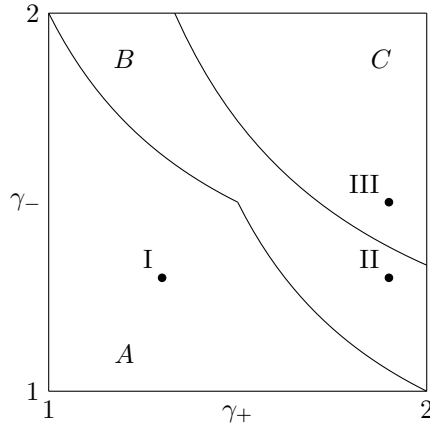
9

Figure 6: Plot of the different scaling regimes for ρ_+^-. The scaling terms for each of the three regions can be found in Table 3. The Roman numerals indicate the three different choices of γ_+ and γ_-, used in Figure 7 and 8, to illustrated the different regimes.

Region	$f(\gamma_+, \gamma_-)$
A	$1/(\gamma_+ \wedge \gamma_-) - 1$
B	$(2/\gamma_+) + (2/\gamma_-) - 3$
C	$-1/2$

Table 3: The three scaling terms for ρ_+^- for each of the three regions, displayed in Figure 6

5.2   Phase transitions for the Out-In degree-degree dependency

First we remark that for the CM, the empirical distribution of the degrees on both sides of a randomly sampled edge converges to the distribution of two independent random variables as N^{-1}, see [27]. Because Spearman's rho and Kendall's tau on independent joint measurements are normal statistics [13], the scaling of their average is $N^{-1/2}$. Hence ρ_α^β for CM graphs scales as $N^{-1/2}$. Since an ECM graph is basically a CM graph where multiple edges are merged and self-loops are removed, it follows that the distributions for the degrees on both side of a randomly chosen edge differ from those of the CM by terms of the order $\sum_{i,j=1}^{N} E_{ij}^c/E$. Therefore, the scaling of ρ_+^- is determined by the largest term out of $N^{-1/2}$ and the scaling of $\sum_{i,j=1}^{N} E_{ij}^c/E$. Since the latter undergoes a phase transition, we actually have a three stage phase transition for the scaling of ρ_+^- in the ECM. The first stage has scaling $N^{-1+1/(\gamma_+, \wedge \gamma_-)}$ and holds for all $\gamma_\pm$ for which

$$\frac{1}{\gamma_+ \wedge \gamma_-} - 1 \leq \frac{2}{\gamma_+} + \frac{2}{\gamma_-} - 3,$$

since both correspond to upper bounds. The next region, $\gamma_\pm$ such that $2/\gamma_+ + 2/\gamma_- - 3 \geq -1/2$, has scaling $N^{2/\gamma_+ + 2/\gamma_- - 3}$. Outside this region we have normal scaling, $N^{-1/2}$. The different regions are displayed in Figure 6, while Table 3 shows the three scaling terms. We remark that the phase transitions of the scaling are smooth since they are induced by inequalities on the terms.

