A Weighted Correlation Index for Rankings with Ties

Sebastiano Vigna
Dipartimento di Informatica
Università degli Studi di Milano, Italy

ABSTRACT
Understanding the correlation between two different scores for the same set of items is a common problem in graph analysis and information retrieval. The most commonly used statistics that quantifies this correlation is Kendall’s $\tau$; however, the standard definition fails to capture that discordances between items with high rank are more important than those between items with low rank. Recently, a new measure of correlation based on average precision has been proposed to solve this problem, but like many alternative proposals in the literature it assumes that there are no ties in the scores. This is a major deficiency in a number of contexts, and in particular when comparing centrality scores on large graphs, as the obvious baseline, indegree, has a very large number of ties in social networks and web graphs. We propose to extend Kendall’s definition in a natural way to take into account weights in the presence of ties. We prove a number of interesting mathematical properties of our generalization and describe an $O(n \log n)$ algorithm for its computation. We also validate the usefulness of our weighted measure of correlation using experimental data on social networks and web graphs.

1. INTRODUCTION
In the analysis of social networks and web graphs, one is often faced with different scores assigned by different centrality measures (e.g., indegree and Bavelas’s closeness [1]) to the same graph. Analogously, in information retrieval one has the lists of documents returned by different search engines and their associated relevance scores, or the lists of query recommendation returned by different algorithms.

In most of the literature, the scores are assumed to be without ties, thus inducing a ranking of the elements. At that point, correlation statistics such as Spearman’s rank correlation coefficient [25] and Kendall’s $\tau$ [12] can be used to evaluate the similarity of the rankings. Spearman’s correlation coefficient is equivalent to the traditional linear correlation coefficient computed on ranks of items. Kendall’s $\tau$, instead, is proportional to the number of pairwise adjacent swaps needed to convert one ranking into the other.

For a number of reasons, Kendall’s $\tau$ has become a standard statistic to compare the correlation between two ranked lists. Such reasons include fast computation ($O(n \log n)$, where $n$ is the length of the list, using Knight’s algorithm [14]), and the existence of a variant that takes care of ties [13].

The explicit treatment of ties is of great importance when comparing scores assigned by different graph centrality measures. The baseline of such scores is indegree. More sophisticated approaches include Bavelas's closeness [1], Katz's index [10], PageRank [22], and countless variants. Due to the highly skewed indegree distribution, a very large number of documents share the same indegree, and the same happens of many other scores: it is thus of utmost importance that the evaluation of correlation takes into account ties as first-class citizens. Breaking ties randomly, for instance, leads to the paradoxical results that the score made by half zeroes, half ones has with any other equivalent score an expected $\tau$ of $n/(2(n - 1)) \approx 0.5$ instead of one.

On the other hand, Kendall’s $\tau$ has some known problems that motivated the introduction of several weighted variants. In particular, a striking difference often emerges between the anecdotal evidence of the top elements by different scores being almost identical, and the $\tau$ value being quite low. This is due to a known phenomenon: the scores of important items tend to be highly correlated in all reasonable rankings, whereas most of the remaining items are ranked in slightly different ways, introducing a large amount of noise, yielding a low $\tau$ value.

This problem motivates the definition of correlation statistics that consider more important correlation between highly ranked items. In particular, recently Yilmaz, Aslam and Robertson introduced a statistics, named AP (average precision) correlation [27], which aims at considering more important swaps between highly ranked items. The need for such a measure is very well motivated in the introduction of their paper, and we will not repeat here their detailed discussion.

In this paper, we aim at providing a measure of correlation in the same spirit of the definition of Yilmaz, Aslam and Robertson, but taking smoothly ties into account. We will actually define a general notion of weighting for Kendall’s $\tau$, and develop its mathematical properties. Since it is important that such a statistics is computable on very large data sets, we will provide a generalization of Knight’s algorithm that can be applied whenever the weighting depends additively or multiplicatively on a weight assigned to each
2. RELATED WORK

Shieh [24] wrote the one of the first papers proposing a generic weighting of Kendall’s $\tau$. She assumes from the very start that there are no ties, and assign to the exchange between $i$ and $j$ a weight $w_{ij}$. Her motivation is the fidelity evaluation of software packages for structural engineering, in which a set of variables is ranked in two different ways, and one would like to emphasize agreement on the most important ones. In particular, she concentrates on weights given by the product of two weights associated with the elements participating in the exchange. Our work can be seen as a generalization of her approach, albeit we combine weights differently.

Kumar and Vassilvitskii [16] study a definition that extends Shieh’s taking into account position weights and similarity between elements. Again, they assume that ties are broken arbitrarily, which is an unacceptable assumption if large sets of elements have the same score. Fagin, Kumar and Sivakumar [5] use instead penalty weights to apply Kendall’s $\tau$ just to the top $k$ elements of two ranked lists (with no ties). Exchanges partially or completely outside the top $k$ elements obtain different weights.

