A framework for evaluating statistical dependencies and rank correlations in power law graphs

Yana Volkovich
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
y.volkovich@ewi.utwente.nl

Nelly Litvak*
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
n.litvak@ewi.utwente.nl

Bert Zwart
Georgia Tech
765 Ferst Drive, NW Atlanta, Georgia 30332-0205
bertzwart@gatech.edu

ABSTRACT
We analyze dependencies in power law graph data (Web sample, Wikipedia sample and a preferential attachment graph) using statistical inference for multivariate regular variation. To the best of our knowledge, this is the first attempt to apply the well developed theory of regular variation to graph data. The new insights this yields are striking: the three above-mentioned data sets are shown to have a totally different dependence structure between different graph parameters, such as in-degree and PageRank. Based on the proposed methodology, we suggest a new measure for rank correlations. Unlike most known methods, this measure is especially sensitive to rank permutations for top-ranked nodes. Using this method, we demonstrate that the PageRank ranking is not sensitive to moderate changes in the damping factor.

Categories and Subject Descriptors
E.1 [Data structures]: Graphs and networks; G.3 [Probability and Statistics]: Multivariate statistics

General Terms
Algorithms, Experimentation, Measurement

Keywords
Regular variation, PageRank, Web, Wikipedia, Preferential attachment

1. INTRODUCTION
What do we know about the Web structure? There is a vast literature on the subject but we are still far from complete understanding. One point where most researchers agree is the presence of power laws. In simple words, a power law with exponent \( \alpha \) means that a probability of obtaining a value greater than \( x \) is proportional to \( x^{-\alpha} \), where \( \alpha > 0 \) is the power law exponent. The standard example of a power law is a Pareto distribution

\[
P(X > x) = cx^{-\alpha} \quad \text{where } x > x_0.
\]

For excellent surveys on history, properties, modeling, and mining of power laws, and their role in complex networks we refer to e.g. [5, 14, 17, 18, 19].

A natural mathematical formalism for analyzing power laws is provided by the theory of regular variation. This theory has been developed in the context of analysis of extremes [6], financial time series [16], and traffic in communication networks [21]. By definition, the distribution \( F \) has a regularly varying tail with index \( \alpha \), if

\[
P(X > x) = x^{-\alpha} L(x), \quad x > 0,
\]

where \( L(x) \) is a slowly varying function, that is, for \( x > 0 \),

\[
L(tx)/L(t) \rightarrow 1 \quad \text{as } t \rightarrow \infty,
\]

for instance, \( L(x) \) may be equal to a constant or \( \log(x) \). Clearly, a power law can be modeled as an instance of regular variation.

In the present work, we employ statistical inference designed for regular variation, as described in Resnick [22], to analyze the dependencies in power law graphs. To the best of our knowledge, most of the proposed methods have never been applied to massive graph data. We consider in-degrees, out-degrees and PageRank scores in three large data sets: an EU-2005 Web sample, a Wikipedia sample and a Growing Network graph based on the preferential attachment model by Albert and Barabási [2]. The data sets are described in detail in Section 2.

It has become common knowledge that in-degree and PageRank in the Web graph obey power laws [3, 7, 20, 23]. The power law exponents can deviate depending on a data set and an estimator but are believed to satisfy \( \alpha \approx 1.1 \). Similar behavior of in-degree and PageRank has been observed in Wikipedia [4, 23]. There is however no common agreement on the distribution of out-degrees in the Web. Whereas Broder et al. [3] observe a power law with exponent about 1.6, Donato et al. [7] claim that the out-degrees do not follow a power law. Remarkably, the conclusion on whether or not the data follows a power law is often seem to be made purely by determining whether or not the log-log plot resembles the signature straight line. This however can be misleading especially when a size-frequency plot is used [14]. Although one may agree with Li et al. [14] that a cumulative (size-rank) plot is enough to reveal a power law to an experienced eye, for more reliable conclusions on realistic noisy data, we need more than just a glance at the log-log plots. Chakrabarti and Faloutsos [5] mention two goodness-
The analysis of extremal dependence lead us to propose a new rank correlation measure which seems particularly suitable for bivariate power law data. The measure has the appealing property that small values in the data set are of limited influence. Thus, the measure is less sensitive to the choice of the number of considered upper order statistics, as is the case for other statistics used in the analysis of heavy tails. Moreover, unlike most known methods for evaluating rank correlation, our proposed measure is especially sensitive to rank permutations for top-ranked nodes. We discuss our ideas on this matter in Section 5.

