The Definition of a New Correlation Variant for Rankings With Ties

Exploratory Definitions of the w-variant in τ, τ_{AP}, τ_h

Mikołaj Jakub Gazeel[†]

Supervised by:
Prof. Julián Urbano[†]

[†]Faculty of Electrical Engineering,
Mathematics and Computer Science,
Delft University of Technology, The Netherlands

22 June 2025

A Thesis Submitted to EEMCS Faculty of the Delft University of Technology, in Partial Fulfillment of the Requirements for the Bachelor of Computer Science and Engineering.

Name of the student: .... Mikołaj Jakub Gazeel
Final project course: .... `CSE3000` Research Project
Thesis committee: .... Prof. Julián Urbano, Prof. Lilika Markatou

This work does not utilize Large Language Models nor generative Artificial Intelligence in any written, theoretical, or experimental part.

The Definition of a New Correlation Variant for Rankings With Ties

Mikołaj Jakub Gazeel
Technische Universiteit Delft
Delft, The Netherlands
mgazeel@tudelft.nl

ABSTRACT

Rankings are simply orderings given to a set of elements; They are a widely used mathematical object in information retrieval. This creates the need for some means of comparing them. Rank Similarity Measures are used exactly for this. They constitute a large research area where many different such measures are defined. A ranking may possibly contain ties. This in turn raises the question of what these ties represent and how to treat them in the calculation of a measure. The treatment of ties in current theory is approached with the a and b variants of the measures. Both a and b stem from a statistical approach to ties; they consider tied elements to represent uncertainty about their real order in the ranking. There is, however, a different interpretation of what ties could represent, namely that the tied elements really occur at the same place in the ranking, that is, there is no intrinsic order in which they should appear. This has been considered in one of the nonconjoint measures and has been coined the w-variant after Webber et al. In this work, we consider the problem of defining this very variant for a family of three commonly used ranking similarity measures, these being τ defined by Kendall, τ_{AP} defined by Yilmaz et al., and τ_h defined by Vigna. We approach this problem by establishing what the variant should represent and defining a set of axioms that any definition of w has to follow. Thereafter, we show that there is only one definition which can possibly satisfy these, with a small exception. We show that this definition coincides with the distance considered by Kemeny in 1959. We use this to create a definition of the w-variant for all three of the measures. Likewise, we investigate the behaviour of this new variant in relation to the existing a and b variants. Moreover, we identify the shortcomings of our definition and evaluate it on real world data. Finally, we lay the groundwork for rigorously proving parts of our definition and other measures which may consider ties to represent occurrence at the same rank.

CCS CONCEPTS

• Information systems → Evaluation of retrieval results; • Mathematics of computing → Exploratory data analysis

KEYWORDS

Rankings, Rank similarity, Ties, Variant, Kendall Tau, Sports Rankings, Rank Correlation Coefficient

ACM Reference Format:

Mikołaj Jakub Gazeel. 2025. The Definition of a New Correlation Variant for Rankings With Ties. Bachelor thesis. Delft University of Technology, Delft, The Netherlands.

This work is a bachelor thesis submitted to EEMCS faculty of the Delft University of Technology in Delft, The Netherlands. To the end of partial fulfillment of the requirements for the Bachelor of Science in Computer Science and Engineering.
It consists of a total of 10 pages of which 8 constitute the main matter.
Accessible at: `https://repository.tudelft.nl/`

1 INTRODUCTION

A *ranking* is an assignment of an order to some predefined set of elements. Figure 1 contains an example of such a ranking on 5 elements.

1st ranking	B	C	A	E	D
2nd ranking	C	A	E	D	B

Figure 1: Two example rankings. With elements: $\{A, B, ...\}$

Rankings are an important part of today's world of data. Just one example of these, that almost everyone interacts with daily are search engines, but rankings represent much more than that as identified in [1], [9]. From athletics to research institutes - these are a ubiquitous part of how we rationalize information about many topics.

To have the ability to propose changes or improvements in ranking algorithms, one needs some quantification of similarity between their outputs. This presents the need for comparisons between rankings.

1.1 Similarity measures and definitions

Similarity measures, sometimes referred to as *rank correlation measures*[1], attempt to address exactly this. They serve as a measure of 'closeness' or 'similarity' between rankings.

There is no one objective approach to defining such a measure. Existing approaches are, however, classified based on their properties into categories. *Weighted* measures are ones which put more importance on some parts of the ranking than others. Meanwhile, *unweighted* measures consider all places in the ranking to have the same impact on the result. We say that a measure is *conjoint* if, for the calculation of its score, we have to present the entire ranking on all elements. *Nonconjoint* measures on the other hand can compare partial rankings, where only some top part of the raking is seen. One could imagine a comparison of first page results from two search engines, some of them may overlap and some may not but, neither ranks all the web pages on the internet.

One of the first similarity measures was defined by Spearman [10] and coined ρ, while another by Kendall [5] and coined τ. τ is widely used today and is a conjoint unweighted similarity measure.

Yilmaz et al. [17] introduced τ_{AP} - a conjoint weighted measure based on τ. It puts more importance on differences closer to the top of the ranking. Building on this work Vigna [12] introduced τ_h which produces a score given an arbitrary weighting scheme, and is a superset of both τ and τ_{AP}. In the context of this work it is also important to mention one nonconjoint measure, namely Rank-Biased Overlap (RBO) introduced by Webber et al. [15].

These measures, given two rankings, are meant to return a value between −1 and 1. 1 means perfect agreement between the rankings and −1 means that they are negatively correlated

[1] Although talking about correlation, depending on what is being compared, may not always make sense.

(perfect disagreement). As an example of this, the score for the rankings of Figure 1 is $\rho(1^{\text{st}}, 2^{\text{nd}}) = 0$, $\tau(1^{\text{st}}, 2^{\text{nd}}) = 0.2$, $\tau_{AP}(1^{\text{st}}, 2^{\text{nd}}) \approx -0.042$, $\tau_{AP}(2^{\text{st}}, 1^{\text{nd}}) = 0.5$.

In this work to denote rankings we will use $\langle\rangle$, for example: 1^{st} from Figure 1 becomes $\langle B, C, A, E, D\rangle$, B occupies the first place, C the second, and so on. Note that the choice of the letters $A, B, \ldots$ is completely arbitrary, there is just a need for some way to identify the elements common to both of the rankings being compared. In fact we will iterate on these letters as if they were elements of $\mathbb{N}$.

Our focus will be τ, τ_{AP}, and τ_h. These are some of the simplest and most deployed measures and we endeavor to define our contribution in terms of these.

1.2 Tied elements

Ties may appear in a ranking. A tie is a situation when two distinct elements are ranked at the same place. Ties will be denoted by [], so for example: $\langle [B, C], D, [A, E, F]\rangle$ - the first two and last three elements are tied.

Let R_n be the set of all possible rankings of length n without ties. Whereas, the set of all rankings of length n possibly including ties will be denoted by $\hat{R}_n$. This also gives: $R_n \subset \hat{R}_n$

A key question that Pearson first identified[2], is how to interpret ties when comparing rankings, and how they should influence the final score of the ranking. For conjoint rankings two widely adopted approaches exist to this end, they will referred to as *variants*.

Unlike other works referenced before, to avoid confusion between the different measures, variants, and parameters, we will denote the variant, as in [2], by a superscript. So for example: the a variant of τ_{AP} will be denoted by τ_{AP}^a.

