A Weighted Correlation Index for Rankings with Ties

Sebastiano Vigna*
Università degli Studi di Milano, Italy

October 31, 2014

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 τ; 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 while 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 information retrieval, one is often faced with different scores[1] for the same set of items. This includes the lists of documents returned by different search engines and their associated relevance scores, the lists of query recommendation returned by different algorithms, and also the score associated to each node of a graph by different centrality measures (e.g., indegree and Bavelas's closeness [1]).

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 [24] and Kendall's τ [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 τ, 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 τ 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 global *exogenous* relevance scores in large collections of web documents. The baseline of such scores is indegree—the number of documents containing hypertextual link to a given document. More sophisticated approaches include Katz's index [10], PageRank [21], 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 uttermost importance that the evaluation of correlation takes into account ties as first-class citizens.

On the other hand, Kendall's τ 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 τ value being quite low. This is due to a known phenomenon: the scores

*Sebastiano Vigna has been supported by the EU-FET grant NADINE (GA 288956).

[1] We purposely and consistently use "score" to denote real numbers associated to items, and "rank" to denote ordinal positions. The two terms are used somewhat interchangeably in the literature, but in this paper the distinction is important as we assume that scores of different items can be identical.

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 τ 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 oh Yilmaz, Aslam and Robertson, but taking smoothly ties into account. We will actually define a general notion of weighting for Kendall's τ, 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 item. The same algorithm can be used to compute AP correlation in time $O(n \log n)$.

All data and software used in this paper are available as part of the LAW software library under the GNU General Public License.[2]

2 Related work

Shieh [23] wrote the one of the first papers proposing a generic weighting of Kendall's τ. 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 [6] use instead *penalty weights* to apply Kendall's τ 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 of 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.

We remark that analogous research exists in association with Spearman's correlation: Iman and Conover [9], for example, study the usage of *Savage scores* [22] instead of ranks when comparing ranked lists. Savage scores for a ranked list of n elements are given by $\sum_{j=i}^{n} 1/j$, 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.

Recently, Webber, Moffat and Zobel [26] have described a similarity measure for *indefinite rankings*—rankings that might have different lengths and contain different elements. Their work has some superficial resemblance with the approach of [16, 27] and our work, as it give preeminence to differences at the top of ranked lists, but it is not technically a correlation index, as it is based on measuring overlaps of infinite lists, rather than on exchanges. Thus, the basic condition for a correlation index (i.e., that inverting the list one obtains the minimum possible correlation, usually standardized to -1), is not even expressible in their framework. Moreover, their measure, being defined on infinite lists, needs the fundamental assumption that the weight function applied to overlaps must be *summable*; in particular, they make importance decrease exponentially. As we will discuss in Section 4.2, and verify experimentally in Section 6, such a choice is a reasonable framework for very short lists, or when only very first elements are relevant (e.g., because one is modeling user behavior), but it would completely flatten the results of our correlation index on large examples, depriving it from its discriminatory power, even if the weight function would decrease just quadratically.

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 [7]. In this proposal, Kemeny's distance between two permutations π and σ is characterized as the minimum number of *adjacent transpositions* (i.e., transpositions of the form $(i\, i+1)$) that turn π into σ. 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

[2] `http://law.di.unimi.it/`

to be the minimum sum of weights of a sequence of adjacent transposition that turn π into σ. 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 [23, 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 focus on the case of centrality measures for graphs. Consider the graph of English Wikipedia[3], which has about four million nodes and one hundred million arcs. In this graph, 99.95% of the nodes have the same indegree of some other node—for example, more than a half million node has indegree one. It is clearly mandatory, when computing the correlation of other scores with indegree, that ties are taken into consideration in a systematic way (e.g., not broken arbitrarily).

We will consider four other commonly used scores based on the adjacency matrix A of the Wikipedia graph. One is PageRank [21], which is defined by

$$\mathbf{1}/n \sum_{k\geq 0} (\alpha \bar{A})^k,$$

where $\alpha \in [0\,..\,1)$ is a *damping factor* and $\bar{A}$ is a stochasticization of A: every row not entirely made of zeroes is divided by its sum, so to have ℓ_1 norm one.

The other index we consider is Katz's [10], which is defined by

$$\mathbf{1} \sum_{k\geq 0} (\alpha A)^k,$$

where $\alpha \in [0\,..\,1/\lambda)$ is an attenuation factor depending on λ, the dominant eigenvalue of A [19]. In both cases, we take α in the middle of the allowed interval (using different values does not change the essence of what follows, unless they are extreme).

A different kind of score is provided by Bavelas's *closeness*. The closeness of x is defined by

$$\frac{1}{\sum_{d(y,x)<\infty} d(y,x)},$$

where $d(-,-)$ denotes the usual graph distance. Note that we have to eliminate nodes at infinite distance to avoid zeroing all scores. By definition the closeness of a node with indegree zero is zero. Finally, we consider *harmonic centrality* [2], a modified version of Bavelas's closeness designed for directed graphs that are not strongly connected; the harmonic centrality of x is defined by

$$\sum_{y\neq x} \frac{1}{d(y,x)}.$$

These scores provide an interesting mix: indegree is an obvious baseline, and entirely local. PageRank and Katz are similar in their definition, but the normalization applied to A makes the scores quite different (at least in theory). Finally, closeness and harmonic centrality are of a completely different nature, having no connection with dominant eigenvectors or Markov chains.

Our first empirical observation is that, looking just at the very top pages of Wikipedia (Table 1; entries in boldface are unique to the list they belong to, here and in the following), we perceive these scores as almost identical, except for closeness, which displays almost random values. The latter behavior is a known phenomenon: nodes that are almost isolated obtain a very high closeness score (this is why harmonic centrality was devised). We note also that harmonic centrality has a slightly different slant, as it is the only ranking including Latin, Europe, Russia and the Catholic Church in the top 20.