Analysis of dependencies in real-life graph and synthetic data contributes towards a better understanding and modeling of complex graph structures. Clearly, for adequate modeling, it is not sufficient to maintain power laws. For instance, it was already argued in [8] that robustness of Internet power law router graph is in strong disagreement with a preferential attachment model. Likewise, our analysis clearly reveals a striking disagreement of the preferential attachment graph with dependence structure of the Web and Wikipedia. Better models have to be sought and existing models have to be thoroughly analyzed before we can conclude that they adequately reflect important features of complex networks.

2. DATA SETS

We chose three data sets that represent different network structures. As the Web sample, we used the EU-2005 data set with 862,664 nodes and 19,235,140 links, that was collected by LAW [1]. We also performed experiments on the Wikipedia (English) data, whose structure is known to be different from the Web graph [4]. This data set contains 4,881,983 nodes and 42,062,836 links. Finally, we simulated a Growing Network by using preferential attachment rule for 90% of new links [2]. The graph consists of 10,000 nodes with constant out-degree $d = 8$. In Figure 1 we show the cumulative log-log plots for in-degrees, out-degrees and PageRank scores in all data sets. The PageRank scores in the

![Figure 1: Cumulative log-log plots for in/(out)-degree, PageRank (c=0.5) and PageRank (c=0.85): (a) EU-2005, (b) Wikipedia, (c) Growing Network](image-url)
Newman [19] suggests to use a bootstrap method for estimating independent random variables having an identical regression line indicating power laws. The log-log plots for in-degree and PageRank in Figure 1 resemble the signature straight line indicating power laws. However, several techniques should be combined in order to establish the presence of heavy tails and to evaluate the power law exponent. Using QQ plots, Hill and altHill plots as well as Pickands plots we will confirm that the in-degree and PageRank (c=0.85) follow power laws with similar exponents for all three data sets. We will also conclude that the out-degree can be modeled reasonably well as a power law with exponent around 2.5-3.

Although all plots in Figure 1 look alike, it does not imply that the three networks have identical structure. One of the goals of the present work is to rigorously examine the dependencies between the network parameters.

3. EVALUATING THE POWER LAWS

Consider non-negative observations $X_1, \ldots, X_n$ and write $X_{(i)}$ for the $i$th largest value of $X_1, \ldots, X_n$, where $1 \leq i \leq n$:

$$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}. \quad (3)$$

In the next sections, we will provide a review of some estimation techniques designed under assumption that $X_1, \ldots, X_n$ are independent random variables having an identical regularly varying distribution with tail index $\alpha$, as defined in (2). The idea is to apply several different procedures and make sure that they lead to the same conclusion.

3.1 Hill plot

The Hill’s estimator $H_{k,n}$ is a widely used estimator of $1/\alpha$, that is based on $k$ upper order statistics:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{X_{(i)}}{X_{(k+1)}} \right).$$

It was proved (see e.g. [22]) that $H_{k,n}$ converges in probability to $1/\alpha$ as $n, k \to \infty$, $k/n \to 0$. An obvious problem with the Hill estimator is choosing the value $k$ so that $X_{(k)}$ corresponds to a ‘beginning’ of the power law tail. This can be mitigated by constructing a so-called Hill plot.

To make a Hill plot for $\alpha$ we graph $\{k, H_{k,n}^{-1}, 1 \leq k \leq n\}$ and if the plot looks stable around a certain horizontal line, we can pick the corresponding value of $\alpha$. This sometimes works beautifully, especially for data close to pure Pareto tails. However, if $L(x)$ in (2) deviates considerably from a constant there may be enormous errors. The Hill plot, as well as the Hill estimator, is also not location invariant. Theoretically, a shift does not affect the power law exponent, however, it drastically distorts the Hill plot. Clearly, in case when the Hill plot does not look stable, the Hill estimator can not be used for the evaluation of $\alpha$.