Finally, the recent quoted work by Yilmaz, Aslam and Robertson [27] on AP correlation is the closest to ours in motivation and methodology, albeit targeted at ranked lists with no ties. Melucci [2] extends their definition to include a probabilistic weight assigned to items, remarking that both approaches are actually specific instances of Shieh’s.

We note that analogous research exists in association with Spearman’s correlation: Iman and Conover [9], for example, study the usage of Savage scores [23] instead of ranks when comparing ranked lists. Savage scores for a ranked list of $n$ elements are given by $\sum_{i=1}^{n} \frac{1}{j}$, where $i$ is the rank (starting at one) of an element. Spearman’s correlation applied to Savage scores considers more important elements at the top of a ranked list.

A fascinating proposal, entirely orthogonal to the ones we discussed, is the idea of weighting Kemeny’s distance between permutations proposed by Farnoud and Milenkovic [6]. In this proposal, Kemeny’s distance between two permutations $\pi$ and $\sigma$ is characterized as the minimum number of adjacent transpositions (i.e., transpositions of the form $(i \ i+1)$) that turn $\pi$ into $\sigma$. At this point, one can define a weight associated to each adjacent transposition, and by assigning larger weights to adjacent transpositions with smaller indices one can make differences in the top part of the permutations more important than differences in the bottom part. The right notion of weighted distance turns out to be the minimum sum of weights of a sequence of adjacent transposition that turn $\pi$ into $\sigma$. The interesting property of this approach is that avoids the need for a ground truth (an intrinsic notion of importance of an element), which is necessary, implicitly or explicitly, to weigh an exchange in the approaches of [24, 27] and in the one discussed in this paper.

The main drawbacks, presently, are that the weight assignment is not very intuitive (as it is related to positions, rather than to elements) and that more work is needed to extend the distance into a proper correlation index in the case of ties.

3. MOTIVATION

The need for weighted correlation measures in the case of ranked list has been articulated in detail in previous work. Here we will fo-

2http://law.di.unimi.it/
Table 3: Kendall’s τ between Wikipedia centrality measures.

<table>
<thead>
<tr>
<th>Indegree</th>
<th>PR</th>
<th>Katz</th>
<th>Harm.</th>
<th>Cl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indegree</td>
<td>1</td>
<td>0.75</td>
<td>0.90</td>
<td>0.62</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.75</td>
<td>1</td>
<td>0.75</td>
<td>0.61</td>
</tr>
<tr>
<td>Katz</td>
<td>0.90</td>
<td>1</td>
<td>0.75</td>
<td>0.70</td>
</tr>
<tr>
<td>Harmonic</td>
<td>0.62</td>
<td>0.61</td>
<td>0.70</td>
<td>1</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.55</td>
<td>0.56</td>
<td>0.62</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 4: Kendall’s τ between Wikipedia centrality measures, restricted to nodes of indegree 1 and 2.

<table>
<thead>
<tr>
<th>Indegree</th>
<th>PR</th>
<th>Katz</th>
<th>Harm.</th>
<th>Cl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indegree</td>
<td>0.31</td>
<td>0.31</td>
<td>0.63</td>
<td>0.24</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.31</td>
<td>1</td>
<td>0.27</td>
<td>0.10</td>
</tr>
<tr>
<td>Katz</td>
<td>0.63</td>
<td>0.27</td>
<td>1</td>
<td>0.50</td>
</tr>
<tr>
<td>Harmonic</td>
<td>0.24</td>
<td>0.10</td>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.06</td>
<td>0.10</td>
<td>0.20</td>
<td>0.65</td>
</tr>
</tbody>
</table>

with the exception of the correlation between indegree and Katz, all other correlation value fail to surpass the 0.9 threshold, usually considered for the threshold for considering two rankings equivalent [26]. Actually, they are below the threshold 0.8, under which we are supposed to see considerable changes. The correlation of closeness with harmonic centrality, moreover, is even more pathological: it is the largest correlation.

An obvious observation is that, maybe, the score is lowered by a large discordance in the rest of the rankings. Table 2 tries to verify this intuition by listing the top pages that are associated with the Wordnet category “scientist” in the Yago2 ontology data [8]. These pages have considerably lower score (their rank is below 300), yet the first three rankings are almost identical. Harmonic centrality is still slightly different (Linnaeus is absent, and actually ranks 21), which tells us that the Kendall’s τ is not giving completely unreasonable data. Nonetheless, closeness continues to provide apparently random results.

We have actually to delve deep into Wikipedia, beyond rank 100000 using the category “cocktail” to see that, finally, things settle down (Table 5). While closeness still displays a few quirks, the rankings start to stabilize.