1.2.1 The meaning of ties

These two widely adopted approaches have been described in the original work of Kendall [5]. Kendall assumes that tied items represent some kind of uncertainty about the actual order of the elements. In other words, during the comparison, we assume that the tied elements have a certain objective order but the creator of the ranking was not able to set it and hence the measure does not have access to it. Both of the variants which are introduced here stem from this base interpretation. This assumption is crucial for the contribution in this work.

These two variants are, as shown by [11]:

a) Based on the work of Woodbury [16], [6] identifies τ^a, it is the value of τ averaged over all permutations of the tied items. So in other words each possible uncertain ordering is weighed equally.
b) Based on the work of "Student" [18], [6] identifies τ^b, which is similar to τ^a but with the denominator of the original adjusted to still reach the extreme values for fully concordant or discordant rankings which contain ties.

1.2.2 A new meaning of ties

In the process of introducing RBO Webber et al. [15] consider its behaviour, when the ranking includes ties. Before the modification to the agreement function of RBO they argue about the interpretation of ties for the definition:

> *«Ties may be handled by assuming that, if t items are tied for ranks d to* $d + (t - 1)$*, they all occur at rank d. To support this, we modify the definition…»*

[2] "Student" makes this claim in [18].

Moreover, as seen in [11], even Kendall [6] mentions this interpretation of ties, however, he never expands on it further.

What can be observed here is that the assumption from Section 1.2.1 may not be necessary to consider ties. Indeed if we instead assume that ties represent items which are **meant to be exactly** at the same rank, we arrive at a new variant.

As an illustration, one could imagine an ice skating competition and two distinct juries each awarding a discrete amount of points to the participants based on their form, speed, etc. At the end of the competition these will produce two rankings of the participants, but what if two participants are tied? How do we compare the output of these juries? The a and b variants are not useful in this case, since both of them assume that there is an inherent order to the tied participants, whereas in our case the 'tying' of participants itself provides information on their relationship.

Urbano & Marrero [11], after defining τ_{AP}^a, τ_{AP}^b, go further and clearly identify the gap in knowledge:

> *«For future work we will consider a third scenario that Kendall […] mentioned implicitly but did not consider explicitly […] he assumed that a tie was given when the observer was unable to discern a difference, but it may be the case that the tied elements are in fact equal in the true ranking.»*

Corsi & Urbano [2] reiterate the arguments of Webber et al. [15] and their definition, identify this as a new variant, and coin it the w-variant after Weber. This way they provide a definition of RBO^w.

To the best of our knowledge this variant has not been considered in any work in terms of conjoint rankings, and hence leads us to the question:

How can the w-variant be defined for conjoint measures, specifically τ, τ_{AP}, and τ_h?

2 APPROACH AND DESCRIPTION

Two different approaches were identified in regards of creating the definition for w.

2.1 RBO-based

Webber et al. [15] and Corsi & Urbano [2] define RBO^a, RBO^b, and RBO^w. To define RBO^a the interpretation of τ^a was used, analogous steps were taken for RBO^b. Would it be possible to define τ^w based upon RBO^w? This would essentially reverse the direction taken in [2] for RBO^a.

In the end this approach was deemed inadequate. Both the definitions and behaviour of RBO^w differ so significantly from all τ measures that no easy mapping or parallel can be created. RBO^w, as a whole inherently gives more significance to items which are higher in the ranking. τ is completely unweighted. RBO^w, in its agreement function, is based on the intersection of sets for a given prefix of the ranking. This makes it inadequate for defining τ, which considers pairs of items instead.

In this case even the interpretation of ties in conjoint rankings is not clear.

2.2 Axiomatic

Instead we consider a different approach, one where we stipulate **how** τ^w should behave, **what it should represent** by choosing a set of axioms. Thereafter, we continue and create a definition which fits all of our axioms. We will further see that the axioms lead to a specific solution.

2.2.1 Background on τ and its variants

Kendall defines τ over R_n in terms of *concordant* and *discordant* pairs. A pair is concordant when two elements appear in the same order in both rankings, and vice-versa for discordant pairs:

$$\tau(l, r) = \frac{\sum_{i<j} \operatorname{sgn}(l_i - l_j)\operatorname{sgn}(r_i - r_j)}{D} \tag{1}$$

With a certain scaling factor $D = \frac{n(n-1)}{2}$ which ensures that the result always falls in $[-1, 1]$.

l_i represents the rank (from 1 to n) of element i in ranking l. For example if we have rankings $l = \langle A, B, C\rangle$ and $r = \langle B, A, C\rangle$ the pair (A, B), is discordant with respect to l, r and hence it will count as: $\operatorname{sgn}(1-2)\operatorname{sgn}(2-1) = -1$ in the summation.

We will refer to the function $\operatorname{sgn}(l_i - l_j)\operatorname{sgn}(r_i - r_j)$ as $g_{l,r}(i, j)$. This function, given two elements, tells us if and how much they agree between the two rankings l and r. τ can now be written as:

$$\tau(l, r) = \frac{\sum_{i<j} g_{l,r}(i, j)}{D} = \frac{N}{D} \tag{2}$$

τ^a can be represented by the same Equation 1, however it is defined over $\hat{R}_n$. This means that $g^a_{l,r}(i, j)$, is the same as $g_{l,r}(i, j)$, but immediately returns 0 if the pair is tied in either ranking. We define N^a analogously. It just so happens that this definition is the average value of τ across all permutations.

Moreover, in τ^b, we have $g^b_{l,r}(i, j) = g^a_{l,r}(i, j)$, but the denominator is a value dependent on the raking. We will elaborate on this further when talking about τ_h.

There is a common theme in these definitions, all of them consist of a *numerator* which is the measure of similarity, and a *denominator* which is a way to scale this measure to the required range.

2.2.2 The interpretations of τ

τ has many interpretations, the amount of concordant and discordant pairs is just one of them. Yilmaz et al. [17] present τ as an expectation of a random variable, this being that a random pick of two elements is concordant.

More importantly, the numerator[3] of τ can also be used to define a metric space over the set R_n by a slight modification [7]. To create a distance function[4] from τ we have to consider that distance is a non-negative measure of *dissimilarity*, hence we have to first negate N to $-N$ and then add 1. This will give us a distance between l and r as required:

$$d\tau(l, r) = 1 - N = \frac{n(n-1)}{2} - \sum_{i<j} g_{l,r}(i, j) \tag{3}$$

$$= \sum_{i<j} 1 - g_{l,r}(i, j) = \sum_{i<j} d_{l,r}(i, j) \tag{4}$$

The new function $d_{l,r}(i, j)$ is just negated and 'shifted' by 1: for a concordant pair it returns 0 and for a discordant pair it returns 2. The distance interpretation is central to our approach.

2.2.3 The usefulness of τ_h and introduction to τ_{AP}

Vigna [12] introduces a weighted variant of τ, which is now referred to as τ_h. This variant besides the two rankings requires a weighting scheme. It stems from a dot product interpretation of the numerator of τ. Therefore first a dot product is defined[5] for any $l, r \in R_n$:

$$l \star_w r = \sum_{i<j} g_{l,r}(i, j) w_{l,r}(i, j) \tag{5}$$

The function w is the weighting function. For a pair of items it determines their contribution to the result. This function may depend on the rankings themselves[6].

Thereafter, a 'norm' is introduced, and τ_h is defined as:

$$\|x\|_w = \sqrt{x \star_w x} \tag{6}$$

$$\tau_{h,w}(l, r) = \frac{l \star_w r}{\|l\|_w \cdot \|r\|_w} = \frac{N_{h,w}}{D_{h,w}} \tag{7}$$

We observe: If $w_{l,r}(i, j) = 1$ is chosen this just becomes τ, since $D_{h,1}$ will be exactly D and $N_{h,1} = N$.

