The Case for Kendall's Assortativity

Paolo Boldi Sebastiano Vigna
Dipartimento di Informatica, Università degli Studi di Milano

January 22, 2020

Abstract

Since the seminal work of Litvak and van der Hofstad [LvdH13], it has been known that Newman's assortativity [New02, New03], being based on Pearson's correlation, is subject to a pernicious size effect which makes large networks with heavy-tailed degree distributions always unassortative. Usage of Spearman's ρ, or even Kendall's τ was suggested as a replacement [vdHL15], but the treatment of ties was problematic for both measures. In this paper we first argue analytically that the tie-aware version of τ solves the problems observed in [vdHL15], and we show that Newman's assortativity is heavily influenced by tightly knit communities. Then, we perform for the first time a set of large-scale computational experiments on a variety of networks, comparing assortativity based on Kendall's τ and assortativity based on Pearson's correlation, showing that the pernicious effect of size is indeed very strong on real-world large networks, whereas the tie-aware Kendall's τ can be a practical, principled alternative.

1 Introduction

Assortativity (or *assortative mixing*) is a property of networks in which similar nodes are connected. More in detail, here we consider the *degree assortativity* of (directed) networks, that looks at whether the indegree/outdegree of x is correlated to the indegree/outdegree of y over all the arcs $x \to y$ of the network. One has thus four types of assortativity (denoted by $+/+$, $-/+$, $-/-$ and $+/-$), and for each type one has to choose which measure of correlation should be used between the lists of degrees at the start/end of each arc. The classical definition of assortativity given by Newman [New02, New03] employed *Pearson's correlation*.

In a seminal paper, Litvak and van der Hofstad [LvdH13] have shown analytically that on pathological examples and on heavy-tailed undirected networks Pearson's degree-degree assortativity tends to zero as the network gets large, because of a size effect induced by the denominator of the formula. They go on to suggest to use of Spearman's ρ (which is Pearson's correlation on the rank of the values) to correct this defect.

A subsequent paper [vdHL15] has shown that the same problem plagues Pearson's degree-degree correlation in the directed case: the authors explore also briefly, besides the ρ coefficient, the possibility of using Kendall's τ, but they do not completely resolve the problem of ties. In their work, they suggest to use averages or randomization to correct for the presence of ties in the case of Spearman's ρ, and simply neglect them in the case of τ, noting that this choice constraints (in a negative way) the possible values that τ can assume.

In this paper, first we extend analytically some of the results above, showing that the version of Kendall's τ that takes ties into consideration (sometimes called τ_b) has the same (better) behavior as

Spearman's ρ on the pathological examples, and thus does not suffer from the limitations highlighted in [vdHL15]. Moreover, we show that tightly knit communities can influence in pernicious ways assortativity. Then, we perform several computational experiments on various web graphs and other types of complex networks, computing both Pearson's and Kendall's degree-degree correlations: by looking inside the data, we confirm the bias in the former when large networks are involved, and also give empirical evidence of the effect of tightly knit communities.

All data used in this paper are publicly available from the LAW, and the code used for the experiments is available as free software as part of the LAW Library.[1]

2 Definitions and conventions

In this paper, we consider directed graphs defined by a set N of n nodes and a set $A \subseteq N \times N$ of arcs; we write $x \to y$ when $a = \langle x, y \rangle \in A$ and call x (respectively, y) the source (respectively, target) of the arc a, denoted by $s(a)$ (respectively, $t(a)$). Our graphs can contain *loops* (arcs a such that $s(a) = t(a)$). A *successor* of x is a node y such that $x \to y$, and a *predecessor* of x is a node y such that $y \to x$. The *outdegree* $d^+(x)$ of a node x is the number of its successors, and the *indegree* $d^-(x)$ is the number of its predecessors.

A *symmetric graph* is a graph such that $x \to y$ whenever $y \to x$; such a graph can be identified with an undirected graph, that is, a graph whose arcs (usually called *edges*) are subsets of one or two nodes. In fact, in the following, all definitions are given for directed graphs, and apply to undirected graphs through their loopless symmetric representation.