To understand what happens in the very low-rank region, in Table 4 we provide Kendall’s τ as in Table 3, but restricting the computation to nodes of indegree 1 and 2. As it is immediately evident, after stabilization the low-rank region is fraught with noise and all correlation values drop significantly.

The very high correlation between closeness and harmonic centrality is, actually, not strange: on the nodes reachable from giant connected component of our Wikipedia snapshot (89% of the nodes) they agree almost exactly, as closeness is the reciprocal of a normalized arithmetic mean, whereas harmonic centrality is the reciprocal of a normalized harmonic mean [2]. Even if the remaining 11% of the nodes is completely out of place, making closeness useless, Kendall’s τ tells us that it should be interchangeable with harmonic centrality. At the same time, Kendall’s τ tells us that indegree is very different from PageRank, which again goes completely against our empirical evidence (and also against theoretical considerations [7]).

In the rest of the paper, we will try to approach in a systematic manner these problems by defining a new weighted correlation index for scores with ties.

4. A WEIGHTED CORRELATION INDEX

In his 1945 paper about ranking with ties [13], Kendall, starting from an observation of Daniels [4], reformulates his correlation index using a definition similar in spirit to that of an inner product, which will be the starting point of our proposal: we consider two real-valued vectors \( r \) and \( s \) (to be thought as scores) with indices in \( n \); then, let us define

\[
\langle r, s \rangle := \sum_{i < j} \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j),
\]

where

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

Indices of score vectors in summations belong to \( n \) throughout the paper. Note that the expression above is actually an inner product in a larger space of dimension \( n(n - 1)/2 \): each score vector \( r \) is mapped to the vector with coordinate \((i, j), i < j \), given by \( \text{sgn}(r_i - r_j) \).

We have the property

\[
\langle r, r \rangle = \langle s, s \rangle = \text{sgn}(s) \langle r, s \rangle,
\]

which reminds of the analogous property for inner products, and that \( \langle r, - \rangle = \langle s, r \rangle = 0 \) if \( r \) is constant. Following the analogy, we can define

\[
\| r \| = \sqrt{\langle r, r \rangle},
\]

so

\[
\| r \| \geq \| s \|.
\]

This property makes it possible to define Kendall’s τ between two vectors \( r \) and \( s \) with nonnull norm as a normalized inner product, in a way formally identical to cosine similarity:

\[
\tau(r, s) := \frac{\langle r, s \rangle}{\| r \| \cdot \| s \|}.
\]

We recall that if \( r \) and \( s \) have no ties, the definition reduces to the classical “normalized difference of concordances and discordances”, as the denominator is exactly \( n(n - 1)/2 \). The definition above is exactly that proposed by Kendall [13], albeit we use a different formalism.

The form of (1) suggests that to obtain a weighted correlation index it would be natural to define a weighted inner product

\[
\langle r, s \rangle_w := \sum_{i < j} \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) w(i, j),
\]

where \( w(-, -) : n \times n \rightarrow \mathbb{R}_{\geq 0} \) is some nonnegative symmetric weight function. We would have then a new norm

\[
\| r \|_w = \sqrt{\langle r, r \rangle_w}.
\]

Throughout the paper, we use von Neumann’s notation: if \( n \) is a natural number, \( n = \{0, 1, \ldots, n - 1\} \).

We thank Ravi Kumar for this useful observation. Note that the map from the original space to the larger space is not linear.
and a new correlation index

\[
\tau_w(r, s) := \frac{\langle r, s \rangle_w}{\|r\|_w \cdot \|s\|_w}.
\]

Note that still \(\langle r, - \rangle_w = \langle -, r \rangle_w = 0\) if \(r\) is constant.

### 4.1 Properties

We say that two score vectors \(r\) and \(s\) are equivalent if \(\text{sgn}(r_i - r_j) = \text{sgn}(s_i - s_j)\), opposite if \(\text{sgn}(r_i - r_j) = -\text{sgn}(s_i - s_j)\) for all \(i\) and \(j\). Since \((r, s)\) is a semi-definite inner product on a larger space (due the possibility of zero weights), the Cauchy-Schwarz inequality still holds, even if we need positive definiteness for the necessary condition on equality:

**Theorem 1** \(\|r, s\|_w \leq \|r\|_w \cdot \|s\|_w\). A sufficient condition for equality to hold is that the two vectors are equivalent or opposite. The condition is necessary if \(w\) is strictly positive and \(\|r\|_w \cdot \|s\|_w \neq 0\) or \(\|r\|_w \cdot \|s\|_w = 0\).

Note that the two last conditions are necessary: when \(w\) is the constant zero weight we have equality for all vectors, and if one of the vector has null norm while the other has not the necessary linearity condition for equality is moot.