To construct confidence intervals for the Hill estimator, Newman [19] suggests to use a bootstrap method for estimating the variance of $H_{k,n}^{-1}$. A simpler way is to use the convergence of $\sqrt{k}H_{k,n}$ to a normal random variable with mean $1/\alpha$ and variance $1/\alpha^2$ as $n, k \to \infty$, $k/n \to 0$ [22, p.304]. Thus, one can obtain confidence intervals based on the quantiles of the standard normal distribution.

One can also display the Hill plot in the alternative form $\{\theta, H_{k,n}^{-1}, 0 \leq \theta \leq 1\}$, where $\lceil x \rceil$ is the smallest integer greater or equal to $x \geq 0$. This plot is called the alternative Hill plot, altHill. Compared to the Hill plot, the altHill shows the largest order statistics more prominently. According to [22], if the distribution is not exactly Pareto, then the altHill spends more time in the small neighborhood of $\alpha$ than the Hill plot.

Below we display Hill and altHill plots for EU-2005 (Figure 2), Growing Network (Figure 3) and Wikipedia (Figure 4). The saw-type picture for in-degrees and out-degrees reflects the fact that we deal with integer values that are the same for quite large groups of nodes.

![Figure 2: EU-2005 data set: Hill plot (left) and altHill plot (right) for (a) in-degree, (b) out-degree, (c) PageRank (c=0.5), and (d) PageRank (c=0.85).](image-url)
nice. The plot for in-degree is more stable as it spends significant time around the line $\alpha = 1.1$. The plot for PageRank ($c=0.85$) also behaves well and seems to suggest a slightly smaller tail index, around 1.05. From the plots we see that the estimator for $\alpha$ is very sensitive to the choice of $k$. Thus, constructing a Hill plot is a helpful step when applying a Hill estimator.

The Hill and altHill plots suggest that the in-degree and PageRank in the Web and in the Growing Networks are heavy-tailed but not exactly a Pareto. Indeed, the plots look relatively stable but it is difficult to single out $\alpha$.

For the out-degree in the Web data, the altHill plot oscillate considerably. However, the Hill plot does not behave as nearly as badly as it would, for instance, for the exponential distribution (see example in [22, p.96]). Based on the Hill plot, one may therefore conclude that the out-degree has a power law.

Finally, Wikipedia turns out to be an example of perfect Hill plots whereas altHill shows large oscillations. We conclude that in-degree and PageRank ($c=0.85$) in Wikipedia follow closely a Pareto distribution with index 1.2. The index of PageRank ($c=0.5$) distribution is around 1.4. The out-degree is also Pareto, with index about 1.6.

### 3.2 Pickands plot

A Pickands estimator as presented in [22], is another way to evaluate $\alpha$ and reveal the presence of power laws. We first introduce the extreme-value distributions, defined as

$$G_\gamma = \exp\left(-\frac{1}{1+\gamma x}\right), \quad \gamma \in \mathbb{R}, \quad 1+\gamma x > 0.$$  

The power law case corresponds to $\gamma > 0$ and then $\gamma = 1/\alpha$.

Suppose $\{X_i, i \geq 1\}$ are i.i.d. with common distribution $F$. The Pickands estimator is derived under the condition that the distribution $F$ is in the domain of attraction of the extreme-value distribution $G_\gamma$, that is, there exist $a(n) > 0$, $b(n) \in \mathbb{R}$ such that $nP[X_1 > a(n)x + b(n)] \rightarrow -\log G_\gamma(x)$ as $n \rightarrow \infty$, for $\gamma > 0$, $x \in (-1/\gamma, \infty)$.

The Pickands estimator of $\gamma$ uses differences of quantiles, where the latter are estimated by means of three upper statistics, $X_{(k)}, X_{(2k)}, X_{(4k)}$, from a sample size $n$. The estimator is defined as

$$\gamma_{k,n}^{(\text{Pickands})} = \frac{1}{\log 2} \log \left( \frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}} \right).$$

Determining an appropriate of $k$ is again an important issue. Unlike the Hill estimator, the Pickands estimator is both location and scale invariant.