3 Assortativity

Degree-degree assortativity measures the propensity of nodes to create links to nodes with similar degrees. Since we consider directed graphs, there are four types of assortativity: outdegree/outdegree, indegree/outdegree, indegree/indegree, and outdegree/indegree, denoted by $+/+$, $-/+$, $-/-$, and $+/-$, respectively. As noted in [vdHL15], the only case in which an arc contributes to the degrees on both of its sides is $+/-$, which makes the $+/-$ case the natural generalization of the undirected case [LvdH13]. The assortativity is defined as a measure of correlation between the appropriate list of degrees at the two sides of the list of all arcs of the graph. More precisely, for any given correlation index c, and every choice of $\alpha, \beta \in \{+, -\}$, we define the *c-assortativity of type* (α, β) as

$$c_\alpha^\beta(G) = c\left(\left[d^\alpha(s(a))\right]_{a\in A}, \left[d^\beta(t(a))\right]_{a\in A}\right)$$

where $[-]_{a\in A}$ are used to denote a list ranging over all arcs of the graph (the order is immaterial, provided that it is coherent, i.e., the same for both lists).

Newman's definition of assortativity [New02, New03] uses Pearson's correlation coefficient for c. However, there are many other possibilities to measure the correlation between degrees. One can use, as an alternative, Spearman's ρ [Spe04], which is Pearson's correlation between the *ranks* of the values in the lists: this choice has some advantages, but it does not provide a solution for *ties*, that is, duplicate degrees.

Correct handling of ties in degree lists is of utter importance because in a real-world graph a large percentage of the arcs is involved in a tie (e.g., the outdegree of the target of all arcs pointing at the

[1] http://law.di.unimi.it/

same large-indegree node are the same). In [vdHL15], the authors resort to typical solutions such as averaging or randomization, which however have been shown to be detrimental and, in fact, decrease the amount of correlation [Vig15].

An alternative that handles ties in a principled way is Kendall's τ correlation index, when formulated properly (see next section). However, the authors of [vdHL15] used a formulation that had been designed for lists without ties, obtaining pathological results. We are going to show that this problem can be fixed using a proper version, thus providing more proper handling of degree ties.

4 Kendall's τ, 1945

The original and most commonly known definition of Kendall's τ is given in terms of concordances and discordances. We consider two real-valued vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ (to be thought of as scores) of n elements, and assume that no score appears twice in a vector (i.e., there are no *ties*). We say that a pair of indices $\langle i, j\rangle$, $0 \le i < j < n$, is *discordant* if $s_i < s_j$ and $t_i > t_j$, or $s_i > s_j$ and $t_i < t_j$, *concordant* otherwise. Then, the τ between the two vectors is given by the number of concordances, minus the number of discordances, divided for the number of pairs. Note that all scores must be distinct.

In his 1945 paper about ranking with ties [Ken45], Kendall, starting from an observation of Daniels [Dan43], reformulates his correlation index using a definition similar in spirit to that of an inner product. Let us define

$$\langle \boldsymbol{r}, \boldsymbol{s}\rangle := \sum_{i<j} \operatorname{sgn}(r_i - r_j)\operatorname{sgn}(s_i - s_j),$$

where

$$\operatorname{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Note that the expression above is actually an inner product in a larger space of dimension $n(n-1)/2$: each score vector $\boldsymbol{r}$ is mapped to the vector with coordinate $\langle i, j\rangle$, $i < j$, given by $\operatorname{sgn}(r_i - r_j)$. Thus, we can define

$$\tau(\boldsymbol{r}, \boldsymbol{s}) := \frac{\langle \boldsymbol{r}, \boldsymbol{s}\rangle}{\sqrt{\langle \boldsymbol{r}, \boldsymbol{r}\rangle} \cdot \sqrt{\langle \boldsymbol{s}, \boldsymbol{s}\rangle}}. \tag{1}$$

Essentially, we are defining a cosine similarity, which we can compute easily as follows: given a pair of distinct indices $0 \le i, j < n$, we say that the pair is

- *concordant* iff $r_i - r_j$ and $s_i - s_j$ are both nonzero and have the same sign;
- *discordant* iff $r_i - r_j$ and $s_i - s_j$ are both nonzero and have opposite signs;
- a *left tie* iff $r_i - r_j = 0$;
- a *right tie* iff $s_i - s_j = 0$.