Interestingly, even if our “inner product” is neither additive nor linear, we can still prove directly the triangular inequality for the induced “norm”:

**Theorem 2** \(\|r + s\|_w \leq \|r\|_w + \|s\|_w\).

**Proof.**

\[
\|r + s\|_w = \langle r + s, r + s \rangle_w = \sum_{i < j} \text{sgn}(r_i + s_i - r_j - s_j)^2 w(i, j)
\]

\[
\leq \sum_{i < j} (|\text{sgn}(r_i - r_j)| + |\text{sgn}(s_i - s_j)|)^2 w(i, j)
\]

\[
= \langle r, r \rangle_w + \langle s, s \rangle_w + \sum_{i < j} |\text{sgn}(r_i - r_j) \cdot \text{sgn}(s_i - s_j)| w(i, j).
\]
We now notice that
\[
\sum_{i<j} \left| \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) w(i, j) \right| \leq \sum_{i<j} \text{sgn}(r_i - r_j)^2 w(i, j) = \|r\|_w^2,
\]
and analogously for \(\|s\|_w\). We conclude that \(\|r + s\|_w^2 \leq \|r\|_w^2 + \|s\|_w^2 + 2\|r\|_w\|s\|_w = (\|r\|_w + \|s\|_w)^2\).

The triangular inequality has a nice combinatorial interpretation: adding score vectors can only decrease the amount of “untieness”. There is no way to induce in a sum vector more untieness than the amount present in the summands.

Finally, we gather systematically the properties of \(\tau_w\):

**Theorem 3** Let \(w : n \times n \rightarrow \mathbb{R}\) be a nonnegative symmetric weight function. The following properties hold for every score vector \(t\) and for every \(r, s\) with nonnull norm:

- if \(t\) is constant, \(\|t\|_w = 0\);
- \(-1 \leq \tau_w(r, s) \leq 1\);
- if \(r\) and \(s\) are equivalent, \(\tau_w(r, s) = 1\);
- if \(r\) and \(s\) are opposite, \(\tau_w(r, s) = -1\).

Moreover, if \(w\) is strictly positive:

- if \(\|t\|_w = 0\), \(t\) is constant;
- if \(\tau_w(r, s) = 1\), \(r\) and \(s\) are equivalent;
- if \(\tau_w(r, s) = -1\), \(r\) and \(s\) are opposite.

As a result, if \(w\) is strictly positive and we obtain correlation \(\pm 1\) the equivalence classes formed by tied scores are necessarily in a size-preserving bijection that is monotone (increasing or decreasing, depending on the sign) on the scores.

### 4.2 Decoupling rank and weight

The reader has probably already noticed that the dependence on the weight on the indices associated to the elements has no meaning: a trivial request (see, for instance [11]) on a correlation measure is that, like Kendall’s \(\tau\), it is invariant by isomorphism, that is, it does not change if we permute the indices of the vector. This currently doesn’t happen because we are using the numbering of the element as ground truth to weigh the correlation between \(r\) and \(s\). While there is nothing bad in principle (we can stipulate that elements are indexed in order of importance using some external source of information), we think that a more flexible approach decouples the problem of the ground truth from the problem of weighting. We thus define the ranked-weight product

\[
(r, s)_{\rho,w} := \sum_{i<j} \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) w(\rho(i), \rho(j)),
\]

where \(\rho : n \rightarrow n\{\infty\}\) is a ranking function associating with each index a rank, the highest rank being zero. We admit the possibility of rank \(\infty\), given that the weight function provides a meaningful value in such a case, to include also the case of partial ground truths. The definition of the ranked-weighted product induces, as in (1), a correlation index \(\tau_{\rho,w}\), and the machinery we developed applies immediately, as \(w(\rho(\cdot), \rho(\cdot))\) is just a different weight function.

What if there is no ground truth to rely on? Our best bet is to use the rankings induced by the vectors \(r\) and \(s\). Let us denote by \(\rho_{r,s}\) the ranking defined by ordering elements lexicographically with respect to \(r\) and then \(s\) in case of a tie (in descending order), and analogously for \(\rho_{s,r}\). If two elements are at a tie in both vectors, their can be placed in any order, as their rank does not influence the value of \(\tau_{\rho,w}\). We just take the average of the two possible correlations and define

\[
\tau_{\rho,w}(r, s) := \frac{\tau_{\rho_{r,s},w}(r, s) + \tau_{\rho_{s,r},w}(r, s)}{2}. \tag{2}
\]

The same approach has been used in [27] to make AP correlation symmetric.

This is the definition used in the rest of the paper. It should be noted that the theory developed in Section 4.1 about \(\|\cdot\|_w\) has no counterpart for \(\tau_{\rho,w}\), but the relevant parts of Theorem 3 can be trivially extended, and those are the basic properties that we want to hold.