5.3   Simulations

In order to show the phase transitions we plotted the empirical cumulative distribution function of ρ_+^- for the specific choices of $\gamma_\pm$, corresponding to the points I, II and III in Figure 6. For each of the three points we shifted the empirical data by its average and multiplied it by $N^{-f(\gamma_+, \gamma_-)}$, for any of the three coefficients from Table 3, corresponding to the different scaling areas A, B and C. The results are shown in Figure 7. When the correct scaling is applied, the corresponding cdf plots should almost completely overlap and resemble the cdf of some limiting distribution. We
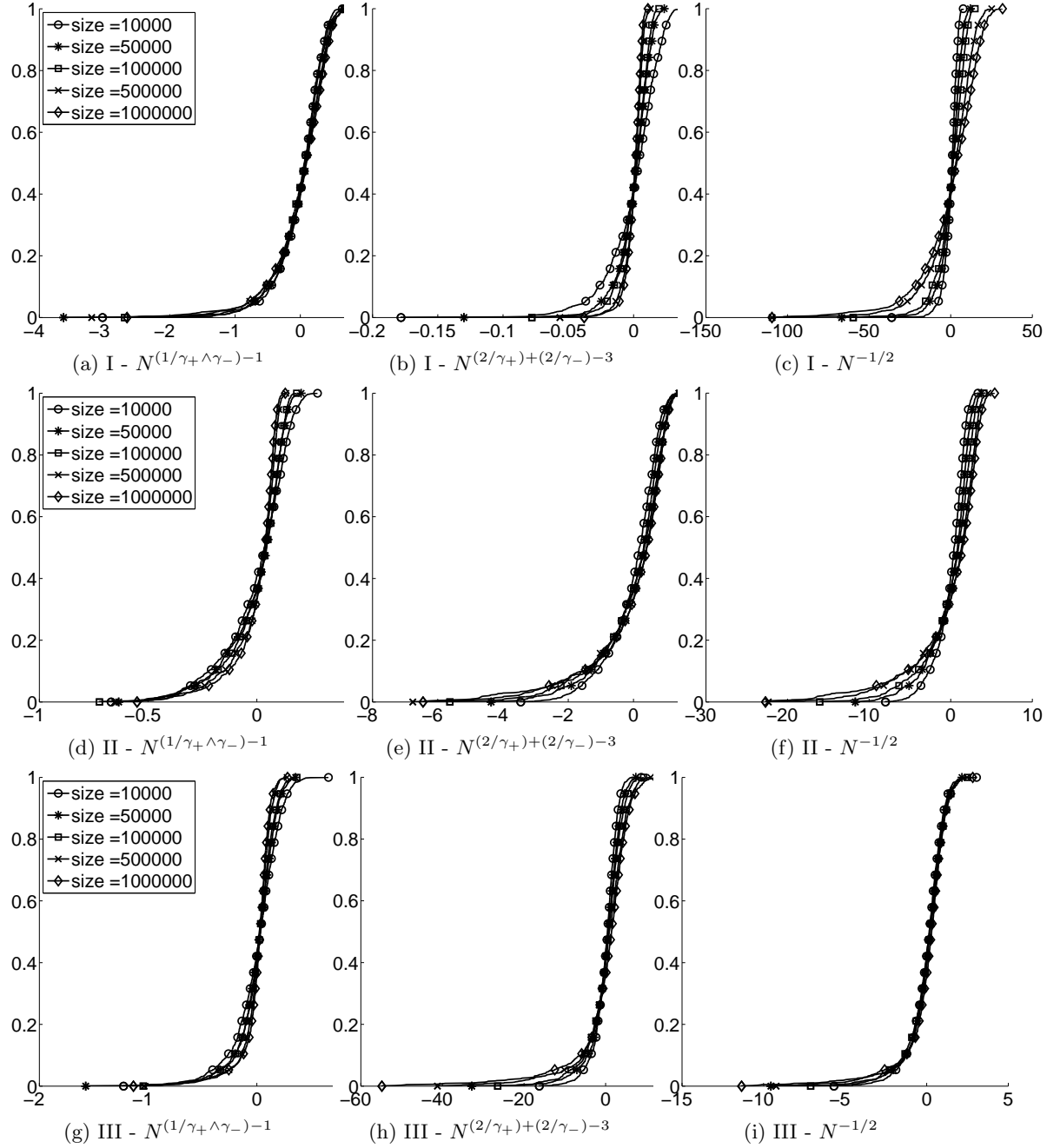
Figure 7: Plots of the empirical cumulative distribution function of ρ_+^- using different scaling and for different choices of $\gamma_\pm$. The left column is scaled by $N^{1/\min(\gamma_+,\gamma_-)}$, the center column by $N^{2/\gamma_+ + 2/\gamma_- - 3}$ and the right column by $N^{-1/2}$. The first row is for ECM graphs with $\gamma_\pm = 1.3$, the second for $\gamma_+ = 1.9$, $\gamma_- = 1.3$ and the third for $\gamma_+ = 1.9$, $\gamma_- = 1.5$, corresponding to points I, II and III, respectively in Figure 6.

observe that for each of the three choices I, II and III, this is the case when the corresponding scaling from its area, respectively A, B and C, is chosen.

6 Scaling of degree-degree dependencies for the other cases

In the previous section we completely characterized the scaling behavior of ρ_+^- for ECM graphs with infinite variance of the degrees. Here, we first discuss the remaining correlation types, ρ_+^+, ρ_-^- and ρ_-^+ in the infinite variance regime and lastly, we consider all four types in the finite variance regime.