Let C, D, T_r, T_s be the number of concordant pairs, discordant pairs, left ties, right ties, and joint ties, respectively. We have

$$\tau(\boldsymbol{r}, \boldsymbol{s}) = \frac{C - D}{\sqrt{\binom{n}{2} - T_r}\sqrt{\binom{n}{2} - T_s}}.$$

Note that a pair that is at the same time a left tie and a right tie (a so-called *joint tie*) will be counted both in T_r and in T_s.

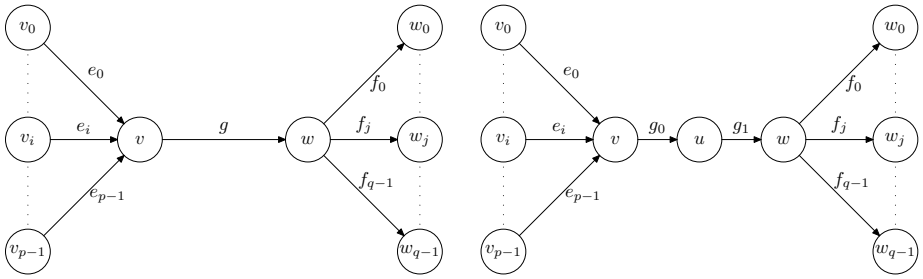

Figure 1: The graphs $G(p, q)$ and $\hat{G}(p, q)$.

5 The Case for Ties in Kendall's Assortativity

As an example of the importance of ties in the computation of Kendall's assortativity, we consider the graphs $G(p, q)$ and $\hat{G}(p, q)$ defined in [vdHL15][Section 5.1] and shown in Figure 1. The graphs are made by a directed 1-path or 2-path (whose arcs we will call g, or g_0 and g_1, respectively), with p further nodes $v_0, v_1, \ldots, v_{p-1}$ pointing to the source of the path (the p corresponding arcs are named $e_0, e_1, \ldots, e_{p-1}$) and q nodes $w_0, w_1, \ldots, w_{q-1}$ pointed by the target of the path (the q corresponding arcs are named $f_0, f_1, \ldots, f_{q-1}$). Overall, the graph has $G(p, q)$ has $p + q + 1$ arcs, whereas $\hat{G}(p, q)$ has $p + q + 2$ arcs.

On these graphs, Pearson's assortativity of type $-/+$ behaves in a completely pathological way: for $G(n, an)$, it tends to 1 as $n \to \infty$, because the mass associated with the arc g becomes very large, due to its very large indegree, whereas the assortativity of $\hat{G}(n, an)$ tends to 0 [vdHL15]. This is extremely counterintuitive, because the two networks are almost identical. In particular, in both cases one expects to measure a significant disassortativity[2], because a large fraction of arcs are disassortative (both the e_i's and the f_j's are such).

Using Spearman's ρ makes assortativity correctly tend to -1 in both cases if one solves ties by giving equal rank to equal elements; one has, however, widely different results with different methods for ranking ties.

The limit of Kendall's τ as $n \to \infty$ without taking ties into consideration is $-2a/(a + 1)^2$, which tends to zero as a grows. The authors comment that this is due to the influence of ties, and indeed we are going to show that using the correct tie-aware version Kendall's τ solves the problem: the network becomes disassortative, as the fraction of concordant pairs goes to zero whereas the fraction of discordant pairs does not. This happens naturally, without having to choose a policy for ties.

Looking again at the $-/+$ assortativity, we see that in both graphs there are p arcs (the arcs e_i) with degree pairs $\langle 0, 1\rangle$ and q arcs (the arcs f_j) with degree pairs $\langle 1, 0\rangle$. Finally, in $G(p, q)$ we have one arc with degree pair $\langle p, q\rangle$, whereas in $\hat{G}(p, q)$ we have one arc with degree pair $\langle p, 1\rangle$ and one arc with degree pair $\langle 1, q\rangle$. We will assume $2 < p < q$ in the following.

In $G(p, q)$ we have:

- $p + q$ concordances, given by the pairs $\langle e_i, g\rangle$ and $\langle f_j, g\rangle$.
- pq discordances, given by the pairs $\langle e_i, f_j\rangle$.
- $\binom{p}{2} + \binom{q}{2}$ left ties, given by distinct pairs of e_i's, and distinct pairs of f_j's.

[2]We use "unassortative" for networks with a correlation close to 0, and "disassortative" for networks with a correlation close to -1.

- $\binom{p}{2} + \binom{q}{2}$ right ties, given by the same pairs.