### 4.3 Choosing a weighting scheme

There are of course many ways to choose \(w\). For computational reasons, we will see that it is a good idea to restrict to a class of weighting schemes in which \(w\) is obtained by combining additively or multiplicatively a one-argument weighting function \(f : n \rightarrow \mathbb{R}_{\geq 0}\) applied to each element of a pair.

Shieh [24], for instance, combines weights multiplicatively, without giving a motivation. We have, however, two important motivations for adding weights. First and foremost, unless weights are scaled in some way that depends on \(n\) (which we would like to avoid), the largest weight will be some constant, and then weight...
will decrease monotonically with importance. As a result, an exchange between the first and the last element would be assigned an extremely low weight. Second, adding weights paves the way to a natural measure for top k correlation [5] by assigning rank \( \sim \) to elements after the first k. The definition of such a measure in the multiplicative case is quite contrived and ends up being case-by-case.

For what matters \( f \), we are particularly interested in the hyperbolic weight function \( f(r) := 1/(r + 1) \). This function gives more importance to elements of high rank, and weights zero only pairs in which both index have infinite rank. Using a hyperbolic weight has a number of useful features. First, it reminds the well-motivated weight given to exchanges by AP correlation. Second, it guarantees that as \( n \) grows the mass of weight grows indefinitely. Using a function with quadratic decay, for instance, might end up in making the influence of low-rank element vanish too quickly, as it is summable. For the opposite reason, a logarithmic decay might fail to be enough discriminative to provide additional information with respect to the standard \( r \).

We try to make this intuition more concrete in Figure 1, where we display a number of scatter plots showing the correlation between Kendall’s \( \tau \) and the additive weighted \( \tau \) defined by (2) under different weighting schemes. The left half of the plots correlates all permutations on 12 elements with the identity permutation. The right half correlates all score vectors made of 15 values with skewed distribution (there are \( t + 1 \) elements with score 0 \( \leq t \leq 4 \) with the same vector in descending order. A visual examination of the plots suggests, indeed, that logarithmic weighting restricts too much the possible divergence from Kendall’s \( \tau \), whereas quadratic weighting ends up in providing answers that are too uncorrelated.

We will return to these considerations in Section 6.

## 5. Computing \( \tau_{p,w} \)

Our motivations come from the study of web and social graphs. It is thus essential that our new correlation measure can be evaluated efficiently. We now describe a generalization of Knight’s algorithm [14] that makes it is possible to compute \( \tau_{p,w} \) in time \( O(n \log n) \) under some assumptions on \( w \). Our first observation is that, similarly to the unweighted case, each pair of indices \( i, j \) with \( i < j \) belongs to one of five subsets; it can be

- a joint tie, if \( r_i = r_j \) and \( s_i = s_j \);
- a left tie, if \( r_i = r_j \) and \( s_i \neq s_j \);
- a right tie, if \( r_i \neq r_j \) and \( s_i = s_j \);
- a concordance, if \( \text{sgn}(r_i - r_j) \cdot \text{sgn}(s_i - s_j) = 1 \);
- a discordance, if \( \text{sgn}(r_i - r_j) \cdot \text{sgn}(s_i - s_j) = -1 \).

Let \( J, L, R, C \) and \( D \) be the overall weight of joint ties, left ties, right ties, concordances and discordances, respectively. Clearly,

\[
J + L + R + C + D = \sum_{i < j} w(r(i), r(j)) = T.
\]

The first requirement for our technique to work is that \( T \) can be computed easily. This is possible if weights are computed additively or multiplicatively from some single-argument function \( f \). In the additive case,

\[
T = \sum_{i < j} \left( f(r(i)) + f(r(j)) \right) = (n - 1) \sum_i f(r(i)). \quad (3)
\]

Also the multiplicative case is easy, as

\[
2T = 2 \sum_{i < j} f(r(i))f(r(j)) = \left( \sum_i f(r(i)) \right)^2 - \sum_i f(r(i))^2. \quad (4)
\]

The same observation leads to a simple \( O(n \log n) \) algorithm to compute \( L \): sort the indices in \( n \) by \( r \), and for each block of consecutive \( k \geq 1 \) elements with the same score apply (3) or (4) restricting the indices to the subset. In the same way one can compute \( R \) and \( J \).

We now observe that, as in the unweighted case,

\[
\langle r, s \rangle_{p,w} = C - D = T - (L + R - J) - 2D.
\]

This can be easily seen from the fact that \( C \) is given by the total weight \( T \), minus the weight of discordances \( D \), minus the number of ties, joint or not, which is \( L + R - J \) (we must avoid to count twice the weight of joint ties, hence the \(-J\) term). In particular,

\[
\langle r, r \rangle_{p,w} = T - L \quad \langle s, s \rangle_{p,w} = T - R,
\]
as in this case there are just concordances and all ties are joint.

