The Treatment of Ties in Rank-Biased Overlap

Corsi, Matteo; Urbano, Julián

DOI
10.1145/3626772.3657700

Publication date
2024

Document Version
Final published version

Published in
SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval

The Treatment of Ties in Rank-Biased Overlap

Matteo Corsi
Delft University of Technology
Delft, The Netherlands
m.corsi@tudelft.nl

Julián Urbano
Delft University of Technology
Delft, The Netherlands
j.urbano@tudelft.nl

ABSTRACT

Rank-Biased Overlap (*RBO*) is a similarity measure for indefinite rankings: it is top-weighted, and can be computed when only a prefix of the rankings is known or when they have only some items in common. It is widely used for instance to analyze differences between search engines by comparing the rankings of documents they retrieve for the same queries. In these situations, though, it is very frequent to find tied documents that have the same score. Unfortunately, the treatment of ties in *RBO* remains superficial and incomplete, in the sense that it is not clear how to calculate it from the ranking prefixes only. In addition, the existing way of dealing with ties is very different from the one traditionally followed in the field of Statistics, most notably found in rank correlation coefficients such as Kendall's and Spearman's. In this paper we propose a generalized formulation for *RBO* to handle ties, thanks to which we complete the original definitions by showing how to perform prefix evaluation. We also use it to fully develop two variants that align with the ones found in the Statistics literature: one when there is a reference ranking to compare to, and one when there is not. Overall, these three variants provide researchers with flexibility when comparing rankings with *RBO*, by clearly determining what ties mean, and how they should be treated. Finally, using both synthetic and TREC data, we demonstrate the use of these new tie-aware *RBO* measures. We show that the scores may differ substantially from the original tie-unaware *RBO* measure, where ties had to be broken at random or by arbitrary criteria such as by document ID. Overall, these results evidence the need for a proper account of ties in rank similarity measures such as *RBO*.

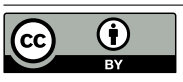
CCS CONCEPTS

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

KEYWORDS

Rank correlation, rank similarity, rank-biased overlap, ties

ACM Reference Format:
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 (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657700



SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657700

1 INTRODUCTION

Rankings are part of our everyday lives: music albums are ranked by sales, universities are ranked by research output, cities are ranked by livability, etc. In Information Retrieval (IR) and Recommender Systems (RecSys), rankings are essential: search engines rank documents by likelihood of relevance to a query, and recommenders rank for instance books by likelihood of purchase. But rankings can be made following alternative criteria, such as music albums by replays, universities by alumni success, cities by pollution, etc. One way to understand the differences and commonalities of ranking criteria is to compare the rankings they produce. In IR this happens when we compare the rankings of documents returned by different systems, topics that documents are estimated to fit to, terms for query expansion, or rankings of systems sorted by different evaluation metrics or different relevance assessors.

Comparing rankings requires a rank similarity measure. Some of the most well-known examples are τ by Kendall [12], ρ by Spearman [27], and D by Kolmogorov [15], while others developed in the context of IR and related disciplines are τ_* by Melucci [21], d_{rank} by Carterette [9], τ_{ap} by Yilmaz et al. [37], K^* and F'^* by Kumar and Vassilvitskii [16], and τ_w by Vigna [33]. While some of these are top-weighted and thus assign more importance to similarities at the top of rankings than at the bottom, none of them can compare non-conjoint rankings that have only some items in common. This is very often the case in IR and RecSys when comparing the results from search engines that have different indexes, recommenders that have different catalogs, or simply cases where the rankings are truncated after a certain depth. The problem of rank similarity under non-conjointness has received far less attention, with works inspired by Spearman's footrule [3, 11], the Hoeffding distance [30], and even IR metrics [5, 31]. Most notably, Webber et al. [35] proposed Rank-Biased Overlap (*RBO*), which has become popular in IR research for example to compare search engine results [8, 23], measure topic similarity [1, 19], assess consistency of systems to query variations [2], or compare rankings of documents in general [10, 22, 26, 38]. Beyond IR, it is also used for example in RecSys [7, 34], Network Science [18, 25] and Neuroscience [4, 28].

RBO is top-weighted, and it handles non-conjointness as well as incomplete rankings, even of different lengths. Incomplete rankings appear for instance after truncation, because their very top-weighted nature implies that, after a sufficiently deep rank, what items appear next is negligible. For example, a search engine may return only the top 20 documents in response to a query, a recommender may suggest only the top 5 items, and a typical TREC run consists of only the top 1,000 documents per topic. This means that rankings may actually consist of a *seen* part or prefix, and an *unseen* part that extends further, potentially up to infinity. Therefore, *RBO* scores have to be computed from a prefix only, ideally accompanied by some quantification of the uncertainty due to the unseen items.

Table 1: Summary statistics of the adhoc runs in the last editions of TREC Web. Fifty topics were used in all cases.

Year	Groups	Runs	Runs with ties	Rankings with ties	Docs tied	Avg. tie group size
2009	25	71	90%	88%	22%	19.3
2010	21	56	96%	87%	30%	17.7
2011	14	37	89%	79%	27%	3.6
2012	11	27	63%	63%	20%	2.8
2013	14	34	62%	57%	35%	26.6
2014	10	30	60%	60%	24%	5.1
Avg.	16	43	77%	72%	26%	12.5

Webber et al. [35] showed how to compute upper and lower bounds for this uncertainty, as well as a point estimate, called RBO_{EXT}. This is the score typically reported when using *RBO*.