All in all, we thus have

$$\tau_{G(p,q)} = \frac{p + q - pq}{\sqrt{\left(\binom{p+q+1}{2} - \binom{p}{2} - \binom{q}{2}\right) \cdot \left(\binom{p+q+1}{2} - \binom{p}{2} - \binom{q}{2}\right)}}.$$

In $\hat{G}(p, q)$, by an analogous analysis we have

$$\tau_{\hat{G}(p,q)} = \frac{p + q - pq - 1}{\sqrt{\left(\binom{p+q+2}{2} - \binom{p}{2} - \binom{q}{2} - q\right) \cdot \left(\binom{p+q+2}{2} - \binom{p}{2} - \binom{q}{2} - p\right)}}.$$

If we consider the case $p = n$, $q = an$ with a constant and $n \to \infty$, as in [vdHL15], we have

$$\tau_{G(n,an)} = \frac{n + an - an^2}{\sqrt{\left(\binom{n+an+1}{2} - \binom{n}{2} - \binom{an}{2}\right) \cdot \left(\binom{n+an+1}{2} - \binom{n}{2} - \binom{an}{2}\right)}} \to -1$$

$$\tau_{\hat{G}(n,an)} = \frac{n + an - an^2 - 1}{\sqrt{\left(\binom{n+an+2}{2} - \binom{n}{2} - \binom{an}{2} - an\right) \cdot \left(\binom{n+an+2}{2} - \binom{n}{2} - \binom{an}{2} - n\right)}} \to -1$$

Thus, the proper definition aligns on this example Kendall's τ with the results from Spearman's ρ, using constant ranks for ties.[3]

6 The Tightly Knit Community Effect, Again

We are now going to discuss another, and possibly more pernicious, effect of size on Pearson's assortativity. This phenomenon is akin to the well-known *tightly knit community* (TKC) effect on certain ranking algorithms such as HITS [Kle99]: a small group of tightly connected users ends up being ranked unfairly high. For this section, we consider undirected graphs, as it is much simpler to compute Pearson's assortativity using the formulae from [VMWG+10], but the same considerations apply to directed graphs.

Let us start from the graph $H(p, q)$ defined as a (p, q) complete bipartite graph, in which the q nodes are further completely connected (i.e., they form a q-clique). This graph (the red and blue nodes in Figure 2) is highly unassortative, in the sense that its Pearson's assortativity, which is $p/(1 - p - q)$, goes to zero if $q, p \to \infty$, as long as $p = o(q)$.

Let us now consider a graph $\hat{H}(p, q, k)$ formed by $H(p, q)$ plus a (disjoint) clique of k vertices (see Figure 2). Our interest is in measuring how much the clique (the most assortative graph) will influence the unassortative graph $H(p, q)$. In particular, we will look at $\hat{H}(p, p^2, p^{1/2+\varepsilon})$: this is the very unassortative graph $H(p, p^2)$, with $p + p^2$ nodes, to which we are adding a clique of size $p^{1/2+\varepsilon}$ (note that the number of added nodes for small ε is negligible in the size of $H(p, p^2)$).

[3] We mention that also Goodman–Kruskal's γ [GK54], defined as the difference between concordances and discordances divided by their sum, provides a principled treatment of ties. However, Kendall's τ has some advantages, and in particular the possibility of defining tie-aware weighted versions [Vig15].

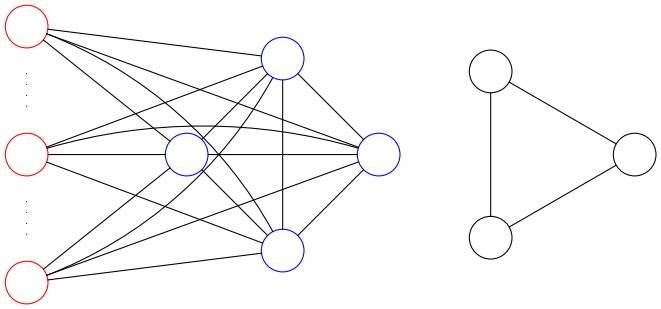

Figure 2: The graph $H(p, q, k)$. There are p red nodes (left), q blue nodes (center) and k black nodes (right).