Yilmaz et al. [17] introduce τ_{AP}. While the original definition stems from the probabilistic interpretation of τ, for the purposes of this work we note that $\tau_{AP}(l, r)$ is exactly $\tau_{h,w}(l, r)$ with $w_{l,r}(i, j) = \frac{1}{\max(l_i, l_j) - 1}$. As seen from the weighting function τ_{AP} is asymmetric and weighs items based on their position in one of the rankings, here we chose l to be the ranking upon which weights are based but this is arbitrary.

We have learned that τ and τ_{AP} can both be defined by a certain choice of weighting in τ_h. We will leverage this very fact to provide a definition for τ^w_h which can then be easily appropriated in order to produce τ^w_{AP}.

To this end we have to first investigate the behaviour of τ_h in the presence of ties: $g_{l,r}(i, j)$ in Equation 5 is replaced by $g^a_{l,r}(i, j)$. The numerator stays the same as both τ^a and τ^b for $w(i, j) = 1$, and the denominator is the same as in τ^b. We will not focus on $D_{h,w}$ since it is not central to this work.

2.2.4 Final method

The definition of τ^w will be based on the distance interpretation of τ. In particular, note that $1 - N^a$ is not a valid distance for $\hat{R}_n$, as will be shown later.

We claim: To treat tied items as really occurring at the same place the numerator of τ^w must result in a metric for $\hat{R}_n$. This ensures that a valid distance is assigned to any pair of rankings with ties, and as such ties are taken into account accordingly as distinct entities.

We will use this very fact to define a distance metric $d\tau^w$ and from it arrive at τ^w. The axiomatic approach continues as follows:

1. A set of axioms which describe the behaviour of $d\tau^w$ (the distance metric based on N^w), and τ^w, will be chosen.
2. $d\tau^w$ will be defined based on those axioms, and we will show that the definition is both necessary and sufficient.
3. τ^w will be defined based on $d\tau^w$ and the axioms from 1.
4. τ^w_h will be defined as a weighted variant of τ^w, just like [12].
5. τ^w_{AP} will follow immediately from τ^w_h.

2.2.5 Kemeny distance

Before proceeding with the method, we mention a contribution which is the base of our approach to the axioms and $d\tau^w$.

The very question of a distance metric on rankings with ties has already been investigated by Kemeny [4]. At the end of his 1959 article a definition for distance function on $\hat{R}_n$ is provided. Further, Kemeny uses an approach where the choice of axioms leads to a specific solution.

[3]The denominator does not matter here, it just scales the whole distance space. This is akin to choosing a unit of measure.

[4]We will denote the distance function based on τ as $d\tau$.

[5]In [12] this operation is written as $\langle x, y\rangle$, to avoid confusion with rankings here we use $\star$.

[6]This is expressed by ρ in the original work.

Kemeny does not provide proofs for the necessity or sufficiency of the axioms stating that they have been omitted. Moreover, Kemeny does not consider this distance in the context of a similarity measure.

Our approach to defining $d\tau^w$ builds on that of Kemeny and we adopt *conditions* 1 to 3 in our axioms as 1 to 3 (reordered), with the exception that we reject the notion of *betweenness* from *condition 1* and instead introduce stipulations which make sense in the context of a correlation measure.

To ease the comprehension of the next section we will not reference Kemeny [4] in each of the axioms which stem from his conditions. We want to stress that [4] had significant impact on these choices as well as the idea behind the axiomatic approach. This work can be seen as an extension of that into the world of similarity measures.

3 THE AXIOMS AND THE DEFINITION

We will start with the axioms concerning the distance measure. This follows immediately from our claim:

Axiom 1. $d\tau^w(l,r)$ *should be a valid distance function for the metric space over* $\hat{R}_n$.

From the definition of a metric space [3] we can expand this:

Axiom 1.1. $d\tau^w(l,r) \geq 0$ *for all* l, r.

Axiom 1.2. $d\tau^w(l,r) = 0$ *if and only if* $l = r$.

Axiom 1.3. $d\tau^w(l,r) = d\tau^w(r,l)$.

Axiom 1.4. $d\tau^w(x,z) \leq d\tau^w(x,y) + d\tau^w(y,z)$ *for all* x, y, z.

Just like in τ, the identities of the elements should not influence the final result:

Axiom 2. Given two rankings l *and* r *if in both of them we pick two elements and exchange their places creating* l' *and* r', *we have:* $d\tau^w(l,r) = d\tau^w(l',r')$.

The definition of our distance must be restricted[7] to resemble that of $d\tau$. It should only depend on the relative ordering of pairs of items.

Axiom 3. $d\tau^w(l,r) = \sum_{i<j} d^w_{l,r}(i,j)$. *Where* $d^w_{l,r}(i,j)$ *is some real function dependent only on* $\operatorname{sgn}(l_i - l_j)$ *and* $\operatorname{sgn}(r_i - r_j)$.

We must establish whether fully discordant elements or a tie mismatch leads to a higher distance between two rankings. Logically, a tie mismatch should result in a smaller penalty in the similarity measure than a full discordance:

Axiom 4. $d\tau^w(\langle A,B\rangle, \langle B,A\rangle) \geq d\tau^w(\langle A,B\rangle, \langle[A,B]\rangle)$

Now let us consider the behaviour of τ^w. First we define the numerator:

Axiom 5. Let m *be the maximal possible value of* $d\tau^w(l,r)$. *We have:* $N^w = \frac{m}{2} - d\tau^w(l,r)$.

The denominator should scale the numerator to the appropriate range of $[-1, 1]$.

Axiom 6. $\tau^w(l,r) = \frac{N^w}{D^w}$, *with* D^w *such that:*

1. *For any* l *and* r, $\tau^w(l,r) \in [-1,1]$
2. *For any* x, $\tau^w(x,x) = 1$.

As with τ^a and τ^b, in the absence of ties τ^w should equal τ.

Axiom 7. If $l, r \in R_n$ *then* $\tau^w(l,r) = \tau(l,r)$.

Finally, correlation measures like τ are expected to return 0 if the rankings are uncorrelated. In other words for two rankings chosen uniformly at random we should expect a 0.

Axiom 8. Define two independent variables L *and* R *where for all* $x \in \hat{R}_n$ *we have* $\Pr[L = x] = \Pr[R = x] = \frac{1}{|\hat{R}_n|}$. *Now it should be the case that:* $\mathbb{E}[\tau^w(L,R)] = 0$.

[7]This axiom sounds very specific but it is equivalent to the more general *Condition 3* of [4] combined with *Axiom 2*. In fact *Axiom 2* is weaker version of *Axiom 3*. Both are included for completeness.

3.1 Arriving at the definition by necessity

All of the axioms are defined as things that the w variant must satisfy. We will see that they lead to a certain necessary definition.

From *Axiom 3* we know that we need to have:

$$d\tau^w = \sum_{i<j} d^w_{l,r}(i,j) \tag{8}$$

Can we define the function $d^w_{l,r}(i,j)$ such that all the other axioms are satisfied? As a first step consider just using $d^a_{l,r}(i,j)$ to arrive the same definition as $d\tau^a$.

Claim: $d\tau^a$ violates axioms.

Proof: Consider the ranking $x = \langle[A,B]\rangle$. Now: $d\tau^a(x,x) = 1$ but $x = x$, which violates *Axiom 1.2.* ■

We can represent $d^w_{l,r}(i,j)$ as a decision table (Table 1), where for a certain combination of relative rankings a given value is output. Meanwhile, $d^a_{l,r}(i,j)$ can be represented as Table 2. This covers all possible definitions of d^w pursuant to *Axiom 3.*

Here '$<$' denotes that $\operatorname{sgn}(x_i - x_j) = -1$ in the given ranking, '$=$' that the elements are tied etc.