The problem is that these facts are not reflected in any way in the values of Kendall's τ shown in Table 3. If we exclude closeness, with the exception of the correlation between indegree and Katz, all other correlation value fail to

[3]More precisely, a specific snapshot of Wikipedia that will be made public by the author. The graph does not contain template pages.

Indegree	PageRank	Katz	Harmonic	Closeness
United States	United States	United States	United States	**Kharqan Rural District**
List of sovereign states	Animal	List of sovereign states	United Kingdom	**Talageh-ye Sofla**
Animal	List of sovereign states	United Kingdom	World War II	**Talageh-ye Olya**
England	France	France	France	**Greatest Remix Hits (Whigfield album)**
France	Germany	Animal	Germany	**Suzhou HSR New Town**
Association football	Association football	World War II	Association football	**Suzhou Lakeside New City**
United Kingdom	England	England	English language	**Mepirodipine**
Germany	India	Association football	China	**List of MPs ... M–N**
Canada	United Kingdom	Germany	Canada	**List of MPs ... O–R**
World War II	Canada	Canada	India	**List of MPs ... S–T**
India	Arthropod	India	**Latin**	**List of MPs ... U–Z**
Australia	Insect	Australia	World War I	**List of MPs ... J–L**
London	World War II	London	England	**List of MPs ... C**
Japan	Japan	Italy	Italy	**List of MPs ... F–I**
Italy	Australia	Japan	**Russia**	**List of MPs ... A–B**
Arthropod	Village	New York City	**Europe**	**List of MPs ... D–E**
Insect	Italy	English language	Australia	**Esmaili-ye Sofla**
New York City	Poland	China	**European Union**	**Esmaili-ye Olya**
English language	English language	Poland	**Catholic Church**	**Levels of organization (ecology)**
Village	**National Reg. of Hist. Places**	World War I	London	**Jacques Moeschal (architect)**

Table 1: Top 20 pages of the English version of Wikipedia following five different centrality measures.

surpass the 0.9 threshold, usually considered the threshold for considering two rankings equivalent [25]. 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 100 000 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 denormalized *arithmetic* mean, whereas harmonic centrality is the reciprocal of a denormalized *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.

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 Definitions and Tools

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 $\boldsymbol{r}$ and $\boldsymbol{s}$ (to be thought as scores) with indices in $[n]$; then, let us define

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

Indegree	PageRank	Katz	Harmonic	Closeness
Carl Linnaeus	Carl Linnaeus	Carl Linnaeus	Aristotle	**Noël Bernard (botanist)**
Aristotle	Aristotle	Aristotle	Albert Einstein	**Charles Coquelin**
Thomas Jefferson	Thomas Jefferson	Thomas Jefferson	Thomas Jefferson	**Markku Kivinen**
Margaret Thatcher	Charles Darwin	Albert Einstein	Charles Darwin	**Angiolo Maria Colomboni**
Plato	Plato	Charles Darwin	Thomas Edison	**Om Prakash (historian)**
Charles Darwin	Albert Einstein	Karl Marx	**Alexander Graham Bell**	**Michel Mandjes**
Karl Marx	Karl Marx	Plato	**Nikola Tesla**	**Kees Posthumus**
Albert Einstein	Pliny the Elder	Margaret Thatcher	**William James**	**F. Wolfgang Schnell**
Vladimir Lenin	Vladimir Lenin	Vladimir Lenin	Isaac Newton	**Christof Ebert**
Sigmund Freud	Johann Wolfgang von Goethe	Isaac Newton	Karl Marx	**Reese Prosser**
J. R. R. Tolkien	Margaret Thatcher	Ptolemy	**Charles Sanders Peirce**	**David Tulloch**
Johann Wolfgang von Goethe	Ptolemy	Johann Wolfgang von Goethe	Noam Chomsky	**Kim Hawtrey**
Spider-Man	Sigmund Freud	Pliny the Elder	**Enrico Fermi**	**Patrick J. Miller**
Pliny the Elder	Isaac Newton	Benjamin Franklin	Ptolemy	**Mikel King**
Benjamin Franklin	Benjamin Franklin	J. R. R. Tolkien	**John Dewey**	**Albert Perry Brigham**
Leonardo da Vinci	J. R. R. Tolkien	Thomas Edison	Johann Wolfgang von Goethe	**Gordon Wagner (economist)**
Isaac Newton	Immanuel Kant	Sigmund Freud	**Bertrand Russell**	**George Henry Chase**
Ptolemy	Leonardo da Vinci	Immanuel Kant	Plato	**Charles C. Horn**
Immanuel Kant	**Pierre André Latreille**	Leonardo da Vinci	**John von Neumann**	**Paul Goldstene**
George Bernard Shaw	Thomas Edison	Noam Chomsky	Vladimir Lenin	**Robert Stanton Avery**

Table 2: Top 20 pages of Wikipedia following five different centrality measures and restricting pages to Yago2 Wordnet category "scientist". The global rank of these items is beyond 300.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.75	0.90	0.62	0.55
PageRank	0.75	1	0.75	0.61	0.56
Katz	0.90	0.75	1	0.70	0.62
Harmonic	0.62	0.61	0.70	1	0.92
Closeness	0.55	0.56	0.62	0.92	1

Table 3: Kendall's τ between Wikipedia centrality measures.

where

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

Indices of score vectors in summations belong to $[n]$ throughout the paper. Note that the expression above is actually an inner product in a space of dimension $n(n-1)$: each score vector $\boldsymbol{r}$ is mapped the vector with coordinate $\langle i, j\rangle$, $i < j$, given by $\operatorname{sgn}(r_i - r_j)$. We have the property

$$\langle \boldsymbol{r}, \alpha \boldsymbol{s}\rangle = \langle \alpha \boldsymbol{r}, \boldsymbol{s}\rangle = \operatorname{sgn}(\alpha)\langle \boldsymbol{r}, \boldsymbol{s}\rangle,$$

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

$$\|\boldsymbol{r}\| := \sqrt{\langle \boldsymbol{r}, \boldsymbol{r}\rangle},$$

so

$$\|\alpha \boldsymbol{r}\| = |\operatorname{sgn}(\alpha)| \cdot \|\boldsymbol{r}\|.$$

The norm thus defined measures the "untieness" of $\boldsymbol{r}$: it is zero if and only if $\boldsymbol{r}$ is a constant vector, and it has maximum value $\sqrt{n(n-1)/2}$ when all components of $\boldsymbol{r}$ are distinct.