The formula for the Pearson's assortativity of $H(p, q, k)$ is

$$1 - \frac{pq(p-1)^2}{(k(k-1)^3 + pq^3 + qs^3 - \frac{1}{2pq+q(q-1)+k(k-1)}\left(pq^2 + qs^2 + k(k-1)^2\right)^2},$$

where $s = q + p - 1$, and dominant term of $H\left(p, p^2, p^{1/2+\varepsilon}\right)$ as $p \to \infty$ is

$$\frac{p^{2\varepsilon}}{1 + p^{2\varepsilon}}. \tag{2}$$

Thus, we have a threshold effect as $p \to \infty$: for $\varepsilon < 0$, the network becomes unassortative; for $\varepsilon = 0$, assortativity tends to $1/2$; but for $\varepsilon > 0$, the network will become completely assortative. In other words, *a tightly knight community of order* $\Omega\left(n^{1/4+\varepsilon}\right)$ *can drive a large unassortative network to assortativity*. This impressive effect is evidently pathological.

Is there a similar phenomenon for Kendall's assortativity? The value of Kendall's assortativity on $H(p, q)$ is (maybe surprisingly) the same of Pearson's assortativity, whereas the Kendall's assortativity of $\hat{H}(p, q, k)$ is

$$\frac{k(k-1)(2p+q-1) - qp^2}{k(k-1)(2p+q-1) + pq(q+p-1)}.$$

When we examine $H\left(p, p^2, p^{1/2+\varepsilon}\right)$ and let $p \to \infty$ we find a completely different situation, as assortativity tends to zero as $-1/p$. However, the leading term of $H\left(p, p^2, p^{3/2+\varepsilon}\right)$ *is again* (2), and a similar transition effect appears.

In other words, *also Kendall's assortativity is subject to the TKC effect, but one needs a community asymptotically much larger to obtain the same effect*, that is, $\Omega\left(n^{3/4+\varepsilon}\right)$ vs. $\Omega\left(n^{1/4+\varepsilon}\right)$. As an example, the graph $H(100, 10000)$ has a (Pearson's and Kendall's) assortativity of -0.010, but if look at $\hat{H}(100, 10000, 200)$ (i.e., we add a clique of 200 nodes, increasing the size of the graph by less than 2%) we get an impressive increase in Pearson's assortativity (it goes up to 0.997), whereas the Kendall's assortativity is still very small (0.029).

7 Experiments

We consider a set of networks available from the repository of the Laboratory for Web Algorithmics.[4] The graphs are listed and briefly described in Table 1, which shows some of their basic properties:

[4] http://law.di.unimi.it/datasets.php

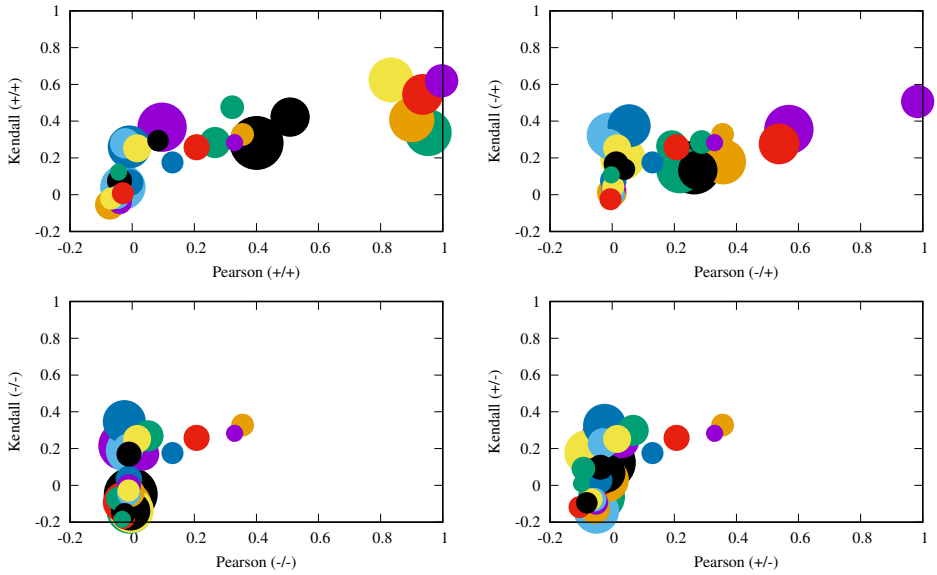

Figure 3: Plots displaying the correlation between Pearson's and Kendall's assortativity.

more information can be found on the repository website. The list includes a variety of types of graphs, both directed and undirected (i.e., symmetric), including web crawls, host graphs, Wikipedia graphs, social networks (e.g., Twitter), telephone-call graphs, and co-authorship/co-starship graphs; their sizes range from a few hundred thousands to billion of edges.