$l \backslash r$	$<$	$=$	$>$
$<$	α	β	γ
$=$	δ	ε	ζ
$>$	μ	θ	λ

Table 1: $d^w_{l,r}(i,j)$

$l \backslash r$	$<$	$=$	$>$
$<$	0	1	2
$=$	1	1	1
$>$	2	1	0

Table 2: $d^a_{l,r}(i,j)$

Claim: All values in Table 1 must be non-negative under the given axioms.

Proof: Assume for the sake of contradiction that one of the values is negative. Without loss of generality let that be β. Produce two rankings l, r of size 2 which make $d^w_{l,r}(A,B)$ output that value. For example for β, these would be $l = \langle A,B\rangle, r = \langle[A,B]\rangle$. Now $d\tau^w(l,r) = \beta$ but $\beta < 0$, this violates *Axiom 1.1.* ■

Claim: $\alpha, \varepsilon, \lambda$ must be 0 under the given axioms.

Proof: Assume for the sake of contradiction that one of $\alpha, \varepsilon, \lambda$ is non-zero. Consider the ranking: $x = \langle C, A, [B,D]\rangle$. $d\tau^w(x,x) = 3\alpha + 2\lambda + \varepsilon$, from the assumption this gives $d\tau^w(x,x) > 0$, since all vales are non-negative, and violates *Axiom 1.2.* ■

Claim: $\mu = \gamma$ under the given axioms.

Proof: $d\tau^w(\langle A,B\rangle, \langle B,A\rangle) = \gamma = d\tau^w(\langle B,A\rangle, \langle A,B\rangle) = \mu$ by *Axiom 1.3* ■

Claim: $\delta = \beta = \theta = \zeta$ under the given axioms.

Proof: $d\tau^w(\langle[A,B]\rangle, \langle A,B\rangle) = \delta = d\tau^w(\langle A,B\rangle, \langle[A,B]\rangle) = \beta$ and $d\tau^w(\langle[A,B]\rangle, \langle B,A\rangle) = \zeta = d\tau^w(\langle B,A\rangle, \langle[A,B]\rangle) = \theta$ by *Axiom 1.3.* Moreover, $d\tau^w(\langle[A,B]\rangle, \langle A,B\rangle) = \delta = d\tau^w(\langle[B,A]\rangle, \langle B,A\rangle) = \zeta$ by *Axiom 2.* ■

We will now rename the variables in Table 1 to Table 3, considering the above claims.

$l \backslash r$	$<$	$=$	$>$
$<$	0	β	α
$=$	β	0	β
$>$	α	β	0

Table 3: $d^w_{l,r}(i,j)$

It is important to note that the actual values of α and β do not matter since this just scales the whole distance metric which will be scaled back by D^w. What is now left to do is to is to establish the relationship between α and β. We continue by inferring some constraints on $\frac{\alpha}{\beta}$ from *Axiom 1.4.*

If the rankings chosen in Appendix A are used, we immediately get $\frac{\alpha}{\beta} \leq 2$ since *Axiom 1.4* has to be true for **all** rankings.

Claim: $\frac{\alpha}{\beta} \geq 1$ under the given axioms.

Axiom 4: $d\tau^w(\langle A,B\rangle, \langle B,A\rangle) \geq d\tau^w(\langle A,B\rangle, \langle [A,B]\rangle)$ this is simply: $\alpha \geq \beta \Rightarrow \frac{\alpha}{\beta} \geq 1$ ■

At this point we mention that $\frac{\alpha}{\beta}$ will be concretely determined by *Axiom 8*. We postpone using *Axiom 8* to empirically decide the value of $\frac{\alpha}{\beta}$ to the very end. However, we now show that any value $1 \leq \frac{\alpha}{\beta} \leq 2$ satisfies all the other axioms. Therefore for now:

$$d^w_{l,r}(i,j) = \begin{cases} \alpha & \text{if } \operatorname{sgn}(l_i - r_i)\operatorname{sgn}(l_j - r_j) = -1 \\ \beta & \text{if } \operatorname{sgn}(l_i - r_i) = 0 \text{ xor } \operatorname{sgn}(l_j - r_j) = 0 \\ 0 & \text{otherwise} \end{cases}$$

With:

$$1 \leq \frac{\alpha}{\beta} \leq 2 \tag{9}$$

3.2 Sufficiency

It was shown that this is how $d\tau^w$ must be defined, now the other direction is shown. Namely, that this definition satisfies all axioms pertaining to the distance function.

Claim: $d\tau^w_{l,r}(i,j)$ satisfies *Axiom 1*.

Proof Axiom 1.1: All entries in Table 3 are non-negative, the sum of non-negative values in non-negative.

Proof Axiom 1.2: $d\tau^w(l,l) = 0$ trivially. For the other direction, consider that $d\tau^w(l,r) = 0$ only if all the elements of the sum for all pairs are 0 since there are no negative values in Table 3. Since each pair of elements is in the same relationship in both rankings $l = r$ necessarily.

Proof Axiom 1.3: Table 3 is symmetric along the diagonal.

Proof Axiom 1.4: We will show that this axiom holds for each possible pair of elements. With this we will be able to sum the inequalities and show the axiom holds for the distance. In Appendix A we show that for any elements i, j the following is true:

$$\forall_{x,y,z} \quad d^w_{x,z}(i,j) \leq d^w_{x,y}(i,j) + d^w_{y,z}(i,j) \tag{10}$$

We can now sum Equation 10 over all pairs of elements $i < j$:

$$\forall_{x,y,z} \quad \sum_{i<j} d^w_{x,z}(i,j) \leq \sum_{i<j}\big(d^w_{x,y}(i,j) + d^w_{y,z}(i,j)\big) \tag{11}$$

$$\forall_{x,y,z} \quad \sum_{i<j} d^w_{x,z}(i,j) \leq \sum_{i<j} d^w_{x,y}(i,j) + \sum_{i<j} d^w_{y,z}(i,j) \tag{12}$$

$$d\tau^w(x,z) \leq d\tau^w(x,y) + d\tau^w(y,z) \tag{13}$$

As required. ■

Claim: $d^w_{l,r}(i,j)$ satisfies *Axioms 2,3*.

Proof Axiom 2: Table 3 is symmetric along the horizontal, vertical, and both diagonal axes. Exchanging two elements in a pair is equivalent to reflecting them along one of these axes in the table.■

Axiom 3 follows immediately from the definition.

3.3 τ^w

Armed with $d\tau^w$ we can define N^w.

To use *Axiom 5* we have to determine the maximal possible value of $d\tau^w$. Since $d\tau^w$ is a sum of non-negative terms, it will be the maximal value of $d^w_{l,r}(i,j)$, which is α since $\alpha \geq \beta$, accounted for $\frac{n(n-1)}{2}$ times since this is the amount of pairs $i < j$. Continuing with *Axiom 5*:

$$N^w = \frac{\alpha\frac{n(n-1)}{2}}{2} - d\tau^w(l,r) \tag{14}$$

$$N^w = \left(\frac{n(n-1)}{2}\right)\frac{\alpha}{2} - \sum_{i<j} d^w_{l,r}(i,j) \tag{15}$$

$$N^w = \sum_{i<j} \frac{\alpha}{2} - d^w_{l,r}(i,j) = \sum_{i<j} g^w_{l,r}(i,j) \tag{16}$$

This leads to $g^w_{l,r}(i,j)$ being Table 4, in contrast to $g^a_{l,r}(i,j)$ (Table 5) from τ^a.