Since $\langle \boldsymbol{r}, \boldsymbol{s}\rangle$ is an inner product on a larger space, we have a Cauchy–Schwartz-like inequality:

$$|\langle \boldsymbol{r}, \boldsymbol{s}\rangle| \leq \|\boldsymbol{r}\|\|\boldsymbol{s}\|$$

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

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

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.31	0.63	0.24	0.06
PageRank	0.31	1	0.27	0.10	0.10
Katz	0.63	0.27	1	0.50	0.20
Harmonic	0.24	0.10	0.50	1	0.65
Closeness	0.06	0.10	0.20	0.65	1

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

We recall that if $\boldsymbol{r}$ and $\boldsymbol{s}$ 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

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

where $w(-,-) : [n] \times [n] \to \mathbf{R}_{\geq 0}$ is some nonnegative symmetric weight function. We would have then a new norm $\|\boldsymbol{r}\|_w = \sqrt{\langle \boldsymbol{r}, \boldsymbol{r} \rangle_w}$ and a new correlation index

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

Note that still $\langle \boldsymbol{r}, - \rangle_w = \langle -, \boldsymbol{r} \rangle_w = 0$ if $\boldsymbol{r}$ is constant.

We say that two score vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ are *equivalent* if $\operatorname{sgn}(r_i - r_j) = \operatorname{sgn}(s_i - s_j)$, *opposite* if $\operatorname{sgn}(r_i - r_j) = -\operatorname{sgn}(s_i - s_j)$ for all i and j. Since $\langle \boldsymbol{r}, \boldsymbol{s} \rangle$ 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 *$|\langle \boldsymbol{r}, \boldsymbol{s} \rangle_w| \leq \|\boldsymbol{r}\|_w \|\boldsymbol{s}\|_w$. 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 $\|\boldsymbol{r}\|_w, \|\boldsymbol{s}\|_w \neq 0$ or $\|\boldsymbol{r}\|_w, \|\boldsymbol{s}\|_w = 0$*

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 inequation for the induced "norm":

Theorem 2 *$\|\boldsymbol{r} + \boldsymbol{s}\|_w \leq \|\boldsymbol{r}\|_w + \|\boldsymbol{s}\|_w$.*

Proof.

$$\begin{aligned}
&\|\boldsymbol{r} + \boldsymbol{s}\|_w^2 = \langle \boldsymbol{r} + \boldsymbol{s}, \boldsymbol{r} + \boldsymbol{s} \rangle_w \\
&= \sum_{i<j} \operatorname{sgn}(r_i + s_i - r_j - s_j)^2 w(i,j) \\
&\leq \sum_{i<j} (|\operatorname{sgn}(r_i - r_j)| + |\operatorname{sgn}(s_i - s_j)|)^2 w(i,j) \\
&= \langle \boldsymbol{r}, \boldsymbol{r} \rangle_w + \langle \boldsymbol{s}, \boldsymbol{s} \rangle_w + \sum_{i<j} |\operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j)| w(i,j).
\end{aligned}$$

We now notice that

$$\sum_{i<j} |\operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j)| w(i,j) \leq \sum_{i<j} \operatorname{sgn}(r_i - r_j)^2 w(i,j) = \langle \boldsymbol{r}, \boldsymbol{r} \rangle_w = \|\boldsymbol{r}\|_w^2,$$

and analogously for $\|\boldsymbol{s}\|_w$. We conclude that

$$\|\boldsymbol{r} + \boldsymbol{s}\|_w^2 \leq \|\boldsymbol{r}\|_w^2 + \|\boldsymbol{s}\|_w^2 + 2\|\boldsymbol{r}\|_w \|\boldsymbol{s}\|_w = (\|\boldsymbol{r}\|_w + \|\boldsymbol{s}\|_w)^2. \blacksquare$$

Indegree	PageRank	Katz	Harmonic	Closeness
Martini (cocktail)	Martini (cocktail)	Irish coffee	Irish coffee	**Magie Noir**
Piña colada	Caipirinha	Caipirinha	Caipirinha	**Batini (drink)**
Mojito	Mojito	Martini (cocktail)	Kir (cocktail)	**Scorpion bowl**
Caipirinha	Piña colada	Piña colada	Martini (cocktail)	**Poinsettia (cocktail)**
Cuba Libre	Irish coffee	Kir (cocktail)	Piña colada	Irish coffee
Irish coffee	Kir (cocktail)	Mojito	Mojito	Caipirinha
Singapore Sling	Cosmopolitan (cocktail)	Mai Tai	Beer cocktail	Kir (cocktail)
Manhattan (cocktail)	Manhattan (cocktail)	Cuba Libre	Shaken, not stirred	Martini (cocktail)
Windle (sidecar)	IBA Official Cocktail	Singapore Sling	Pisco Sour	Piña colada
Cosmopolitan (cocktail)	Beer cocktail	Long Island Iced Tea	Mai Tai	Mojito
Mai Tai	Mai Tai	Shaken, not stirred	Spritz (alcoholic beverage)	Beer cocktail
IBA Official Cocktail	Singapore Sling	Beer cocktail	Long Island Iced Tea	Shaken, not stirred
Kir (cocktail)	Cuba Libre	Manhattan (cocktail)	Sazerac	Mai Tai
Shaken, not stirred	**Tom Collins**	Cosmopolitan (cocktail)	Fizz (cocktail)	Spritz (alcoholic beverage)
Beer cocktail	Long Island Iced Tea	Windle (sidecar)	Flaming beverage	Pisco Sour
Pisco Sour	Sour (cocktail)	Pisco Sour	Cuba Libre	Long Island Iced Tea
Long Island Iced Tea	Shaken, not stirred	White Russian (cocktail)	Wine cocktail	Sazerac
Sour (cocktail)	**Negroni**	IBA Official Cocktail	Singapore Sling	Flaming beverage
White Russian (cocktail)	Flaming beverage	Moscow mule	Moscow mule	Fizz (cocktail)
Vesper (cocktail)	**Lillet**	Vesper (cocktail)	White Russian (cocktail)	Wine cocktail