We computed assortativity values based on Pearson's correlation and on Kendall's τ using the implementations available in the LAW library, and we report the values in Table 2. Large differences (≥ 0.20) are shown in boldface.

First of all, we remark that we can confirm the results about Wikipedia graphs discussed in [vdHL15]: they are unassortative for all types. However, when we consider social networks such as LiveJournal, Twitter and Orkut, where we do expect some kind of assortativity, Kendall's assortativity provides significant larger values, proving for the first time that on real-world networks the size effect is substantial, as it makes such networks appear unassortative according to Pearson's correlation.

In Figure 3 we show a scatter plot of the values obtained by Pearson's correlation versus those obtained by Kendall's τ, sizing the dots depending on the number of arcs of the graph. In the $-/-$ and $+/-$ case one can see immediately the strip of small graphs on the diagonal showing correlation, and the pile of large graphs, all stationing around the value zero of Pearson's correlation.

There is of course another feature that is evident: several web graphs have incredibly large assortativity of type $+/+$ (e.g., `indochina-2004` has Pearson's assortativity 0.9965 and Kendall's assortativity 0.6195), which shows up as the block of large circles in the upper right part of the top left graph of Figure 3. Is it really possible that web pages tend to link mostly to pages with the same number of links?

The answer is no: the preposterously high score we are observing are simply due to the TKC effect. Extremely connected sites (e.g., machine-generated tables or calendars) can have a very strong impact on assortativity. For example, the densest website of non-negligible size (probably a link farm) in `indochina-2004` contains 7 611 nodes and 48 231 874 arcs (a density of 83%!). To study the role of such dense websites, we can try to remove them from the graph (or, in the opposite direction, to add

Table 1: The graphs used in the experiments.

Name	Nodes	Arcs	
arabic-2005	22 744 080	639 999 458	A crawl of Arabic countries [BCSV04]
cnr-2000	325 557	3 216 152	A crawl of the CNR [BCSV04]
dblp-2010	326 186	1 615 400	The co-authorship graph from DBLP
dblp-2011	986 324	6 707 236	The co-authorship graph from DBLP
dewiki-2013	1 532 354	36 722 696	German Wikipedia
enwiki-2013	4 206 785	101 355 853	English Wikipedia
eswiki-2013	972 933	23 041 488	Spanish Wikipedia
frwiki-2013	1 352 053	34 378 431	French Wikipedia
eu-2005	862 664	19 235 140	A crawl of .eu [BMSV19]
gsh-2015-host	68 660 142	1 802 747 600	Host graph of a general crawl [BMSV19]
gsh-2015-tpd	30 809 122	602 119 716	Top domains of a general crawl [BMSV19]
hollywood-2009	1 139 905	113 891 327	The co-starship graph from the IMDB
hollywood-2011	2 180 759	228 985 632	The co-starship graph from the IMDB
hu-tel-2006	2 317 492	46 126 952	Call graph from Hungarian Telekom [KBCL07]
in-2004	1 382 908	16 917 053	A crawl of .in [BCSV04]
indochina-2004	7 414 866	194 109 311	A crawl of Indochina [BCSV04]
it-2004	41 291 594	1 150 725 436	A crawl of .it [BCSV04]
itwiki-2013	1 016 867	25 619 926	Italian Wikipedia
ljournal-2008	5 363 260	79 023 142	LiveJournal [CKL+09]
orkut-2007	3 072 626	234 370 166	The Orkut social network [MMG+07]
sk-2005	50 636 154	1 949 412 601	A crawl of .sk [BCSV04]
twitter-2010	41 652 230	1 468 365 182	Twitter [KLPM10]
uk-2002	18 520 486	298 113 762	A crawl of .uk [BCSV04]
uk-2005	39 459 925	936 364 282	A crawl of .uk [BCSV04]
uk-2014-host	4 769 354	50 829 923	Host graph of .uk [BMSV19]
uk-2014-tpd	1 766 010	18 244 650	Top domains of .uk [BMSV19]
webbase-2001	118 142 155	1 019 903 190	A crawl from the Stanford WebBase

Table 2: Pearson's and Kendall's assortativity values for the graphs of Table 1. Boldfaced entries have a difference larger than 0.20. Note that all values for the undirected graphs (e.g., `orkut-2007`) are all identical.