$l \backslash r$	<	=	>
<	$\frac{\alpha}{2}$	$\frac{\alpha}{2} - \beta$	$-\frac{\alpha}{2}$
=	$\frac{\alpha}{2} - \beta$	$\frac{\alpha}{2}$	$\frac{\alpha}{2} - \beta$
>	$-\frac{\alpha}{2}$	$\frac{\alpha}{2} - \beta$	$\frac{\alpha}{2}$

Table 4: $g^w_{l,r}(i,j)$

$l \backslash r$	<	=	>
<	1	0	−1
=	0	0	0
>	−1	0	1

Table 5: $g^a_{l,r}(i,j)$

Next, in *Axiom 6* we define the denominator. Looking back at the definition of τ_h in our case N^w can serve as the '⋆' operator. This choice indeed satisfies the constraints of *Axiom 6*. The crucial observation here is that, the norm is a square root of the sum of the values on the diagonal of Table 4 over all pairs. Ties or no ties, in the case of D^w this value will always be:

$$D^w = \left(\sqrt{\frac{n(n-1)}{2}\frac{\alpha}{2}}\right)^2 = \frac{\alpha n(n-1)}{4} \tag{17}$$

Without loss of generality we can pick any value for α, since D^w removes its contribution to the result. To simplify we pick $\alpha = 2$. This gives us an almost final definition for τ^w:

$$\tau^w(l,r) = \frac{\sum_{i<j} g^w_{l,r}(i,j)}{\frac{n(n-1)}{2}} \tag{18}$$

Where:

$$g^w_{l,r}(i,j) = \tag{19}$$

$l \backslash r$	<	=	>
<	1	$1-\beta$	−1
=	$1-\beta$	1	$1-\beta$
>	−1	$1-\beta$	1

And from Equation 9:

$$\beta \leq 2 \leq 2\beta \Rightarrow \beta \in [1,2] \tag{20}$$

Axiom 7 is instantly satisfied since the corners of Table 4 and Table 5 agree.

The only degree of freedom left in the definition now is the choice of β. Let us tackle this problem next.

3.4 Independent rankings

We have approached both the necessity and sufficiency of the previous axioms quite rigorously. However, for *Axiom 8* we were unable to find a discrete or general solution for an arbitrary n. The value of β which satisfies *Axiom 8* is different for example between $n = 2$ and $n = 3$. Therefore a definition fully conforming to it without introducing dependence on the length of the ranking is simply impossible.

Instead, we consider an empirical approach where β is explicitly found for small n and its value for larger n is extrapolated. This evaluation shows that the values of β for different n differ by a relatively small amount and the results show a convergent tendency. We will use this extrapolation to approximate the (assumed) convergence value and choose β as such.

We can express *Axiom 8*, for a particular n:

$$\mathbb{E}[\tau^w(L,R)] = \sum_{l,r \in \hat{R}_n} \tau^w(l,r)\Pr[L = l]\Pr[R = r] \tag{21}$$

$$= \frac{1}{|\hat{R}_n|^2} \sum_{l,r \in \hat{R}_n} \frac{N^w}{D^w} = 0 \quad (22)$$

$$\Rightarrow \frac{1}{|\hat{R}_n|^2 \; D^w} \sum_{l,r \in \hat{R}_n} \left(\sum_{i<j} g^w_{l,r}(i,j) \right) = 0 \quad (23)$$

$$\Rightarrow \sum_{l,r \in \hat{R}_n} \left(\sum_{i<j} g^w_{l,r}(i,j) \right) = 0 \quad (24)$$

3.4.1 Empirical experiment

All elements of $\hat{R}_n$ were generated for n from 2 to 7, while ensuring that there are no duplicates. Thereafter, the incidence of each type of pair was counted across all rankings. All of these concrete values can be found in Appendix B. With the incidence values and Equation 24 we can calculate the value of β which would satisfy *Axiom 8* for each length.

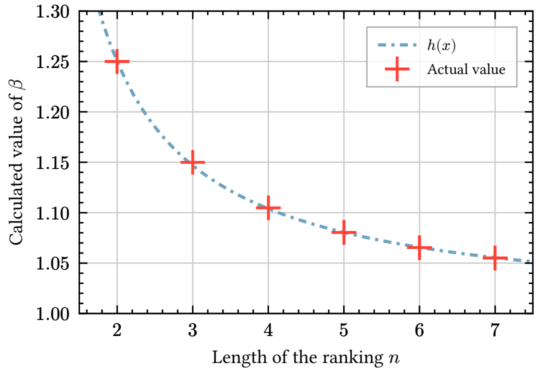

Figure 2: The value of β satisfying *Axiom 8* versus n.

In Figure 2 the function $h(x) = 0.356(\frac{1}{x-0.576}) + 1$ was included as an extrapolation of the data points into larger values of n.

$h(x)$ converges to 1. All the values of β are very close to 1 and (seemingly) converge to it. This prompts our final choice, which is: $\boldsymbol{\beta = 1}$. This is motivated by *Axiom 3* which prevents dependence on the length of the ranking, and the convergence value observed above. The implications of $\beta = 1$ are twofold:

1. τ^w will be slightly biased for small values of n.
2. This results in a definition of $d\tau^w$ which is exactly the same as the aforementioned Kemeny distance.

To reiterate: *Axiom 8* cannot be fully satisfied, therefore as a tradeoff we choose a value of β which satisfies *Axiom 8* when $n \to \infty$.

To tackle the first point: It is important to establish the exact amount of this bias this choice results in. Figure 3 shows τ^w as slightly positively biased, for very short rankings. In reality, rankings above the size of 4 will almost always be the ones considered, the bias for those should be very small and hence we believe this provides a strong base for fixing the value of β at 1.

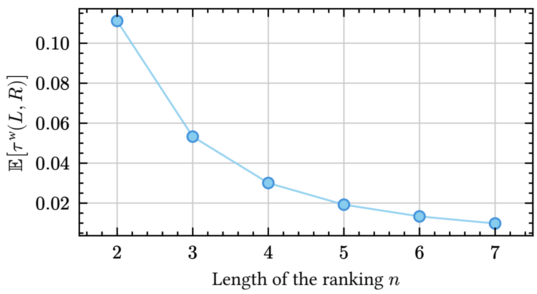

Figure 3: Bias of τ^w for small n.

In relation to the second point: The set of axioms chosen here, the ones which logically state how w should behave, can be mostly found equivalent to Kemenys' constraints. We however rejected the constraint of *betweenness*[8], which was not examined in this work. This in essence gave us the degree of freedom which is the choice of β. Instead we constrain β using the behaviours which are desired for similarity measures, not distance metrics (*Axiom 8*). Even with this key difference in treating ties, the surprising (but necessary) result is a definition which resembles exactly the result of Kemeny. That is: without considering the exact value of β but our choice of $\beta = 1$.

With this we conclude that given our axioms we are lead to the final definition of τ^w:

$$\tau^w(l,r) = \frac{\sum_{i<j} g^w_{l,r}(i,j)}{\frac{n(n-1)}{2}} \quad (25)$$

Where:

$$g^w_{l,r}(i,j) = \begin{cases} 1 & \text{if } \operatorname{sgn}(l_i - r_i) = \operatorname{sgn}(l_j - r_j) \\ -1 & \text{if } \operatorname{sgn}(l_i - r_i)\operatorname{sgn}(l_j - r_j) = -1 \\ 0 & \text{otherwise} \end{cases} \quad (26)$$

Which differs from τ^a only in regard that when the two elements are tied in both rankings $g^w_{l,r}(i,j)$ returns 1.