Table 5: Top 20 pages of Wikipedia following five different centrality measures and restricting pages to Yago2 Wordnet category "cocktail". The global rank of these items is beyond 100 000.

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 τ_w:

Theorem 3 *Let $w : [n] \times [n] \to \mathbf{R}$ be a nonnegative symmetric weight function. The following properties hold for every score vector $\boldsymbol{t}$ and for every $\boldsymbol{r}$, $\boldsymbol{s}$ with nonnull norm:*

- *if $\boldsymbol{t}$ is constant, $\|\boldsymbol{t}\|_w = 0$;*
- *$-1 \leq \tau_w(\boldsymbol{r}, \boldsymbol{s}) \leq 1$;*
- *if $\boldsymbol{r}$ and $\boldsymbol{s}$ are equivalent, $\tau_w(\boldsymbol{r}, \boldsymbol{s}) = 1$;*
- *if $\boldsymbol{r}$ and $\boldsymbol{s}$ are opposite, $\tau_w(\boldsymbol{r}, \boldsymbol{s}) = -1$;*

Moreover, if w is strictly positive:

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

As a result, if w is strictly positive and we obtain correlation ± 1 the equivalence classes formed by tied scores are necessarily in a size-preserving bijection that is monotone decreasing on the scores.

4.1 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 τ, 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 $\boldsymbol{r}$ and $\boldsymbol{s}$. 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

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

where $\rho : [n] \to [n] \cup \{\infty\}$ is a ranking function associating with each index a *rank*, the highest rank being zero. We admit the possibility of rank ∞, 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 $w(\rho(-), \rho(-))$ 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 $\boldsymbol{r}$ and $\boldsymbol{s}$. Let us denote by $\rho_{\boldsymbol{r},\boldsymbol{s}}$ the ranking defined by ordering elements lexicographically with respect to $\boldsymbol{r}$ and then $\boldsymbol{s}$ in case of a tie (in descending order), and analogously for $\rho_{\boldsymbol{s},\boldsymbol{r}}$ (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 define

$$\tau_{w,\bullet}(\boldsymbol{r}, \boldsymbol{s}) := \frac{\tau_{\rho_{\boldsymbol{r},\boldsymbol{s}},w}(\boldsymbol{r}, \boldsymbol{s}) + \tau_{\rho_{\boldsymbol{s},\boldsymbol{r}},w}(\boldsymbol{r}, \boldsymbol{s})}{2}. \tag{2}$$

The same approach has been used in [27] to make AP correlation symmetric. This is the definition used in the rest of the paper.

4.2 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 $f : [n] \to \mathbf{R}_{\geq 0}$ applied to each element of a pair.

Shieh [23], 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* [6] by assigning rank ∞ 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.

$$f(r) := \frac{1}{r+1}.$$

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 τ.

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 τ and the additive weighted τ 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 τ, whereas quadratic weighting ends up in providing answers that are too uncorrelated. We will return to these consideration in Section 6.

5 Computing $\tau_{\rho,w}$

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 $\tau_{\rho,w}$ 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 $s_i = s_j$;