The intuition behind the structural negative Out-In dependencies was that multiple edges are more likely to exist between nodes of large out- and in-degree. The other three types do not show negative correlations, see Figure 5b-5d, which we argued was due to the fact that the in- and out-degree of a node in the ECM are independent. Nevertheless, the spread of both the Out-Out and In-In degree-degree dependency exhibits scaling with the same functions as the Out-In dependency. This is illustrated in Figure 8, where we plotted the empirical cumulative distribution of the Out-Out dependency for ECM graphs, for values of $\gamma_\pm$ corresponding to points I, II and III from Figure 6, scaled by the correct term for each of these points. This is because ρ_+^+ again depends on the number of erased edges, through the out-degree of their target nodes. However, the out-degree of the source node of a removed edge can be both large or small, thus ρ_+^+ in the ECM remains zero on average. By symmetry, the scaling for the In-In dependency is similar.

This non-trivial scaling is typical for the ECM. Recall that in the CM, ρ_α^β is a normal statistic and scales as $N^{-1/2}$, for any, α, β because all degrees are independent random variables. This is exactly what we observe for the In-Out degree-degree dependency, which, in contrast to the other three, is not biased towards removed edges. As we expect, here we have normal, square root, scaling for ECM graphs for any choice of $\gamma_\pm$. This can clearly be observed in Figure 9, where we plotted the empirical cumulative distributions of ρ_-^+ scaled by $N^{-1/2}$.

For the degree-degree dependencies in the finite variance regime we plotted the empirical cumulative distributions of ρ_α^β, scaled by $N^{-1/2}$, in Figure 10. Since these are all completely similar, we took the plot for ρ_+^- for a ECM graph of size 10^6 and compared it to a fitted normal distribution with $\mu = 0$ and $\sigma^2 = 0.8$, see Figure 11. These plots strongly overlap, enforcing the claim that for ECM graphs with finite degree variance all four correlations are normal statistics.

7 Conclusion and Discussion

In this paper we analyzed degree-degree dependencies in the directed Erased Configuration Model. We showed, Figure 5, that in the infinite variance regime only the Out-In dependency exhibits structural negative values, while all correlations behave similar when both degrees have finite variance, Figure 4. We investigated the scaling of the structural negative Out-In correlations. These undergo a phase transition in terms of the exponents $\gamma_\pm$ of the degree distributions (3), which we showed by establishing two upper bounds, (10) and (16), on the total, average, removed number of edges, both of which scale at different rates. Combining this with the square root scaling of Spearman's rho and Kendall's tau, we identified three regions, depending on $\gamma_\pm$, with different scaling, Figure 6, and illustrated their phase transitions in Figure 7. Next, we considered the remaining three dependency types for the infinite variance regime. We showed, Figure 8, that the scaling of the Out-Out and In-In correlations behaves similarly to the Out-In, even though they do not exhibit structural negative values, while the In-Out degree-degree dependency has square root scaling, Figure 9. Finally we investigated the scaling for correlations when the degrees have finite variance. In this case all four types have square root scaling and the plots of the cumulative distributions are very similar, Figure 10. This was confirmed when we compared the plot of ρ_+^- for ECM graphs of size 10^6, with $\gamma_\pm = 2.1$, with that of a fitted normal distribution in Figure 11.