3.5 Weighted measures

Just as discussed in Section 2.2.3, the definition of the operator $\star_w$ for w is now straightforward, since it follows from $g^w_{l,r}(i,j)$.

$$l \star^w_w r = \sum_{i<j} g^w_{l,r}(i,j) w_{l,r}(i,j) \quad (27)$$

τ^w_h is therefore:

$$\tau^w_{h,w}(l,r) = \frac{l \star^w_w r}{\|l\|^w_w \cdot \|r\|^w_w} \quad (28)$$

$$= \frac{\sum_{i<j} g^w_{l,r}(i,j) w_{l,r}(i,j)}{\sqrt{\sum_{i<j} g^w_{l,l}(i,j) w_{l,l}(i,j)} \sqrt{\sum_{i<j} g^w_{r,r}(i,j) w_{r,r}(i,j)}} \quad (29)$$

$$= \frac{\sum_{i<j} g^w_{l,r}(i,j) w_{l,r}(i,j)}{\sqrt{\sum_{i<j} (w_{l,l}(i,j)) \cdot \sum_{i<j} w_{r,r}(i,j)}} \quad (30)$$

By $w_{l,r}(i,j) = \frac{1}{\max(l_i,l_j)-1}$ this leads[9] immediately to τ^w_{AP}:

$$\tau^w_{AP}(l,r) = \frac{\sum_{i<j} g^w_{l,r}(i,j) \frac{1}{\max(l_i,l_j)-1}}{\sqrt{\sum_{i<j} \left(\frac{1}{\max(l_i,l_j)-1} \right) \cdot \sum_{i<j} \frac{1}{\max(r_i,r_j)-1}}} \quad (31)$$

The choice of the weight function also influences the considerations of the previous section. The bias will change depending on the weight function since the weighting changes contributions of each pair of rankings to Equation 23. We cannot evaluate the bias for an arbitrary weight function, but Appendix C includes the results for τ^w_{AP} where the weight function is defined.

4 RESPONSIBLE RESEARCH

This work did not collect or process any data about people or groups of people, directly or indirectly. The definition of τ^w has no direct societal consequences, it is a proposal in a fully theoretical sense. We believe however that it can be helpful in the field of information retrieval and expand on this further in Section 5.3. This in turn could allow for more accurate ranking systems and, in the long term, benefit society.

[8]Part of *Condition 1* of [4].
[9]See also: Appendix D.

The results are not without flaws and we investigate their shortcomings further and offer a critical look in Section 5.

We want to reiterate the impact of Kemeny [4], this cannot be understated. Our work differs in one significant choice pertaining to the selection of core axioms of the distance measure. We still arrive at essentially the same result as Kemeny.

In order to ensure the reproducibility and complete transparency of our results, all code is open source and available. This guarantees that the considerations of Section 3.4.1 and Section 5 are verifiable. The code can be found in a version control repository at: `https://codeberg.org/mgazeel/tau_w`.

5 DISCUSSION

In this section our definition will be examined again, some toy examples will be provided, and most importantly we try to identify the shortcomings of our approach to the w-variant.

5.1 Negative correlation with ties

The resulting measures behave as required for a lot of cases due to our choice of axioms. For example rankings which are the same, or in other words perfectly positively correlated, yield a correlation of exactly 1 with all three measures. This is a direct result of *Axiom 6*.

$$\tau^w\begin{pmatrix}\langle A, B, C, D, E\rangle \\ \langle A, B, C, D, E\rangle\end{pmatrix} = 1 \qquad \tau^w\begin{pmatrix}\langle A, [B, C] D, E\rangle \\ \langle A, [B, C] D, E\rangle\end{pmatrix} = 1$$

Moreover, due to *Axiom 8*, for unrelated rankings or ones picked uniformly at random from all the possible rankings we are guaranteed to expect a value close to 0. However, none of the axioms consider the other end of the range: perfect negative correlation. For rankings without ties this works as intended[10]. But what if ties are considered?

$$\tau^w\begin{pmatrix}\langle A, B, C, D, E\rangle \\ \langle E, D, C, B, A\rangle\end{pmatrix} = -1 \qquad \tau^w\begin{pmatrix}\langle A, [B, C] D, E\rangle \\ \langle E, D, [B, C] A\rangle\end{pmatrix} = -0.8$$

But we would expect the result to be the −1 for both of these rankings with the w-variant! This is a problem which cannot be resolved by adding any axioms since τ^w is unique for our set of axioms. Furthermore, this problem cannot be resolved by a simple modification to the denominator D^w, even if we consider D^w to be dependent on the structure of ties (like in τ^b). We cannot reach both: −1 for the pair presented above while also keeping the result 1 for the identity comparison with ties. This is because D^w can only scale the numerator and the tie structure in both of these comparisons is the same.

Pursuant to the reasons above, this behaviour is accepted as a shortcoming of our definition. Its resolution is left to future work.

5.2 Comparison with a and b

As a point of comparison we utilize synthetic data generated with the simulator courtesy of Corsi & Urbano [2]. To observe a meaningful difference between w and the other variants a very high proportion of ties is required. This is a part of the definition, but the extent to which ties must be prevalent in the ranking may be too high.

An indirect result of *Axiom 7* is that the less ties ties are in the rankings the more similar τ^w becomes to τ.

Figure 4 and Figure 5 are runs of the two measures over rankings $n = 20$ using a high prevalence of ties (`frac_ties`[11] ≈ 0.95) and a low prevalence of ties (`frac_ties` ≈ 0.3).

In the comparison against τ^b we can observe, that for rankings where τ^b returns a highly uncorrelated judgment, τ^w tends to show more correlation. Furthermore this effect does not occur on the right-hand side of the graph. We believe this to be a direct result of the considerations in Section 5.1.

In both comparisons we see a large amount of points in the upper part where $\tau^a = 0$ or $\tau^b = 0$. This is a result of the fact that both of these measures return 0 if any of the rankings is completely tied. Our measure is the only one that can handle this case and attempt to perform a comparison. However, all the results for such a comparison return values higher than 0 which may be unwanted.

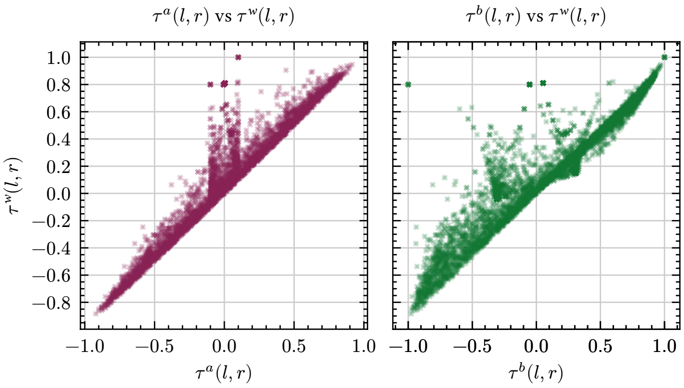

Figure 4: Comparison using synthetic data, ties high.