	+/+		−/+		−/−		+/−	
Name	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
hu-tel-2006	−0.0063	0.0660	0.0037	0.0768	−0.0107	0.0349	−0.0418	0.0266
ljournal-2008	0.2665	0.2835	0.1919	0.2656	**0.0508**	**0.2689**	**0.0674**	**0.2989**
twitter-2010	−0.0301	0.0410	**−0.0089**	**0.3232**	**−0.0121**	**0.1886**	−0.0506	−0.1395
orkut-2007	**0.0158**	**0.2528**	**0.0158**	**0.2528**	**0.0158**	**0.2528**	**0.0158**	**0.2528**
dblp-2010	0.3300	0.2827	0.3300	0.2827	0.3300	0.2827	0.3300	0.2827
dblp-2011	0.1296	0.1757	0.1296	0.1757	0.1296	0.1757	0.1296	0.1757
hollywood-2009	0.3555	0.3278	0.3555	0.3278	0.3555	0.3278	0.3555	0.3278
hollywood-2011	0.2073	0.2585	0.2073	0.2585	0.2073	0.2585	0.2073	0.2585
enwiki-2013	−0.0715	−0.0542	−0.0007	0.0126	−0.0077	−0.0385	−0.0553	−0.1287
frwiki-2013	−0.0469	−0.0136	0.0028	0.0037	−0.0110	−0.0458	−0.0564	−0.0778
dewiki-2013	−0.0398	−0.0395	0.0052	0.0295	−0.0109	−0.0058	−0.0518	−0.0934
eswiki-2013	−0.0301	0.0078	−0.0054	−0.0244	−0.0261	−0.1774	−0.1049	−0.1183
webbase-2001	0.4005	0.2817	0.2635	0.1483	−0.0048	−0.0486	−0.0107	0.1234
arabic-2005	**0.9350**	**0.5456**	**0.5378**	**0.2766**	−0.0288	−0.0916	−0.0539	0.0406
indochina-2004	**0.9965**	**0.6195**	**0.9837**	**0.5068**	0.0333	0.1704	**0.0332**	**0.2397**
eu-2005	**0.0832**	**0.2942**	0.0394	0.1388	−0.0239	−0.1541	−0.0815	−0.0947
in-2004	0.3219	0.4761	0.2883	0.2865	−0.0458	−0.0722	−0.0925	0.0908
it-2004	**0.9003**	**0.4083**	0.3582	0.1790	−0.0113	−0.1033	−0.0191	0.0304
sk-2005	**0.9534**	**0.3375**	0.2177	0.1353	−0.0078	−0.1351	−0.0343	−0.0633
uk-2002	0.5083	0.4219	0.2755	0.1338	−0.0050	−0.1383	−0.0209	0.0679
uk-2005	**0.8334**	**0.6246**	0.0337	0.2010	−0.0031	−0.1364	**−0.0818**	**0.1781**
cnr-2000	−0.0439	0.1237	−0.0027	0.1079	−0.0317	−0.1850	−0.0986	0.0099
itwiki-2013	−0.0678	−0.0209	0.0034	0.0429	−0.0114	−0.0284	−0.0666	−0.0758
uk-2014-host	**−0.0221**	**0.2792**	**−0.0093**	**0.2743**	**−0.0123**	**0.2005**	**−0.0296**	**0.2282**
gsh-2015-host	**0.0962**	**0.3674**	**0.5688**	**0.3547**	**−0.0297**	**0.2159**	**−0.0205**	**0.2558**
uk-2014-tpd	−0.0406	0.0755	0.0128	0.1679	−0.0100	0.1695	−0.0376	0.0988
gsh-2015-tpd	**−0.0092**	**0.2595**	**0.0544**	**0.3749**	**−0.0252**	**0.3471**	**−0.0243**	**0.3263**

Table 3: Pearson's and Kendall's assortativity values for variants of the graph G of `indochina-2004`. Here H_1 and H_2 are the densest and second-densest non-negligible websites of this web graph, whereas K_t is a clique of size t.