Our analysis shows that degree-degree dependencies in directed networks display non-trivial behavior in terms of scaling when the degrees have infinite variance. This scaling is important
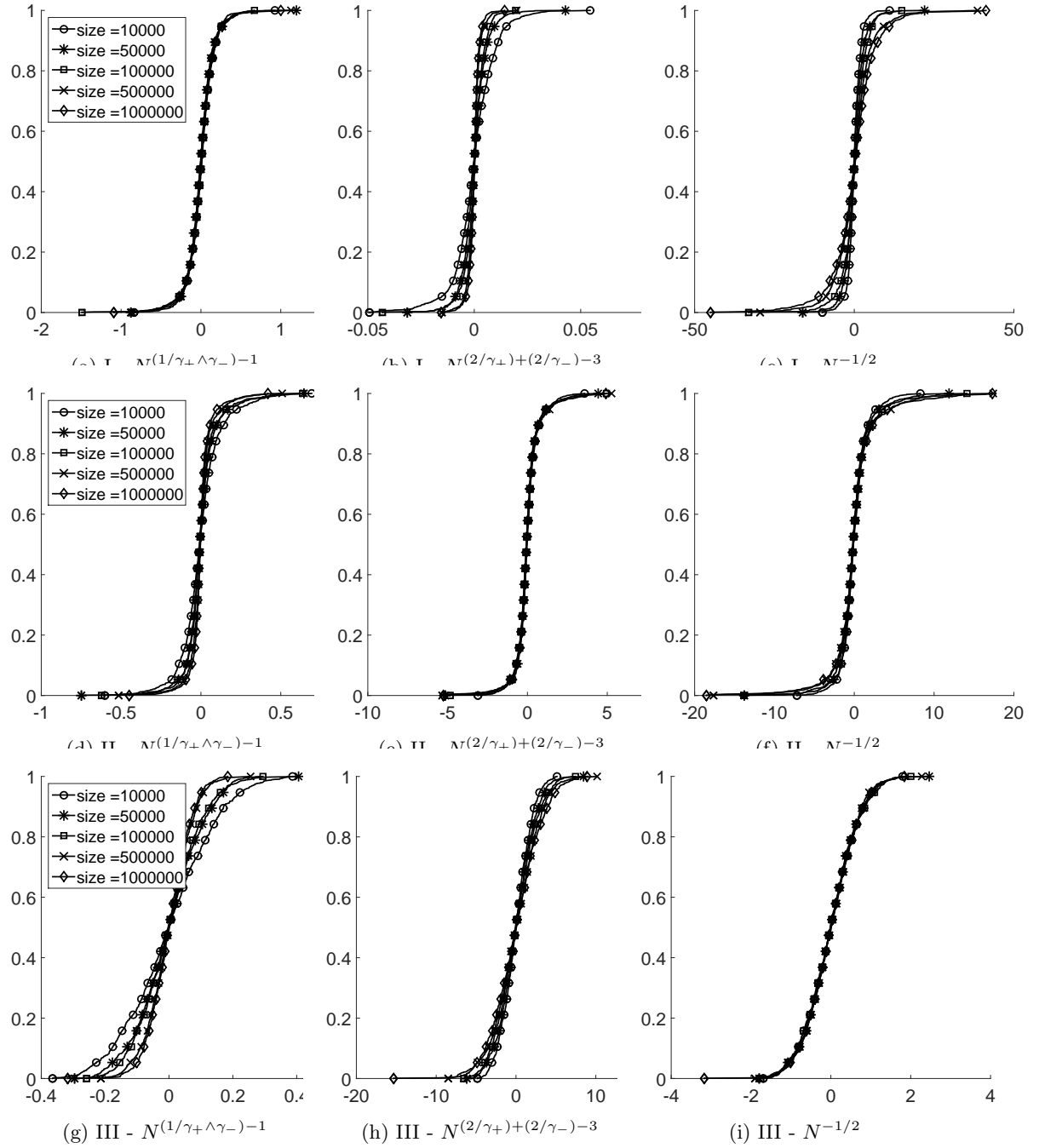
Figure 8: Plots of the empirical cumulative distribution function of ρ_+^+ for choices of $\gamma_\pm$ corresponding to points I, II and III from Figure 6, using the corresponding scaling.
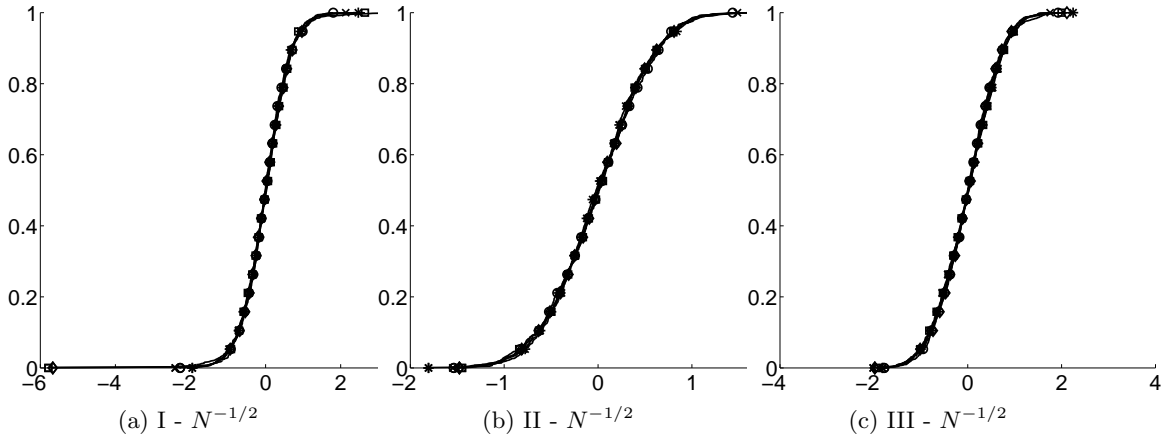
13

(a) I - $N^{-1/2}$ (b) II - $N^{-1/2}$ (c) III - $N^{-1/2}$

Figure 9: Plots of the empirical cumulative distribution function of ρ_-^+ for choices of $\gamma_\pm$ corresponding to points I, II and III from Figure 6, using square root scaling.

when doing statistical analysis of these measures or their impact on other processes on networks, for it determines their spread and hence enables to asses the significance of measurements.