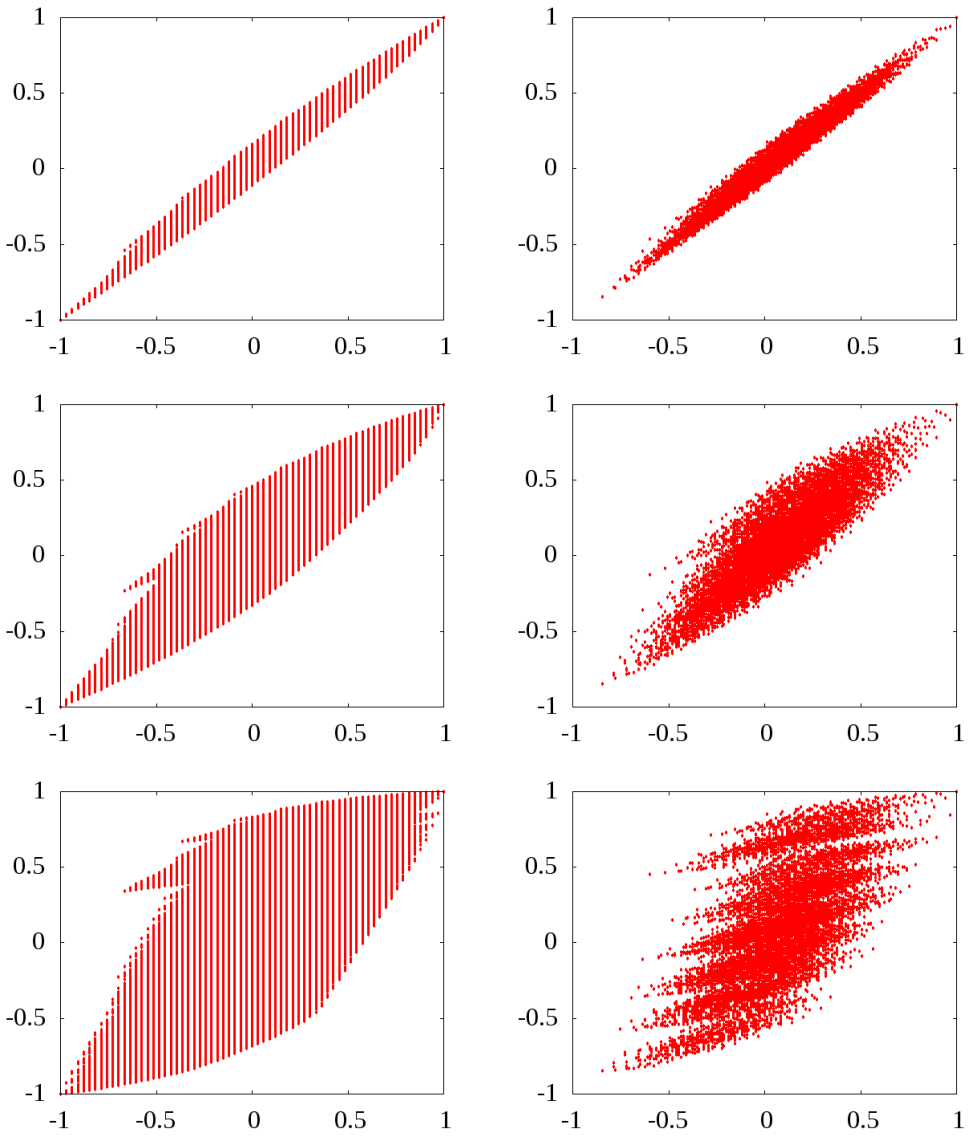

Figure 1: Scatter plots between Kendall's τ and the additive weighted τ. The rows, from top to bottom, represent logarithmic, hyperbolic and quadratic weighting. The plots are generated correlating a permutation of 12 elements versus the identity permutation (left), or a permuted set of scores with skewed distribution w.r.t. the same scores in descending order (right).

- 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 $\operatorname{sgn}(r_i - r_j)\operatorname{sgn}(s_i - s_j) = 1$;
- a *discordance*, if $\operatorname{sgn}(r_i - r_j)\operatorname{sgn}(s_i - s_j) = -1$.

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

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

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,

$$T = \sum_{i<j} \big(f(\rho(i)) + f(\rho(j))\big) = (n-1)\sum_i f(\rho(i)). \tag{3}$$

Also the multiplicative case is easy, as

$$2T = 2\sum_{i<j} f(\rho(i))f(\rho(j)) = \left(\sum_i f(\rho(i))\right)^2 - \sum_i f(\rho(i))^2. \tag{4}$$

The same observation leads to a simple $O(n \log n)$ algorithm to compute L: sort the indices in $[n]$ by $\boldsymbol{r}$, and for each block of consecutive $k > 1$ 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,

$$\langle \boldsymbol{r}, \boldsymbol{s} \rangle_{\rho,w} = C - D = T - (L + R - J) - 2D.$$

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,

$$\langle \boldsymbol{r}, \boldsymbol{r} \rangle_{\rho,w} = T - L \qquad \langle \boldsymbol{s}, \boldsymbol{s} \rangle_{\rho,w} = T - R,$$

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 $\preceq$ on the elements of the list computes the number of exchanges that are necessary to $\preceq$-sort the list. The algorithm is a modified MergeSort [15][4]: 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 discordances 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 weight of the exchanges generated by the movement of an element from the second list.

If weights are computed multiplicatively or additively starting from a single-argument function f this is not difficult: it is sufficient to let r be the sum of the values of f applied to the elements currently in the first list. In the additive case, moving an element i from the second list increases the weight of exchanges by the residual r plus the weight $f(\rho(i))$ multiplied by the length of the first list. In the multiplicative case, we must instead use the weight $f(\rho(i))$ multiplied by the residual r. When we move an element from the first list we update the residual by subtracting its weight.

The resulting recursive procedure (for the additive case) is Algorithm 1. The final layout of the computation of $\tau_{\rho,w}$ is thus as follows:

- Consider a list $\mathscr{L}$ initially filled with the integers $[0\,..\,n)$.

[4] 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.

- Sort stably $\mathscr{L}$ using $\boldsymbol{r}$ as primary key and $\boldsymbol{s}$ as secondary key.
- Compute T and L using $\mathscr{L}$ to enumerate elements in the order defined by $\boldsymbol{r}$ and $\boldsymbol{s}$.
- Apply Algorithm 1 to $\mathscr{L}$ using $\boldsymbol{s}$ to define the order $\preceq$, thus computing D and sorting $\mathscr{L}$ by $\boldsymbol{s}$.
- Compute R using $\mathscr{L}$ to enumerate elements in the order defined by $\boldsymbol{s}$.
- Compute T and put everything together.

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

Input: A list $\mathscr{L}$, a comparison function $\preceq$ for the elements of $\mathscr{L}$, a rank function ρ, and a single-argument weight function f that will be combined additively. e is a global variable initialized to 0 that will contain the weight of exchanges after the call weigh(0, $|\mathscr{L}|$). The procedure works on a sublist specified by its starting index $0 \le s < |\mathscr{L}|$ and its length ℓ. $\mathscr{T}$ is a temporary list.
Output: the sum of $f(\rho(-))$ on the specified sublist.

```
0   function weigh(s : integer, ℓ : integer)
1     if ℓ = 1 then return f(ρ(ℒ[s])) fi
2     ℓ₀ ← ⌊ℓ/2⌋
3     ℓ₁ ← ℓ − ℓ₀
4     m ← s + ℓ₀
5     r ← weigh(s, ℓ₀)
6     w ← weigh(m, ℓ₁) + r
7     i, j, k ← 0
8     while j < ℓ₀ and k < ℓ₁ do
9       if ℒ[s + j] ⪯ ℒ[m + k] then
10        𝒯[i] = ℒ[s + j++]
11        r ← r − f(ρ(𝒯[i]))
12      else
13        𝒯[i] = ℒ[m + k++]
14        e ← e + f(ρ(𝒯[i])) · (ℓ₀ − j) + r
15      fi
16      i++
17    od
18    for k = ℓ₀ − j − 1, ..., 0 do
19      ℒ[s + i + k] ← ℒ[s + j + k]
20    od
21    for k = 0, ..., i − 1 do ℒ[s + k] ← 𝒯[k] od
22    return w
23  end
```