Similarly to the Hill plot, a Pickands plot consists of the points $\{(k, \gamma_{k,n}^{(\text{Pickands})})\}, \; 1 \leq k < n/4$. A difficulty in constructing Pickands plots for integer-valued observations such as in-degrees and out-degrees in the networks, is that the values of order statistics might be identical, resulting in division by zero. To fix this problem we introduce a randomization of the data by adding uniformly (0,1) distributed random variables to each of the observations.

The Pickands plots for our data sets are presented in Figure 5 below. We note that we plot the values of $\gamma_{k,n}^{(\text{Pickands})}$ that estimates $1/\alpha$. The results for in-degree and PageRank in all three data sets are in good agreement with Hill plots. The new information we find by looking at the plot for out-degree in the Web data. Here a large part of the Pickands plot shows $\gamma < 0$ which signals light tails. This is in agreement with Donato et al. [7] and other papers that claim

![Figure 3: Growing Network data set: Hill plot (left) and altHill plot (right) for (a) in-degree, (b) PageRank ($c=0.5$), and (c) PageRank ($c=0.85$).](image)

![Figure 4: Wikipedia data set: Hill plot (left) and altHill plot (right) for (a) in-degree, (b) out-degree, (c) PageRank ($c=0.5$), and (d) PageRank ($c=0.85$).](image)
that the out-degree data does not follow a power law. On the other hand, the Pickands plot goes below zero only for quite large values of $k$, so we still can not exclude the power law tail.

3.3 QQ plot

Suppose we have a hypothesis that the true distribution function producing the data is $F(x)$. A goodness of fit test provides the rigorous way to verify such hypothesis, whereas the QQ plot is a more informal but convenient alternative. To construct a QQ plot we graph the theoretical quantiles of $F$ versus the sample quantiles:

$$\left\{ \left( F^{-1}\left( \frac{i}{n+1} \right), X_{(n-i+1)}\right) : 1 \leq i \leq n \right\},$$

where $F^{-1}(y) = \inf \{ x : F(x) \geq y \}$ is the inverse of distribution function $F$. If our hypothesis is true then the result should fall roughly on the straight line $\{ (x, x), x > 0 \}$. One potential problem is how to decide what we consider ‘close enough’ to linear.

To apply QQ plots to power laws, suppose that our null hypothesis is that for some $x_0 > 0$, distribution of random variable $X$ satisfies

$$P(X > x) = \left( \frac{x}{x_0} \right)^{-\alpha},$$

so it follows that $P(\log X > y) = e^{-\alpha(y - \log x_0)}$. Hence, using quantiles of exponential distribution we plot

$$\left\{ \left( -\log \left( 1 - \frac{i}{n+1} \right), \log X_{(n-i+1)} \right) : 1 \leq i \leq n \right\}.$$

The slope of the least-squared line fitted to the QQ plot is an estimate of $1/\alpha$. Thus, if $\{(x_i, y_i), 1 \leq i \leq n\}$ are $n$ points on the plane, we can calculate the slope in standard way

$$SL\{(x_i, y_i), 1 \leq i \leq n\} = S_{xy}/S_{xx},$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\bar{x}$ means mean value of $x$. Now we can define the QQ estimator for $1/\alpha$ based on $k$ upper order statistics as

$$SL\left\{ \left( -\log \left( 1 - \frac{i}{n+1} \right), \log X_{(n-i+1)} \right) : n - k + 1 \leq i \leq n \right\}.$$ Clearly, there remains the problem of choosing $k$.

The QQ plots for our data are presented in Figures 6 and 7 for two choices of $k$. Again, the data on in-degree and PageRank resulted in QQ plots similar to straight line, and the estimates for $\alpha$ are close to what we expected. Thus, in these case all techniques point to the same result.

With a certain amount of tolerance, we can accept that the QQ plot for out-degrees in the Web data in Figure 7(b)(left) is close enough to a straight line. Moreover, the estimated $\alpha = 2.95$ is in good agreement with the Hill plot. We also note that $\alpha > 2$ implies a finite variance while power law models are especially important in case when the variance is infinite, reflecting high variability [14, 21]. Hence, in case of a finite variance, it is not really crucial whether the data obeys a power law. To exclude the possibility of exponential tail of out-degree, we also constructed a QQ plot with exponential quantiles by plotting $-\log (1 - i/(n+1))$ against $X_{(n-i+1)}$. The result that we do not present here is not any close to a straight line. To summarize, the out-degree has a finite variance and a tail heavier than exponential, so it can
be modeled reasonably well as a power law with exponent around 2.5-3, according to our estimates.