We showed that degree-degree dependencies for degrees with finite variance, scaled by $N^{-1/2}$, converge to a normal distribution with zero mean. We have not yet been able to determine the variance of these distributions as a function of the tail exponents $\gamma_\pm$ which would completely characterize their behavior.

For three of the four correlation types in the infinite variance regime, we did not determine the limiting distributions. This is mainly due to the fact that we expect these to be *stable distributions*, since one of the three scaling regions is due to the Central Limit Theorem for stable random variables. Although these distributions have a well defined characteristic function, their density function, in general, does not have an analytical expression. Moreover, we are dealing with discrete data and simulation of such distributions is a field of it's own. Nevertheless, we do expect that Central Limit Theorems for degree-degree dependencies can be formulated and proven, which would fully complete their statistical analysis.

Finally, our empirical results clearly show the, analytically derived, phase transitions. However, the region with the $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ scaling is less distinct than the other two. One of the possible reasons for this is that within the area where this scaling applies, the difference in value with the other two terms is small. We therefore picked point II in Figure 6 such that this difference was large enough to distinctly show this scaling visually in the plots.

We close by strongly suggesting to use the ECM as a null model for analysis of degree-degree dependencies, both for determining their impact on processes as well as significance. Although for the latter, values are often compared to averages, using the rewiring model [17], we emphasize that fixing the degrees imposes strong constraints on the possible simple graphs that can be generated. Moreover, in real-life networks, not only wiring but also the degrees of the nodes, are a result of a random process. Therefore, in a null-model, it seems more natural to fix only general properties of the network, such as degree distributions.

Acknowledgments:

References

[1] Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2011.
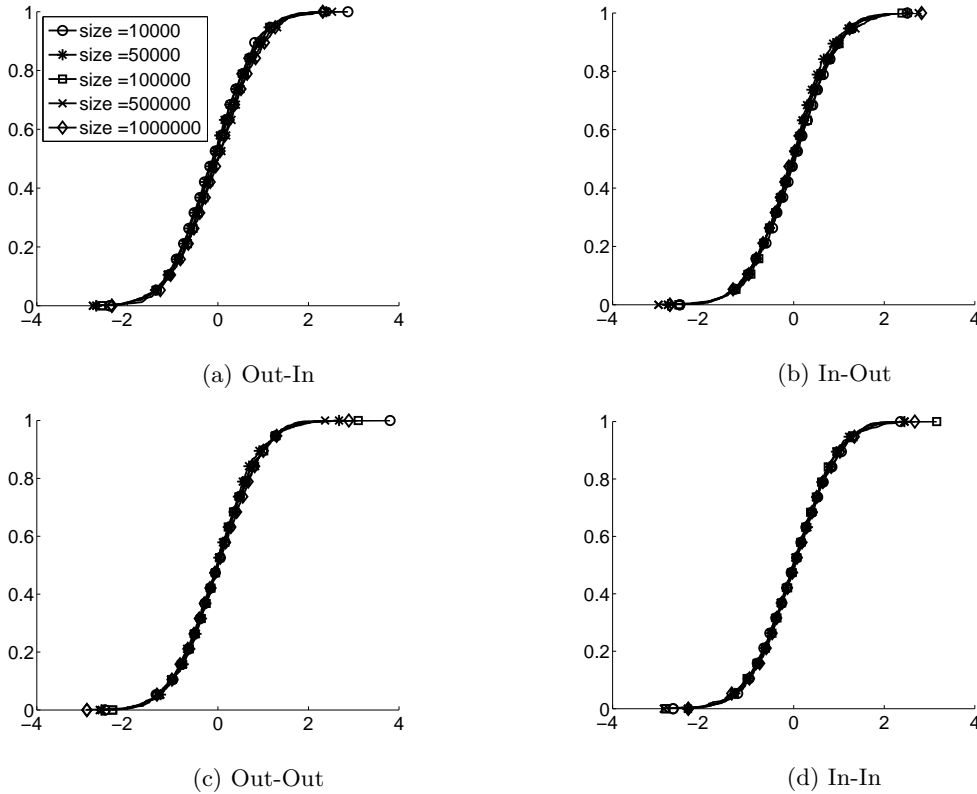
Figure 10: Plots of the empirical cumulative distribution of ρ_α^β for all four degree-degree dependency types for ECM graphs with $\gamma_\pm = 2.1$ of different sizes, scaled by $N^{-1/2}$. Each plot is based on 10^3 realizations of the model.