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

```
9   if ℒ[s + j] ⪯ ℒ[m + k] then
10    𝒯[i] = ℒ[s + j++]
11  else
12    𝒯[i] = ℒ[m + k++]
13    e ← e + (ℓ₀ − j)/ρ(𝒯[i])
14  fi
```

The running time of the computation is dominated by the sorting phases, and it is thus $O(n \log n)$.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.95	0.98	0.90	0.27
PageRank	0.95	1	0.96	0.92	0.65
Katz	0.98	0.96	1	0.93	0.26
Harmonic	0.90	0.92	0.93	1	0.28
Closeness	0.27	0.65	0.26	0.28	1

Table 6: τ_{h} on Wikipedia.

5.1 The asymmetric case and AP Correlation

It is easy to adapt Algorithm 1 for the case in which $w(i, j)$ is given by a combination of *two* different one-argument functions, one, f, for the left index and one, g, for the right index. The only modification of Algorithm 1 is the replacement of f with g at line 14, so that we combine the residual computed with f with a weight computed with g.

The formulae for computing T can be updated easily for the additive case:

$$T = \sum_{i<j} \big(f(\rho(i)) + g(\rho(j))\big) = \sum_{i\neq 0} i(f(\rho(n-1-i)) + g(\rho(i)))$$

and for the multiplicative case:

$$T = \sum_{i<j} f(\rho(i))g(\rho(j)) = \sum_{i} f(\rho(i)) \sum_{i<j} g(\rho(j)).$$

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 $\boldsymbol{r}$ w.r.t. $\boldsymbol{s}$, where both vectors have no ties, is simply $\tau_{w,\rho_{\boldsymbol{s}}}(\boldsymbol{r}, \boldsymbol{s})$, where $\rho_{\boldsymbol{s}}$ is the ranking induced by $\boldsymbol{s}$ 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.[5]

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 $(T - 2e)/T$.

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 experiment that help to corroborate our intuition that τ_{h}, the *additive hyperbolic weighted* τ, 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 τ_{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.

[5]Of course, it is possible to forget that we are computing AP correlation and use the weight matrix just described combined with the machinery of Section 4 to define an "AP correlation with ties". In this case, J, L and R should be computed using the formulae for the asymmetric case, and the probabilistic interpretation would be lost. Such an index would probably give a notion of correlation very similar to τ_{h}, but we find more natural and more in line with Kendall's original definition to introduce the weighted τ as a symmetric index in which both ends of an exchange are relevant in computing the exchange weight.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.76	0.90	0.63	0.55
PageRank	0.76	1	0.76	0.62	0.56
Katz	0.90	0.76	1	0.70	0.62
Harmonic	0.63	0.62	0.70	1	0.91
Closeness	0.55	0.56	0.62	0.91	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	1.00	1.00	1.00	0.22
PageRank	1.00	1	1.00	1.00	0.85
Katz	1.00	1.00	1	1.00	0.18
Harmonic	1.00	1.00	1.00	1	0.07
Closeness	0.22	0.85	0.18	0.07	1

Table 7: The logarithmic (top) and quadratic (bottom) additive τ on Wikipedia.

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 τ (compare with Table 3), and quadratic weighs make the influence of the tail so low that all non-pathological scores collapse.

To gather a better understanding of the behavior of τ_{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[6] in the first half of 2012.[7] 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 τ 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 τ_{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`, `rtalabel.org`, `staff.tumblr.com`, `miibeian.gov.cn`, `phpbb.com`) 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 τ, 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 τ_{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 the

[6] `http://commoncrawl.org/`

[7] The crawl contains 3.53 billion web documents; we are using the associated 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 [18], and the complete host ranking can be accessed at `http://wwwranking.webdatacommons.org/`.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.42	0.93	0.55	0.43
PageRank	0.42	1	0.36	0.10	0.18
Katz	0.93	0.36	1	0.61	0.49
Harmonic	0.55	0.10	0.61	1	0.86
Closeness	0.43	0.18	0.49	0.86	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.90	0.98	0.91	0.10
PageRank	0.90	1	0.88	0.81	0.64
Katz	0.98	0.88	1	0.92	0.11
Harmonic	0.91	0.81	0.92	1	0.18
Closeness	0.10	0.64	0.11	0.18	1

Table 8: Kendall's τ (top) and τ_{h} (bottom) on the Hollywood co-starship graph.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.71	0.89	0.61	0.54
PageRank	0.71	1	0.66	0.50	0.50
Katz	0.89	0.66	1	0.69	0.59
Harmonic	0.61	0.50	0.69	1	0.86
Closeness	0.54	0.50	0.59	0.86	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.91	0.96	0.72	0.20
PageRank	0.91	1	0.90	0.81	0.69
Katz	0.96	0.90	1	0.78	0.15
Harmonic	0.72	0.81	0.78	1	0.35
Closeness	0.20	0.69	0.15	0.35	1

Table 9: Kendall's τ (top) and τ_{h} (bottom) on the on the Common Crawl host graph.

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.

All three pictures show clearly that *PageRank assigns a preposterously high rank to to nodes belonging to components that are unreachable from the giant component*. This behavior is actually related to PageRank's *insensitivity to size*: 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. This explains why small dense components end up being so highly ranked. 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 (even if Kaufman himself has indegree rank 219 and Debbie Rochon 1790).

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 τ.