	$G - H_{1,2}$		$G - H_1$		G		$G + K_{1000}$		$G + K_{20\,000}$	
	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
$+/+$	0.5659	0.4167	0.7784	0.4342	0.9965	0.6195	0.9965	0.6233	0.9998	0.8026
$-/+$	0.2070	0.2005	0.3622	0.2252	0.9837	0.5068	0.9837	0.5118	0.9993	0.7917
$-/-$	-0.0200	-0.1344	-0.0269	-0.1262	0.0333	0.1704	0.0337	0.1717	0.5596	0.7075
$+/-$	-0.0672	0.0027	-0.0665	0.0063	0.0332	0.2397	0.0335	0.2403	0.5597	0.7112

a fictitious large clique) and see how this operation impacts on assortativity. Table 3 shows the results for `indochina-2004`:

- The TKC effect is evident for $+/+$ and $-/+$, but the increase is more dramatic in Pearson's than in Kendall's assortativity (in line with the discussion of Section 6).
- The same observation holds for $-/-$ and $+/-$ but in this case the phenomenon in Pearson's assortativity is diluted by the size effect: this web graph contains pages extremely large indegree and zero outdegree (simply because the crawl was stopped before the outlinks of those pages could be fetched); the arcs toward these pages contribute to a very large second component in $-/-$ and $+/-$ (whereas they do not show up in $+/+$ and $-/+$).
- Once we remove the noise, the Kendall $-/-$ values tend to *disassortativity*, which is actually what we expect from a web graph (highly pointed nodes of influential websites are, by and large, linked by "normal" nodes): in other words, the noise from the TKC and the size effects completely hide the actual disassortative nature of the network.

Other web crawls behave similarly: the take-home message here is that one should be very cautious when using assortativity values measured on noisy large-scale data such as web crawls, and that, in any case, Kendall's τ is more robust and less sensitive to the TKC and size effects. In fact, computing *both* measures is an excellent way to spot anomalous substructures in a network.

8 Conclusions

We have discussed important, practical shortcomings of measures of degree-degree correlation, in particular Newman's assortativity, when applied to large networks. We believe that using Kendall's τ in place of Pearson's correlation might mitigate parts of the problems. More theoretical analysis and experiments are however necessary to understand in detail the sensitivity of these measures to small locally dense graphs. Kendall's τ requires some more computational effort, that is, $O(m \log m)$ (where m is the number of arcs) rather than the $O(m)$ time of Spearman's and Pearson's correlation. However, there are $O(m \log m)$ algorithms based on sorting [Kni66] that are easily parallelized or distributed among multiple computational units, which should help to mitigate the problem.

References

[BCSV04] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. UbiCrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711–726, 2004.

[BMSV19] Paolo Boldi, Andrea Marino, Massimo Santini, and Sebastiano Vigna. BUbiNG: Massive crawling for the masses. *ACM Trans. Web*, 12(2):12:1–12:26, 2019.

[CKL+09] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi, and Prabhakar Raghavan. On compressing social networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 219–228, New York, NY, USA, 2009. ACM.

[Dan43] Henry E. Daniels. The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 33(2):129–135, 1943.

[GK54] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764, 1954.

[KBCL07] Miklos Kurucz, Andras Benczur, Karoly Csalogany, and Laszlo Lukacs. Spectral clustering in telephone call graphs. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 82–91. ACM, 2007.

[Ken45] Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600. ACM, 2010.

[Kni66] William R. Knight. A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436–439, June 1966.

[LvdH13] Nelly Litvak and Remco van der Hofstad. Uncovering disassortativity in large scale-free networks. *Phys. Rev. E*, 87:022801, 2013.

[MMG+07] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[New02] Mark E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, 2002.

[New03] Mark E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[Spe04] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[vdHL15] Pim van der Hoorn and Nelly Litvak. Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet Mathematics*, 11(2):155–179, 2015.

[Vig15] Sebastiano Vigna. A weighted correlation index for rankings with ties. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 24th international conference on World Wide Web*, pages 1166–1176. ACM, 2015.

[VMWG+10] Piet Van Mieghem, Huijuan Wang, Xin Ge, Siyu Tang, and Fernando A Kuipers. Influence of assortativity and degree-preserving rewiring on the spectra of networks. *The European Physical Journal B*, 76(4):643–652, 2010.