We are left with the computation of \( D \). The core of Knight’s algorithm is an exchange counter: an \( O(n \log n) \) algorithm that given a list of elements and an order \( \leq \) on the elements of the list computes the number of exchanges that are necessary to \( \leq \)-sort the list. The algorithm is a modified MergeSort [13]\(^8\) during the merging phase, whenever an element is moved from the second list to the temporary result list the current number of elements of the first list is added to the number of exchanges. The number of discords is then equal to the number of exchanges (as we evaluate whether there is a discordance on \( i \) and \( j \) only for \( i < j \)).

Our goal is to make this computation weighted: for this to happen, it must be possible to keep track incrementally of a residual weight \( r \) associated with the first list, and obtain in constant time the resulting recursive procedure (for the additive case) is Algorithm 1. The final layout of the computation of \( D \) from the first list we update the residual by subtracting its weight.

\[ D = \sum_{i < j} wi;j \]

In principle, any stable algorithm that sorts by comparison could be used. This is particularly interesting as entirely on-disk algorithms, such as polyphase merge [15], could be used to count exchanges using constant core memory.

Algorithm 1 A generalization of Knight’s algorithm for weighing exchanges.

\[ \begin{align*}
0 & \quad \text{function} \text{ weight}(s : \text{integer}, \ell : \text{integer}) \\
1 & \quad \text{if} \ell = 1 \text{ then return } f(r(L(s)))
\end{align*} \]

and for the multiplicative case:

\[ T = \sum_{i < j} f(p(i))g(p(j)) = \sum_i f(p(i)) \sum_{i < j} g(p(j)). \]

Both formulae can be computed in linear time using a suitable loop.

Given this setup, it is easy to compute AP correlation: as it can be checked from the very definition [27], the AP correlation of \( R \) w.r.t. \( s \), where both vectors have no ties, is simply \( t_{w,\rho} (r, s) \), where \( \rho \) is the ranking induced by \( s \) and the weight function \( w \) is computed additively from two weight functions \( f(r) = 0, g(r) = 1/r \). In this case, \( T = n - 1 \), \( J = L = R = 0 \) (we are under the assumption that there are no ties) and Algorithm 1 can be considerably simplified, as the residual \( r \) is always zero.\(^7\)

\[ T = \sum_{i < j} \left( f(p(i)) + g(p(j)) \right) = \sum_i f(p(i)) \sum_{i < j} g(p(j)). \]

Both formulae can be computed in linear time using a suitable loop.

Algorithm 2 The replacement for lines 9–15 of Algorithm 1 to compute AP correlation.

\[ \begin{align*}
9 & \quad \text{if} \ L[s + j] \leq L[m + k] \text{ then} \\
10 & \quad \mathcal{T}[i] = L[s + j + 1] \\
11 & \quad \text{else} \\
12 & \quad \mathcal{T}[i] = L[m + k + 1] \\
13 & \quad e \leftarrow e + (\ell_0 - j)/p(\mathcal{T}[i]) \\
14 & \quad \ell_0 \leftarrow \ell_0 + 1
\end{align*} \]
Algorithm 2 makes explicit the change to the selection statement of Algorithm 1 that is sufficient to compute AP correlation. Since keeping track of the residual is no longer necessary, the recursive function can be further simplified to a recursive procedure that does not return a value. The value $e$ computed by the modified algorithm is all we need to compute AP correlation using the formula $(T - 2e)/T$.

<table>
<thead>
<tr>
<th></th>
<th>Indegree</th>
<th>PR</th>
<th>Katz</th>
<th>Harm.</th>
<th>Cl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indegree</td>
<td>1</td>
<td>0.95</td>
<td>0.98</td>
<td>0.90</td>
<td>0.27</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.93</td>
<td>1</td>
<td>0.96</td>
<td>0.92</td>
<td>0.65</td>
</tr>
<tr>
<td>Katz</td>
<td>0.98</td>
<td>0.96</td>
<td>1</td>
<td>0.93</td>
<td>0.26</td>
</tr>
<tr>
<td>Harmonic</td>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.27</td>
<td>0.65</td>
<td>0.26</td>
<td>0.28</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Indegree</th>
<th>PR</th>
<th>Katz</th>
<th>Harm.</th>
<th>Cl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indegree</td>
<td>1</td>
<td>0.76</td>
<td>0.90</td>
<td>0.63</td>
<td>0.55</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.76</td>
<td>1</td>
<td>0.76</td>
<td>0.62</td>
<td>0.56</td>
</tr>
<tr>
<td>Katz</td>
<td>0.90</td>
<td>0.76</td>
<td>1</td>
<td>0.70</td>
<td>0.62</td>
</tr>
<tr>
<td>Harmonic</td>
<td>0.63</td>
<td>0.62</td>
<td>0.70</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.55</td>
<td>0.56</td>
<td>0.62</td>
<td>0.91</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Indegree</th>
<th>PR</th>
<th>Katz</th>
<th>Harm.</th>
<th>Cl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indegree</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.22</td>
</tr>
<tr>
<td>PageRank</td>
<td>1.00</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>Katz</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>1.00</td>
<td>0.18</td>
</tr>
<tr>
<td>Harmonic</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.22</td>
<td>0.85</td>
<td>0.18</td>
<td>0.07</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6: $\tau_h$ on Wikipedia.