4. **EXTREMAL DEPENDENCIES**

The goal of this section is to measure the dependencies between in-degree and PageRank (c=0.5 and 0.85), in-degree and out-degree, and out-degree and PageRank (c=0.85) in our data sets. In Sections 4.1 we explain the methodology and perform preliminary computations. The results on dependence structure in our three data sets are presented in Section 4.2.

4.1 **Angular Measure**

Suppose we are interested in analyzing the dependencies between two regular varying characteristics of a node, X and Y. Let \( X_j \) and \( Y_j \) be observations of X and Y for the corresponding node \( j \). Following [22], we start by using the rank transformation of \((X, Y)\), leading to \( \{(r^x_j, r^y_j), 1 \leq j \leq n\} \), where \( r^x_j \) is the descending rank of \( X_j \) in \((X_1, \ldots, X_n)\) and \( r^y_j \) is the descending rank of \( Y_j \) in \((Y_1, \ldots, Y_n)\). Next we choose \( k = 1, \ldots, n \) and apply the polar coordinate transform as follows

\[
\text{POLAR} \left( \frac{k}{r^x_j}, \frac{k}{r^y_j} \right) = (R_{j,k}, \Theta_{j,k}).
\]

Now we need to consider the points \( \{(\Theta_{j,k}, R_{j,k}), 1 \leq j \leq n\} \) and make a plot for cumulative distribution function of \( \Theta \). In other words, we are interested in the angular measure, i.e. in the empirical distribution of \( \Theta \) for \( k \) largest values of \( R \). Thus, unlike the correlation coefficient, the angular measure provides a subtle characterization of the dependencies in the tails of X and Y, or, extremal dependencies. If such measure is concentrated around \( \pi/4 \) then we observe a tendency toward complete dependence, when a large value of X appears simultaneously with a large value of Y. In the opposite case, when such large values almost never appear together, we have either large value of X or large value of Y, hence, \( \Theta \) should be around 0 or \( \pi/2 \). The middle case plots can be seen as a tendency to dependency or independency.

It was proved in [22] that the empirical measure converges to a proper distribution on \([0, \pi/2]\) as \( n, k \to \infty, k/n \to 0 \). That is, ideally, we need to consider only a relatively small part of a large data set.

In practice the problem remains: how to choose a suitable value of \( k \)? In the case of bi-variate data, this can be determined by making a Starića plot. We consider radii \( R_1, \ldots, R_{n,k} \) from (4) and rank them in descending order \( R_{(1)} \geq \ldots \geq R_{(n)} \) as before. To get Starića plot we graph

\[
\left\{ \left\{ \frac{R_{1,j}}{R_{(1)}}, \frac{R_{2,j}}{R_{(2)}} \right\}, 1 \leq j \leq n \right\},
\]

or

\[
\left\{ \left\{ \frac{R_{1,j}}{\sum_{i=1}^{n} 1(R_{i,k} \geq 1)} \right\}, 1 \leq j \leq n \right\}.
\]

The idea is that for suitable \( k \) the ratio in the ordinate should be roughly a constant and equal 1 for the values of the abscissa in the neighborhood of 1. The plot looks different for the different parameters \( k \) and one can either find a suitable \( k \) by trial and error or use numerical algorithms to compute optimal \( k \). A Starića plot for good \( k \) will have a region in the right neighborhood of \( x = 1 \) where the plot is hugging the \( y = 1 \) line. If the line is going steep up at \( x = 1 \) then the chosen \( k \) is too large. On the other hand, if the graph stabilizes around \( y = 1 \) for some \( x < 1 \) then it means that \( k \) is too small, and we miss some valuable tail data. We refer to Resnick [22] for more details and references.