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 τ starting from its 1945 definition for scores with ties. We have developed the mathematical properties of our generalization following a mathematical similarity

Indegree	PageRank	Katz	Harmonic	Closeness
Shatner, William	Jeremy, Ron	Shatner, William	Sheen, Martin	**Östlund, Claes Göran**
Flowers, Bess	Hitler, Adolf	Sheen, Martin	Clooney, George	**Östlund, Catarina**
Sheen, Martin	**Kaufman, Lloyd**	Hanks, Tom	Jackson, Samuel L.	**von Preußen, Oskar Prinz**
Reagan, Ronald (I)	Bush, George W.	Williams, Robin (I)	Hopper, Dennis	**von Preußen, Georg Friedrich**
Clooney, George	Reagan, Ronald (I)	Clooney, George	Hanks, Tom	**von Mannstein, Robert Grund**
Jackson, Samuel L.	Clinton, Bill (I)	Reagan, Ronald (I)	Stone, Sharon (I)	**von Mannstein, Concha**
Williams, Robin (I)	Sheen, Martin	Willis, Bruce	Brosnan, Pierce	**von der Busken, Mart**
Hanks, Tom	**Rochon, Debbie**	Jackson, Samuel L.	Hitler, Adolf	**van der Putten, Thea**
Jeremy, Ron	**Kennedy, John F.**	Stone, Sharon (I)	**McDowell, Malcolm**	**de la Bruheze, Joel Albert**
Hitler, Adolf	Hopper, Dennis	Freeman, Morgan (I)	Williams, Robin (I)	**de la Bruheze, Emile**
Willis, Bruce	**Nixon, Richard**	Flowers, Bess	**De Niro, Robert**	**te Riele, Marloes**
Clinton, Bill (I)	**Estevez, Joe**	Brosnan, Pierce	Willis, Bruce	**de Reijer, Eric**
Freeman, Morgan (I)	Shatner, William	Douglas, Michael (I)	**Hopkins, Anthony**	**des Bouvrie, Jan**
Hopper, Dennis	Jackson, Samuel L.	Madonna (I)	Madonna (I)	**de Klijn, Judith**
Stone, Sharon (I)	**Stewart, Jon (I)**	Travolta, John	**Lee, Christopher (I)**	**de Freitas, Luís (II)**
Madonna (I)	**Carradine, David (I)**	Hopper, Dennis	Douglas, Michael (I)	**de Freitas, Luís (I)**
Bush, George W.	Asner, Edward	Ford, Harrison (I)	**Sutherland, Donald (I)**	**Zuu, Winnie Otondi**
Harris, Sam (II)	**Zirnkilton, Steven**	Asner, Edward	Freeman, Morgan (I)	**Zuu, Emmanuel Dahngbay**
Brosnan, Pierce	**Colbert, Stephen**	**MacLaine, Shirley**	**Stallone, Sylvester**	**Zilbersmith, Carla**
Travolta, John	**Madsen, Michael (I)**	Clinton, Bill (I)	Ford, Harrison (I)	**Zilber, Mac**

Table 10: Top 20 pages of the Hollywood co-starship graph.

Indegree	PageRank	Katz	Harmonic	Closeness
wordpress.org	gmpg.org	wordpress.org	youtube.com	**0-p.com**
youtube.com	wordpress.org	youtube.com	en.wikipedia.org	**0-0-0-0-0-0-0.indahiphop.ru**
gmpg.org	youtube.com	gmpg.org	twitter.com	**0-0-1.i.tiexue.net**
en.wikipedia.org	**livejournal.com**	en.wikipedia.org	google.com	**0-00cigarettes.info**
tumblr.com	tumblr.com	tumblr.com	wordpress.org	**0-0mos00.hi5.com**
twitter.com	en.wikipedia.org	twitter.com	flickr.com	**0-0new0-0.hi5.com**
google.com	twitter.com	google.com	facebook.com	**0-0sunny0-0.hi5.com**
flickr.com	**networkadvertising.org**	flickr.com	**apple.com**	**0-1.i.tiexue.net**
rtalabel.org	**promodj.com**	rtalabel.org	vimeo.com	**0-1.sxsy.co**
wordpress.com	**skriptmail.de**	wordpress.com	creativecommons.org	**0-2.paparazziwannabe.com**
mp3shake.com	**parallels.com**	mp3shake.com	**amazon.com**	**0-311.cn**
w3schools.com	**tistory.com**	w3schools.com	**adobe.com**	**0-360.rukazan.ru**
domains.lycos.com	google.com	creativecommons.org	**myspace.com**	**0-5days.com**
staff.tumblr.com	miibeian.gov.cn	staff.tumblr.com	**w3.org**	**0-5days.net**
club.tripod.com	phpbb.com	domains.lycos.com	**bbc.co.uk**	**0-5kalibr.pdj.ru**
creativecommons.org	**blog.fc2.com**	club.tripod.com	**nytimes.com**	**0-9-0-4-4-9.promoradio.ru**
vimeo.com	**tw.yahoo.com**	vimeo.com	**yahoo.com**	**0-9-0-9.dbass.ru**
miibeian.gov.cn	w3schools.com	miibeian.gov.cn	**microsoft.com**	**0-9-0-9.promodj.ru**
facebook.com	wordpress.com	facebook.com	**guardian.co.uk**	**0-9-1125.i.tiexue.net**
phpbb.com	domains.lycos.com	phpbb.com	**imdb.com**	**0-9-7-16.software.informer.com**

Table 11: Top 20 hosts of the Common Crawl host graph.