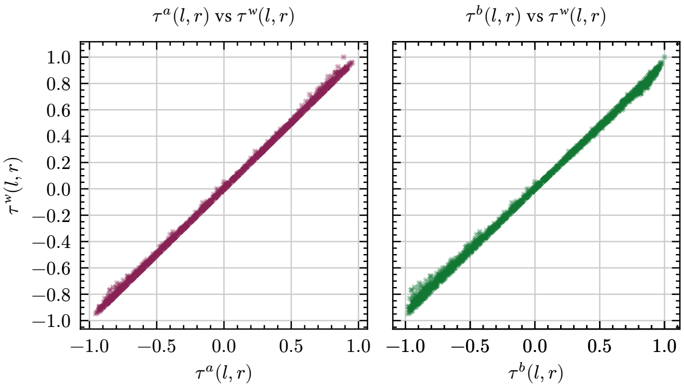

Figure 5: Comparison using synthetic data, ties low.

Finally, to showcase that a real difference between the three variants exists. We use real world data. In Figure 6 we provide the results of comparisons on data from TREC 2014 concerning Web Ad Hoc systems [13] comparing all combinations of topics.

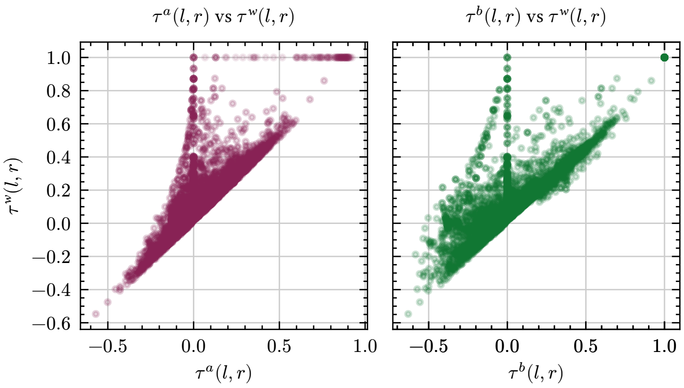

Figure 6: Comparison using data from TREC 2014 evaluation.

[10] because of *Axiom 7*.

[11] An argument to the generation code originally published in [2].

5.3 Usefulness in real world datasets

When evaluating the correlation between rankings of real retrieval systems, τ is used in a specific way. Voorhees et al. [14] provide a description of this approach and evaluate their results using it on the TREC-8 Ad Hoc data. Systems are ranked by their mean performance according to some score (for example $P@10$). Afterwards, τ (or τ^b) is used to calculate the similarity between mean rankings of systems - this allows us to measure the similarity of two scores. This similarity between the means is the number reported in works, but for the end user, it may not reflect the actual performance of a system on a topic. A more informative metric would be the correlation not between the mean rankings but rankings for that particular topic of interest. This, however, is too many values to report. Yet, possibly the mean of correlations of system rankings instead of the correlation of mean rankings would be a better indicator of performance.

The individual rankings of systems on a topic contain a high amount of ties[12]. In the case of metrics such as $P@10$, $P@20$ the ties between two systems in a topic essentially mean the exact same performance, not uncertainty in ranking them. This would require a correlation measure which treats ties not as uncertainty but occurrence at the same rank. This is exactly what τ^w is designed for.

Whereas the mean ranking of system does not contain a high incidence of ties. This makes the choice of variant have much less impact on the result.

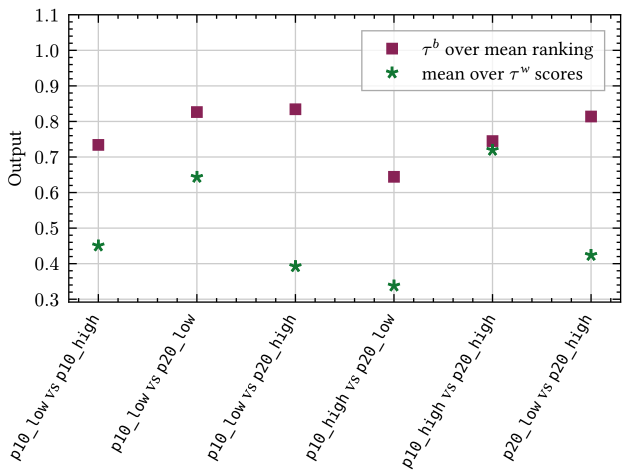

Figure 7: Comparison of computed correlation for TREC 2014.

Figure 7 shows the two aforementioned evaluation methods between different scores. The key insight this provides is that on real world data, such as the results of TREC, calculating the mean of correlations over topic rankings can return lower scores. This is paramount. It points at the possibility that calculating correlation over the mean scores may be **overestimating the actual performance** for end users.

6 CONCLUSION

The premise of the w variant was introduced. A distance based interpretation of this premise was shown, and thereafter we defined logical axioms about the desired behaviour of τ^w.

We have shown that our definition of $d\tau^w$ upon these axioms is unique, with the exception of *Axiom 8* which was approached empirically. If a different definition were to be considered, the choice of these axioms must be challenged, changed, and rationale for this must be provided.

[12]In the context of the TREC 2014 dataset.

Finally, we also arrive at τ^w_{AP} and τ^w_h using the same reasoning that was used by Vigna [12] to define τ_h originally.

The relationship between these definitions and the widely used variants a and b has been investigated on real and synthetic data. Furthermore, we provide results which show that our method may be useful in evaluating system performance in information retrieval.

We conclude by saying that our final results can be found in Equation 25 in the case of τ^w, Equation 31 for τ^w_{AP}, Equation 30 for τ^w_h.

6.1 Future work

The most important problems to consider are the findings of Section 5.1. When negatively correlated rankings with ties are compared, the current formulation of D^w is not able to scale the numerator to -1. From the discussion it is clear the a simple reformulation of D^w is not sufficient. One way of addressing this problem is by both 'offsetting' the numerator and modifying the denominator. This would however violate the premises of *Axiom 8* which is a central part of τ^w. We believe that this problem may be addressed by 'splitting' the scaling of the numerator. That is: introducing a different D^w dependent on the result of N^w being negative or positive. These could be coined D^w_- and D^w_+ and have to be carefully chosen based on the rankings being compared. This way only *Axiom 6* has to modified. This needs much further investigation. Sadly, due to the time constraints imposed on this thesis we were unable to conduct it.

Another area worth expanding are the findings of Section 3.4. Especially, it may be possible to find a rigorous proof that the value of β does indeed converge to 1. If that fails, we have also considered uniformly sampling from $\hat{R}_n$. A preliminary investigation points to this being a hard problem, but if an efficient algorithm could be found for it, it may be possible to probabilistically calculate the values for much higher n.

Another question which is left unanswered is the w variant in the context of different similarity measures. This concerns in particular Spearman [10], since it is another widely used measure.

Finally, as discussed, the choice of axioms can also be challenged, or possibly the approach of basing the definition on that of RBO^w which was rejected in this work could be investigated. This may lead to different results as to the definition.

Bibliography

[1] Ben Carterette, Virgil Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million Query Track 2009 Overview. In *TREC*, 2009.

[2] Matteo Corsi and Julián Urbano. 2024. The Treatment of Ties in Rank-Biased Overlap. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2024. Washington DC, USA, 251–260. https://doi.org/10.1145/3626772.3657700

[3] Felix Hausdorff. 1914. Grundzuge der Mengenlehre.

[4] John George Kemeny. 1959. Mathematics without Numbers. *Daedalus* 88, 4 (1959), 577–591. Retrieved May 23, 2025 from https://www.jstor.org/stable/20026529

[5] Maurice George Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (June 1938), 81–93. https://doi.org/10.2307/2332226

[6] Maurice George Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, (November 1945), 239–251. https://doi.org/10.2307/2332303

[7] Maurice George Kendall. 1948. *Rank Correlation Methods.* C. Griffin.

[8] Karl Pearson. 1907. *On further methods of determining correlation.* Cambridge University Press, London, United Kingdom.