After some experiments, we chose appropriate values of \( k \) for the four pairs (in-degree, PageRank (c=0.85)), (in-degree, PageRank (c=0.5)), (in-degree, out-degree), and (out-degree, PageRank (c=0.85)) in our data sets. The corresponding Starića plots are presented in Figure 8(a,b) and Figure 9(a-d). The good news is that the plots for in-degree/PageRank behaves nicely in all three data sets, which makes our angular measure more reliable. The Growing Network exhibits an ideal Starića plot (Figure 8). A surprisingly bad behavior is on the plot for in-degree/out-degree in Wikipedia (Figure 9(d)(right)), where the Starića curve wonders well off the \( y = 1 \) line.

4.2 **Dependence measurements on the data**

After defining a suitable \( k \), we compute the pairwise angular measure. In Figure 10 we depict \( \theta \in [0, \pi/2] \) against
the fraction of observations where the angle $\Theta$ is greater or equal to $\theta$.

The results are striking. Let us look first at Figure 10(a,b) which characterizes the dependence between in-degree and PageRank. For the Wikipedia data set we observe that about half of observations are concentrated around 0 whereas another half is close to $\pi/2$. This suggests an independence of the tails of in-degree and PageRank ($c=0.85$ and $c=0.5$). That is, in Wikipedia data set an extremely high in-degree almost never implies an extremely high ranking. The picture is completely the opposite for Growing Networks, where the angular measure is entirely concentrated around $\pi/4$ indicating a complete dependence. Thus, in highly centralized preferential attachment graphs, most connected nodes are also most highly ranked.

Finally, the Web graph exhibits a subtle dependence structure that results in angular measure which is almost uniform on $[0, \pi/2]$. This suggests that PageRank popularity measure can not be replaced by in-degree without significant disturbance in the ranking (of course, in-degree can not be used as a popularity measure for many other reasons, for instance, because it is easy to spam by creating link farms; we refer to [13] for further discussion of PageRank and other popularity measures).

The picture is different in Figure 12(c) where we depict the angular measure for in-degree and out-degree in the Web and in Wikipedia. In the Web, the in- and out-degree tend to be independent which justifies the distinction between hubs and authorities [11]. In Wikipedia the in- and out-degrees are dependent but this dependence is not absolute.

Finally, the dependence between out-degree and PageRank in the Web and Wikipedia in Figure 12(d) resembles the patterns observed for in-degree and PageRank.

5. RANK CORRELATION

In this section, we introduce a new method for measuring correlations between ranking orders in power law graphs. The proposed correlation measure is based on the extremal dependencies technique, presented in Section 4.

5.1 The $\Theta$ rank correlation measure

We start by noting that the angular measure described in Section 4.1 is in fact based on a rank transformation. This is clearly seen from formula (4) where only rank of the parameters $X$ and $Y$ plays a role. This observation naturally leads to a new measure for rank correlations.

In summary, our idea is as follows. As before, we define $r^1_i$ and $r^2_i$ as a ranking order of page $i$ in scheme 1 and 2, respectively, where $i = 1 \ldots n$. Now we suggest to represent the difference between the two rank positions of $i$ by the angle

$$\Theta_i = \arctan(r^1_i/r^2_i).$$

For example, in Figure 11, $\Theta_i$ is depicted for a node that has rank 3 in scheme 1 and rank 6 in scheme 2. Note that this is exactly the angle in $(0, \pi/2)$ computed in (4) in order to construct the angular measure. The value $\Theta$ close to $\pi/4$ means a relatively small change in ranking. On the other hand, $\Theta$ around $\pi/2$ means that the node $i$ is much better off with scheme 2, and the value close to 0 says that the node is ranked much higher by scheme 1. Thus, we actually...
measure the rank difference for node $i$ in radians! Having computed $\Theta_i$ for every $i$ (or for a certain set of highly ranked nodes $i$) we construct a corresponding empirical cumulative distribution function for $\Theta$. As in the previous section, the resulting angular measure can be used to characterize the rank correlations.

In order to illustrate the proposed methodology, consider the scatter plot of ranking order 1 against ranking order 2 (see Figure 11). When two ranks are the same (like the node ranked 1 in the example) then the corresponding point lies on the diagonal. On the other hand, if there is a considerable disturbance in ranking (for instance, in the example, the rank 2 and 9 are swapped) then we immediately see considerable deviation from the diagonal.