Table 7: The logarithmic (top) and quadratic (bottom) additive $\tau$ on Wikipedia.

6. EXPERIMENTS

We now return to our main motivation—understanding the correlation between centralities on large graph. In this section, we gather the results of a number of computational experiments that help to corroborate our intuition that $\tau_h$, the additive hyperbolic weighted $\tau$, or in short, the hyperbolic $\tau$, works as expected. We will find also an interesting surprise along the way.

Note that judging whether a new measure is useful for such a purpose is a difficult task: to be interesting, a new measure must highlight features that were previously undetectable or badly evaluated, but those are exactly those features on which a systematic assessment is problematic.

Table 6 reports the value of $\tau_h$ on the Wikipedia graph. We finally see data corresponding to the empirical evidence discussed in Section 3: indegree, Katz and PageRank are almost identical, harmonic centrality is highly correlated but definitely less than the previous triple, which matches our empirical observations. Closeness is not close to any ranking (and in particular, not to harmonic centrality) due to its pathological behavior.

There is of course a value that immediately stands out: the suspiciously high correlation (0.65) between closeness and PageRank. We reserve discussing this value for later.

In Table 7 we show the same data for logarithmic and quadratic weights. The intuition we gathered from Figure 1 is fully confirmed: logarithmic weights provides results almost indistinguishable from Kendall’s $\tau$ (compare with Table 3), and quadratic weights make the influence of the tail so low that all non-pathological scores collapse.

To gather a better understanding of the behavior of $\tau_h$ we extended our experiments to two very different datasets: the Hollywood co-starship graph, an undirected graph (2 million nodes, 229 million edges) with an edge between two persons appearing in the Internet Movie Data Base if they ever worked together, and a host graph (100 million nodes, 2 billion arcs) obtained from a large-scale crawl gathered by the Common Crawl Foundation in the first half of 2012. 9 As (unavoidably anecdotal) empirical evidence we report the top 20 nodes for both graphs.

Table 8 should be compared with Table 10. PageRank and harmonic centrality turns to be less correlated to indegree than Katz in Table 8, and indeed we find many quirk choices in the very top PageRank actors (Ron Jeremy is a famous porn star; Lloyd Kaufman is an independent horror/splatter filmmaker and Debbie Rochon an actress working with him). Harmonic centrality provides unique names such as Malcolm McDowell, Robert De Niro, Anthony Hopkins and Sylvester Stallone, and drops all USA presidents altogether. Kendall’s $\tau$ values, instead, suggest that PageRank and harmonic centrality are entirely uncorrelated (whereas we find several common items), and that harmonic and closeness centrality should be extremely similar.

We see analogous results comparing Table 9 with Table 11. Here $\tau_h$ separates in a very strong way harmonic centrality from the first three, and indeed we see a significant difference in the lists, with numerous sites that have a high indegree and appear in at least two of the three lists because of technical or political reasons (gmpg.org, rralabel.org, staff.tumblr.com, phpbb.com, milbienan.gov.cn), disappearing altogether in favor of sites such as apple.com, amazon.com, myspace.com, microsoft.com, bbc.co.uk, nytimes.com and guardian.co.uk, which do not appear in any other list. If we look at Kendall’s $\tau$, we should expect PageRank and Katz to give very different rankings, whereas more than half of their top 20 elements are in common.

6.1 PageRank and closeness

It is now time to examine the mysteriously high $\tau_h$ between PageRank and closeness we found in all our graphs. When we first computed our correlation tables, we were puzzled by its value. The phenomenon is interesting for three reasons: first, it has never been reported—using standard, unweighted indices this correlation is simply undetectable; second, it was known for techniques based on singular vectors [17]; third, we know exactly the cause of this correlation, because the only real difference between harmonic and closeness centrality is the score assigned to nodes unreachable from a distinctive host graph, which has a node for each host and an arc between two hosts $x$ and $y$ if some page in $x$ points to some page in $y$. More information about the graph can be found in [19], and the complete host ranking can be accessed at [http://wwwranking.webdatacommons.org/](http://wwwranking.webdatacommons.org/)
Hollywood, which is undirected, these nodes form separate components, omitting nodes which have indegree zero, modulo loops (as all measures give the lowest score to such nodes); then, we sorted the nodes in order of decreasing closeness rank, and plotted for each node its rank following the other measures (we average ranks on block of nodes so to contain the number of points in the plots). A point of high abscissa in the figures implies a high rank.