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 $O(n \log n)$ 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 in distributed environment such as MapReduce [5]. 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 τ_{h} (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 τ, τ_{h} 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 [20, 3], as in all our experiments the τ_{h} 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* γ, which in the notation of Section 5 is just $(C - D)/(C + D)$. The problem with γ is that the ranking of ties is only implicit (they

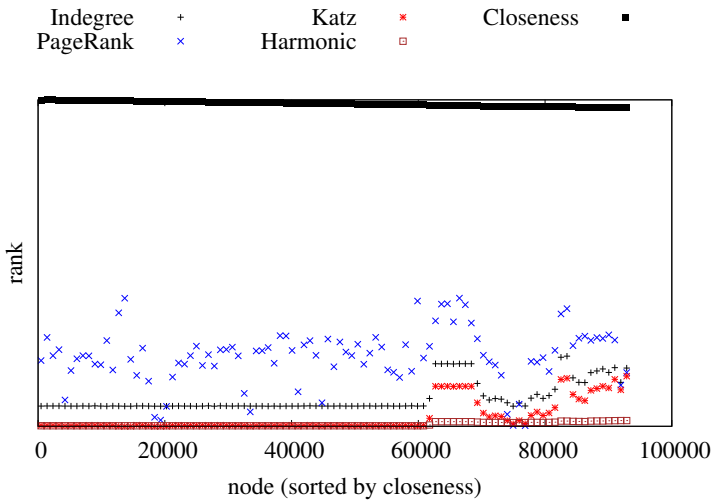

Figure 2: Ranks of components unreachable from the giant component of the Wikipedia graph.

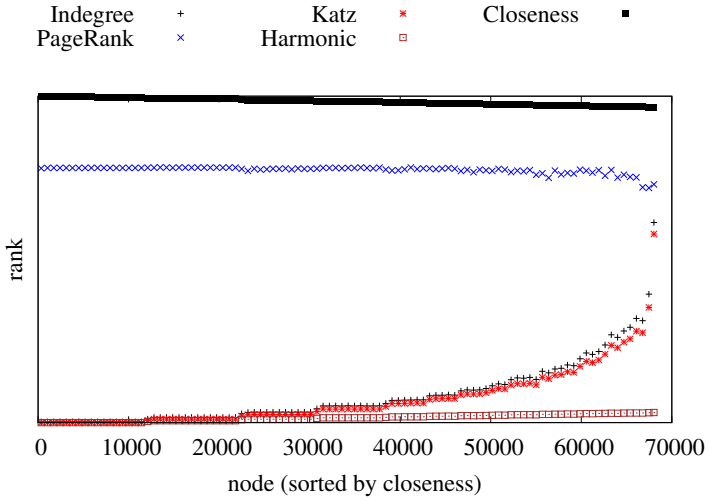

Figure 3: Ranks of components of the Hollywood co-starship graph, excluding the giant component.

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, and suggests the Kendall's τ is a better candidate for this approach.

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 ρ returns ∞ after rank k, we obtain a correlation index that weighs 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 τ proposed in [6], 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.

References

[1] Alex Bavelas. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.*, 22(6):725–730, 1950.

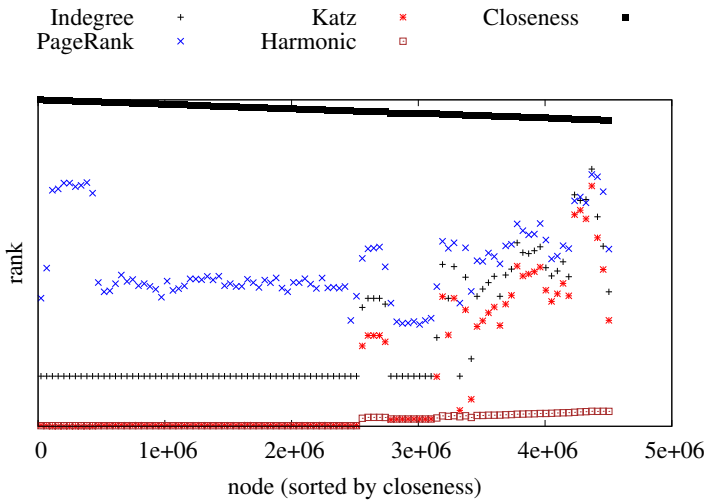

Figure 4: Ranks of components unreachable from the giant component of the Common Crawl host graph.

[2] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *CoRR*, abs/1308.2140, 2013. To appear in *Internet Mathematics*.

[3] Nick Craswell, David Hawking, and Trystan Upstill. Predicting fame and fortune: PageRank or indegree? In *In Proceedings of the Australasian Document Computing Symposium, ADCS2003*, pages 31–40, 2003.

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

[5] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI '04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, 2004.

[6] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[7] F. Farnoud and O. Milenkovic. An axiomatic approach to constructing distances for rank comparison and aggregation. *IEEE Trans. on Information Theory*, 60(10):6417–6439, October 2014.

[8] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[9] Ronald L. Iman and W. J. Conover. A measure of top-down correlation. *Technometrics*, 29(3):351–357, 1987.

[10] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[11] John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

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

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

[14] 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.

[15] Donald E. Knuth. *Sorting and Searching*, volume 3 of *The Art of Computer Programming*. Addison-Wesley, second edition, 1997.

[16] Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, pages 571–580. ACM, 2010.

[17] Ronny Lempel and Shlomo Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*, 33(1):387–401, 2000.

[18] Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer. Graph structure in the web — Revisited, or a trick of the heavy tail. In *WWW'14 Companion*, pages 427–432. International World Wide Web Conferences Steering Committee, 2014.

[19] Carl D. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, 2000.

[20] Marc Najork, Hugo Zaragoza, and Michael J. Taylor. HITS on the web: how does it compare? In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 471–478. ACM, 2007.

[21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, Stanford University, 1998.

[22] I. Richard Savage. Contributions to the theory of rank order statistics—the two-sample case. *The Annals of Mathematical Statistics*, 27(3):590–615, 1956.

[23] Grace S. Shieh. A weighted kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998.

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

[25] Ellen M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 2001.

[26] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst*, 28(4):20:1–20:38, 2010.

[27] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594. ACM, 2008.