Compared to the common rank correlation measures such as Kendall’s $\tau$ and Spearman’s $\rho$, our proposed measure has an important advantage that it is able to reveal the slightest rank disturbance among highly ranked nodes while neglecting even moderate perturbations among lowly ranked nodes. Indeed, if we swap the rank 1 and 10, we get $\Theta = \arctan(1/10) \approx 0.1$, which is close to the $x$-axis, and is a visible deviation from $\pi/4$. On the other hand, swapping the numbers 1000 and 1010 yields $\Theta = \arctan(1000/1010) \approx \pi/4$. In other words, the $\Theta$ rank correlation measure actually evaluates the rank disturbance visible for users. Certainly, the $\arctan(\cdot)$ function makes our measure symmetric with respect to the schemes 1 and 2.

![Figure 11: Rank Correlation.](image)

Naturally, in this framework, it is also possible to compute such angular measure only for the top ranked pages. This can be done along the same lines as in Section 4.1 as follows.

Based on the polar transformation (4) we can separate top ranked pages by considering only points $(\Theta_i,k : R_{ilk} > 1)$. Here the question of choosing $k$ does not arise anymore. Indeed, the technique involving Starica plot was needed to get an idea where the power law behavior starts in order to measure statistical dependency for the heavy-tailed data as in [22]. On the other hand, if we are interested in rank correlations, we may simply pick the $k$ that gives us the top proportion of pages we are interested in. Note that by increasing $k$ we do not change the observed values of $\Theta$, we merely increase their number. As a result, in the angular measure, each observation will simply have less weight. On contrary, decreasing $k$ means ‘zooming in’ the rank perturbations on the top.

One more advantage of the proposed correlation measure is the fast and easy implementation since for each node $i$, only the fraction $r_i^k / r_i^2$ has to be computed.

Below we present the example of the proposed rank correlation measure in Growing Networks, Web and Wikipedia. We rank the three data sets by using the definition of PageRank (2), where the damping factor is equal to $c = 0.5$ and $c = 0.85$. In Figure 12 we plot cumulative functions for angular measures for $k = 100$ and the values of $k$’s that have been chosen according to the Starica plots (see Figure 8(c) and Figure 9(e)). For Growing Network data set we observe the strong correlation between ranking schemes. We can also conclude that in Wikipedia the change in the damping factor affects only about 20% of considered pages, in the top-hundred group as well as in the larger group. For the Web data, the correlation between ranking is not significant for approximately half of the pages. However, for the top pages, the difference in the damping factor mixes up the order of ranking. The results for the top 100 pages are in lines but more informative than the corresponding values of Kendall’s $\tau$: $\tau_{GN} = 0.9967$, $\tau_{WI} = 0.6879$, $\tau_{EU} = 0.4092$.

### 5.2 Discussion

The main idea of the $\Theta$ rank correlation measure is that we characterize the rank correlations by a cumulative distribution of $\Theta_i$’s, where $i = 1, \ldots, n$. This way, one can actually see how many pages change their ranks significantly. Such measure is substantially more informative than just one number, that represents the correlation in the whole graph. For instance, Melucci [15] noticed that Kendall’s $\tau$ tends to grow close to one for large data sets. The author provides an example where Kendall’s $\tau$ for ranking orders of only a few hundred Web pages becomes almost 1, in spite of the large number of rank perturbations. We remark however that if for some reason having one number is necessary, one can always compute, e.g. the expected deviation of $\Theta$ from $\pi/4$.

As mentioned before, the proposed correlation measure is quite harsh with respect to lowly ranked nodes. Indeed, the node ranked 1000 must fall all the way to 2000 to make the same effect as number 1 becoming number 2. We would like to emphasize that such discrepancy is especially suitable for ranking order emerging from a heavy-tailed data, such as PageRank or in-degree. This is because in such data, there is a huge difference between the highest values of the realizations, cf. [9].

### Acknowledgments

The authors would like to thank Debora Donato for her great help with Web and Wikipedia data sets. This article is the
result of joint research in the 3TU Centre of Competence NIRICT (Netherlands Institute for Research on ICT) within the Federation of Three Universities of Technology in The Netherlands.

6. REFERENCES