Table 8: Kendall’s $\tau$ (top) and $\tau_h$ (bottom) on the Hollywood co-starship graph.

The giant component. We thus expect to discover an unsuspected correlation between the way PageRank and closeness rank these nodes.

To have a visual understanding of what is happening, we created Figure 2, 3 and 4 in the following way: first, we isolated the nodes that are unreachable from the giant component (in the case of Hollywood, which is undirected, these nodes form separate components), omitting nodes which have indegree zero, modulo loops (as all measures give the lowest score to such nodes); then, we sorted the nodes in order of decreasing closeness rank, and plotted for each node its rank following the other measures (we average ranks on block of nodes so to contain the number of points in the plots). A point of high abscissa in the figures implies a high rank.

Table 9: Kendall’s $\tau$ (top) and $\tau_h$ (bottom) on the on the Common Crawl host graph.
All three pictures show clearly that PageRank assigns a preposterously high rank to nodes belonging to components that are unreachable from the giant component. This behavior is actually related to PageRank’s well-known insensitivity to size [2]: for instance, in a graph made of two components, one of which is a 3-clique and the other a $k$-clique, the PageRank score of all nodes is $1/(3 + k)$, independently of $k$. The same phenomenon is at work when the community around Lloyd Kaufman’s production company (very small and very dense) is attributed such a great importance that its elements make their way to the very top ranks. Dually, harmonic centrality, the only score sensitive to size among those we are considering, sports consistently the lowest ranks on such nodes.

We remark that the gap in rank is lower in the case of Wikipedia, but this is fully in concordance with the higher baseline value of Kendall’s $\tau$.

7. CONCLUSIONS

In this paper, motivated by the need to understand similarity between score vectors, such as those generated by centrality measures on large graphs, we have defined a weighted version of Kendall’s $\tau$ starting from its 1945 definition for scores with ties. We have developed the mathematical properties of our generalization following a mathematical similarity with internal products, and showing that for a wide range of weighting schemes our new measure behaves as expected, providing a correlation index between -1 and 1, and hitting boundaries only for opposite or equivalent scores.

We have then proposed families of weighting schemes that are intuitively appealing, and showed that they can be computed in time $O(m \log n)$ using a generalization of Knight’s algorithm, which makes them suitable for large-scale applications. The fact that the main cost of the algorithm is a modified stable sort makes it possible to apply standard techniques to run the algorithm exploiting multicore parallelism, or a in distributed environment. The algorithm can be also used to compute AP correlation [27].

In search for a confirmation of our mathematical intuition, we have then applied our measure of choice $\tau_w$ (which uses additive hyperbolic weights) to diverse graph such as Wikipedia, the Hollywood co-starship graph and a large host graph, finding that, contrarily to Kendall’s $\tau$, $\tau_w$ provides results that are consistent with an anecdotal examination of lists of top elements.

Our measure was also able to discover a previously unnoticed correlation between PageRank and closeness on small components that are unreachable from the giant component, providing a quantifiable account of the strong bias of PageRank towards small-sized dense communities. This bias might well be the cause of the repeatedly assessed better performance of indegree w.r.t. PageRank in ranking documents [21, 3], as in all our experiments the $\tau_w$ between PageRank and indegree is above 0.9.

A generalization similar to the one described in this paper can be also applied to Goodman–Kruskal’s $\gamma$, which in the notation of Section 5 is just $(C - D)/(C + D)$. The problem with $\gamma$ is that the contribution of ties is only implicit (they are simply not counted). Thus, the value of $w$ on tied pairs does not appear at all in the above formula. This “forgetful” behavior can lead to unnatural results: for instance, the $\gamma$ correlation between the score vectors $(1, 2, 3)$ and $(1, 2, 2)$ is one, independently of the weights (i.e., even if we let $w_{12}$ go to infinity). This behaviour suggests the Kendall’s $\tau$ is a better candidate for this approach—the $\tau$ correlation is $\sqrt{2}/(2 + w_{12})$, which is 1 when $w_{12} = 0$, $\sqrt{2}/3$ with uniform weights, and goes to zero as $w_{12}$ goes to infinity.

We remark that an interesting application of additive hyperbolic weighting is that of measuring the correlation between top $k$ lists. By assuming that the rank function $\rho$ returns $\infty$ after rank $k$, we obtain a correlation index that weights zero pairs outside the top $k$, weights only “by one side” pairs with just one element outside the top $k$, and weights fully pairs whose elements are within the top $k$. Formula (2) could provide then in principle a finer assessment than, for instance, the modified Kendall’s $\tau$ proposed in [5], as the position of each element inside the list, beside the fact that it appears in the top $k$ or not, would be a source of weight. We leave the analysis of such a correlation measure for future work.
8. REFERENCES