[2] Marián Boguná and Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66(4):047104, 2002.

[3] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical review letters*, 90(2):028701, 2003.

[4] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Cut-offs and finite size effects in scale-free networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):205–209, 2004.

[5] 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.

[6] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[7] Michele Catanzaro, Marián Boguñá, and Romualdo Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Physical Review E*, 71(2):027103, 2005.

[8] Ningyuan Chen and Mariana Olvera-Cravioto. Directed random graphs with given degree distributions. *Stochastic Systems*, 3(1):147–186, 2013.

[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
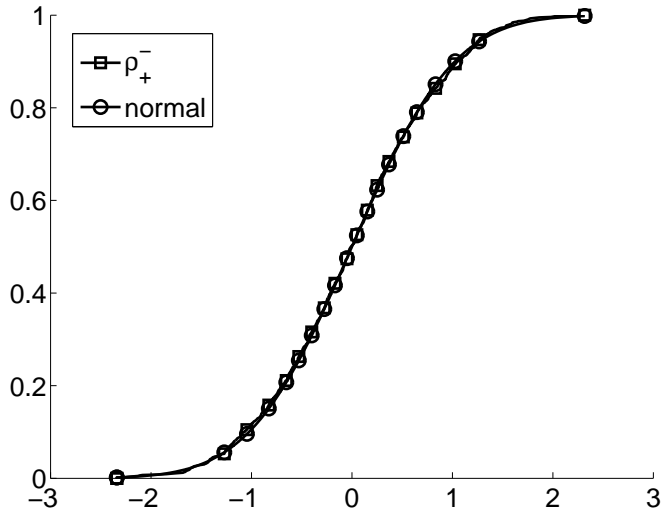
Figure 11: Plot of the empirical cumulative distribution function of ρ_+^- for ECM graphs of size 10^6 with $\gamma_\pm = 2.1$ and a normal cumulative distribution with $\mu = 0$ and $\sigma^2 = 0.8$.

[10] Sergey N Dorogovtsev and Jose FF Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

[11] SN Dorogovtsev, AL Ferreira, AV Goltsev, and JFF Mendes. Zero pearson coefficient for strongly correlated growing trees. *Physical Review E*, 81(3):031135, 2010.

[12] 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.

[13] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325, 1948.

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

[15] Nelly Litvak, Werner RW Scheinhardt, and Yana Volkovich. In-degree and pagerank: Why do they follow similar power laws? *Internet Mathematics*, 4(2-3):175–198, 2007.

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

[17] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[18] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333:529–540, 2004.

[19] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

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

[21] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[22] Juyong Park and Mark EJ Newman. Origin of degree correlations in the internet and other networks. *Physical Review E*, 68(2):026112, 2003.

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

[24] Tiziano Squartini and Diego Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8):083001, 2011.

[25] Animesh Srivastava, Bivas Mitra, Fernando Peruani, and Niloy Ganguly. Attacks on correlated peer-to-peer networks: An analytical study. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 1076–1081. IEEE, 2011.

[26] Remco van der Hofstad, Gerard Hooghiemstra, and Piet Van Mieghem. Distances in random graphs with finite variance degrees. *Random Structures & Algorithms*, 27(1):76–123, 2005.

[27] Pim van der Hoorn and Nelly Litvak. Convergence of rank based degree-degree correlations in random directed networks. *arXiv preprint arXiv:1407.7662*, 2014.

[28] Pim van der Hoorn and Nelly Litvak. Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet Mathematics*, (just-accepted):00–00, 2014.

[29] Alexei Vázquez and Yamir Moreno. Resilience to damage of graphs with degree correlations. *Phys. Rev. E*, 67:015101, Jan 2003.

[30] Ward Whitt. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer, 2002.