[9] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2010. Geneva, Switzerland, 555–562. https://doi.org/10.1145/1835449.1835542

[10] Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (January 1904), 72–101. https://doi.org/10.2307/1412159

[11] Julián Urbano and Mónica Marrero. 2017. The Treatment of Ties in AP Correlation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 2017. Amsterdam, The Netherlands, 321–324. https://doi.org/10.1145/3121050.3121106

[12] Sebastiano Vigna. 2015. A Weighted Correlation Index for Rankings with Ties. In *Proceedings of the 24th International Conference on World Wide Web*, May 2015. Florence, Italy, 1166–1176. https://doi.org/10.1145/2736277.2741088

[13] Ellen M. Voorhees and Angela Ellis (Eds.). 2014. *Proceedings of The Twenty-Third Text REtrieval Conference.* National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

[14] Ellen M. Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models?. Retrieved June 20, 2025 from https://arxiv.org/abs/2201.11086

[15] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28, 4 (2010). https://doi.org/10.1145/3626772.3657700

[16] Max Atkin Woodbury. 1940. Rank Correlation when There are Equal Variates. *The Annals of Mathematical Statistics* 11, 3 (September 1940), 358–362. Retrieved May 23, 2025 from https://www.jstor.org/stable/2235684

[17] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2008. Singapore, Singapore, 587–594. https://doi.org/10.1145/1390334.1390435

[18] "Student". 1921. An Experimental Determination of the Probable Error of Dr Spearman's Correlation Coefficients. *Biometrika* 13, 2/3 (June 1921), 263–282. https://doi.org/10.2307/2331754

APPENDICES

A CASE ANALYSIS

When proving *Axiom 1.4*, we will show that for any configuration of elements:

$$\forall_{x,y,z} \quad d^{w}_{x,z}(i,j) \leq d^{w}_{x,y}(i,j) + d^{w}_{y,z}(i,j)$$

Below we include all possible configurations of a pair between three rankings of size 2, as well as the concrete values of the above inequality in each case. Symmetric configurations are not considered.

$$\begin{array}{c|c|c} x & y & z \\ A & A & A \\ B & B & B \end{array} \qquad 0 \leq 0 + 0$$

$$\begin{array}{c|c|c} x & y & z \\ A & B & A \\ B & A & B \end{array} \qquad 0 \leq 2\alpha$$

$$\begin{array}{c|c|c} x & y & z \\ A & \overline{A} & A \\ B & \underline{B} & B \end{array} \qquad 0 \leq 2\beta$$

$$\begin{array}{c|c|c} x & y & z \\ A & A & B \\ B & B & A \end{array} \qquad 0 \leq 0$$

$$\begin{array}{c|c|c} x & y & z \\ A & B & B \\ B & A & A \end{array} \qquad 0 \leq 0$$

$$\begin{array}{c|c|c} x & y & z \\ A & \overline{A} & B \\ B & \underline{B} & A \end{array} \qquad \alpha \leq 2\beta$$

$$\begin{array}{c|c|c} x & y & z \\ A & A & \overline{A} \\ B & B & \underline{B} \end{array} \qquad 0 \leq 0$$

$$\begin{array}{c|c|c} x & y & z \\ A & B & \overline{A} \\ B & A & \underline{B} \end{array} \qquad 0 \leq \alpha$$

$$\begin{array}{c|c|c} x & y & z \\ A & \overline{A} & \overline{A} \\ B & \underline{B} & \underline{B} \end{array} \qquad 0 \leq 0$$

B PREVALENCE OF PAIRS OF RELATIONS

Here we provide an exhaustive list of the prevalences found in all rankings, with ties, of a given size.

$l \backslash r$	<	=	>
<	1	1	1
=	1	1	1
>	1	1	1

Table 6: Prevalence for $\hat{R}_2$

$l \backslash r$	<	=	>
<	75	45	75
=	45	27	45
>	75	45	75

Table 7: Prevalence for $\hat{R}_3$

$l \backslash r$	<	=	>
<	5766	2418	5766
=	2418	1014	2418
>	5766	2418	5766

Table 8: Prevalence for $\hat{R}_4$

$l \backslash r$	<	=	>
<	542890	174750	542890
=	174750	56250	174750
>	542890	174750	542890

Table 9: Prevalence for $\hat{R}_5$

$l \backslash r$	<	=	>
<	64335615	16806165	64335615
=	16806165	4390215	16806165
>	64335615	16806165	64335615

Table 10: Prevalence for $\hat{R}_6$

$l \backslash r$	<	=	>
<	9531963525	2095197615	9531963525
=	2095197615	460540269	2095197615
>	9531963525	2095197615	9531963525

Table 11: Prevalence for $\hat{R}_7$

C THE BIAS OF τ^w_{AP} AND HYPERBOLIC τ^w_h

We also investigate the bias of $\tau^w_{h,w}$ when the weighting function is an additive hyperbolic function as defined by Vigna [12]. We will refer to this as $\tau^w_{h,h}$.

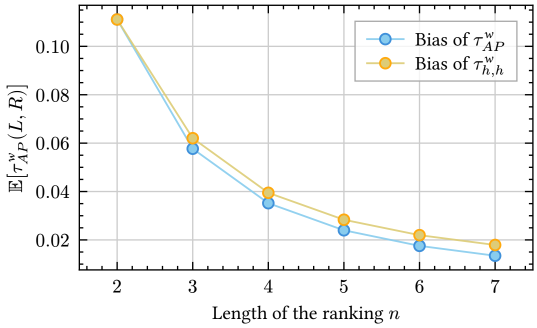

Figure 8: Bias for small n.

n	$\mathbb{E}[\tau^w(L,R)]$	$\mathbb{E}[\tau^w_{AP}(L,R)]$	$\mathbb{E}[\tau^w_{h,h}(L,R)]$
2	$0.\overline{1}$	$0.\overline{1}$	$0.\overline{1}$
3	0.0533	0.0577	0.0620
4	0.0300	0.0352	0.0395
5	0.0192	0.0240	0.0283
6	0.0133	0.0175	0.0220
7	0.0098	0.0135	0.0179

Table 12: Comparison of approximate bias between τ^w, τ^w_{AP}, $\tau^w_{h,h}$.

D NOTES ABOUT THE DENOMINATOR OF τ^w_{AP}

The denominator in the definition of τ^w_{AP} (Equation 31) can be simplified in the absence of ties. To recall:

$$D^w_{AP} = \sqrt{\sum_{i<j}\left(\frac{1}{\max(l_i,l_j)-1}\right)\cdot\sum_{i<j}\frac{1}{\max(r_i,r_j)-1}}$$

If we consider rankings without ties, each rank in r and l appears exactly once meaning that in each of the sums we have the pairs:

Pair	Contribution
$(l_i = 2, l_j = 1)$ or vice-versa	$\frac{1}{\max(2,1)-1} = \frac{1}{1}$
Total contribution of $l_i = 2$	1
$(l_i = 3, l_j = 1)$ or vice-versa	$\frac{1}{\max(3,1)-1} = \frac{1}{2}$
$(l_i = 3, l_j = 2)$ or vice-versa	$\frac{1}{\max(3,2)-1} = \frac{1}{2}$
Total contribution of $l_i = 3$	1
$\vdots$	$\vdots$

The total value of the sum will therefore be:

$$D^w_{AP} = \sqrt{(n-1)(n-1)} = (n-1)$$

This appendix is included for the sake of completeness, and the above analysis breaks down for rankings which do contain ties since the contribution changes based on the rankings of tied items.