1.1 The Problem of Tied Items

Rankings may very well contain tied items. For example, systems with the same *P*@10 score, or documents with the same retrieval score for a query. The latter is a well-known issue in IR research. For instance, Table 1 presents summary statistics from all the adhoc runs in TREC Web between 2009 and 2014, showing that 77% of the runs contained ties in 72% of the rankings, with 26% of the documents tied in groups of 12.5 documents on average. This begs the question: how should these documents be ranked when evaluating effectiveness? Different approaches to this problem, as well as their impact, have been studied for example by Cabanac et al. [6], Lin and Yang [17], McSherry and Najork [20], Raghavan et al. [24].

A similar question may be asked about rank similarity measures such as *RBO*: how should tied items be handled? The first approach is to break the ties, essentially ignoring them. For example, ties may be broken at random, but this would introduce noise. Other, rather arbitrary criteria may be used to break ties, such as by document ID; this is the approach implemented in `trec_eval`, but it has its own issues [6, 17]. For *RBO* specifically, it inflates similarity scores because tied documents would artificially appear in the same order in both rankings. We illustrate this in Figure 1 with the *RBO* scores computed between pairs of the TREC Web runs summarized in Table 1, as well as synthetic rankings (details are presented in Section 5). As expected, these tie-breaking strategies generally lead to different *RBO* scores, but some differences are strikingly large, specially in the TREC data. Differences are larger than normal reporting fidelity (i.e. 2 decimal digits) in 10% of the TREC cases, and 66% of the synthetic ones. In addition, it is clear that breaking ties by document ID does inflate scores.

The second approach is to explicitly handle the ties, in a principled way, in *RBO* itself. Webber et al. [35] followed this line to motivate a tie-aware variant by assuming that all items tied between ranks n and m occur at the same rank n. In essence, they assumed that tied items *really* occur at the same rank. This contrasts with the view typically taken in the Statistics literature, where a tie represents *uncertainty*. Specifically, an integral ranking is assumed to exist, with a strict order between all items, and a tie really represents that it is not known, for whatever reason, which item goes first. This happens when their underlying scores are the same within reporting fidelity, or when they represent observations that

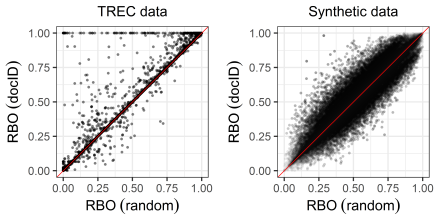


Figure 1: Differences in $RBO(p = 0.9)$ when breaking ties by doc ID or at random, for TREC and synthetic data

are too noisy to discern the actual order. In essence, a tie here represents a loss of information. This notion of ties for rank similarity dates back to more than a century ago, to the work of Student [29] for adapting Spearman's ρ to handle ties. It was later popularized, most notably by Kendall [13, 14], when similarly adapting his τ coefficient, which resulted in two tie-aware variants: τ_a and τ_b.

The a-variant was proposed for cases where one ranking represents a reference and the goal is to compute the *accuracy* of the other ranking with respect to this reference. To this end, τ_a was precisely defined as the expected value of τ when breaking ties at random. The b-variant was proposed for cases where there is no reference and one wants to calculate the *agreement* between the two rankings. To this end, τ_b corrects the measured similarity by the amount of information lost due to ties. Vigna [33] followed the second approach to define his τ_w coefficient, albeit without explicit mention of it. More recently, Urbano and Marrero [32] followed both approaches to define tie-aware variants of Yilmaz et al.'s τ_{ap}.

1.2 Contributions

Webber et al. [35] tackled the issue of ties in *RBO* only in passing: as will be detailed in Sections 3.2 and 4, it is unclear how to calculate RBO_{EXT} and its bounds when dealing with ties. This is very well illustrated by one of the popular implementations available online,[1] where authors argue that the equations in [35] need modifications in order to handle ties, but it turns out that these modifications make the results incorrect when rankings do not have ties.

In this paper we deal with the problem of explicitly treating ties in Rank-Biased Overlap. Specifically:

(1) We show how to compute both a- and b-variants of *RBO*.
(2) We develop a formulation for *RBO* that generalizes all three variants and allows us to derive the point estimate RBO_{EXT} and the bounds (i.e., RBO_{MIN}, RBO_{MAX} and thus RBO_{RES}).
(3) Using both TREC data and synthetic data we illustrate the differences among variants, as well as the importance of following a principled approach to deal with ties, as opposed to arbitrarily breaking them and computing bare *RBO*.
(4) We provide a full implementation of all coefficients,[2] as well as guidelines for when to use each (see Section 6).

In summary, we contribute the theoretical underpinnings for a principled treatment of ties in *RBO*, providing complete formulations and implementations of three variants that align with different notions of ties, namely RBO^w, RBO^a and RBO^b.

[1] https://github.com/dlukes/rbo
[2] https://github.com/julian-urbano/sigir2024-rbo

Table 2: Summary of notation. See Figure 2 for examples.

S, L	Rankings of lengths s and l, where $s \leq l$.
S_d, S_e^{-1}	Item at rank d and rank of item e, in S.
$S_{n:m}$	Set of items from rank n to m in S.
$\Omega = \{S \cup L\}$	Set of all items seen in S or L.
d	Evaluation depth for computing agreement.
X_d, A_d	Overlap and agreement at depth d.
p	Persistence parameter of *RBO*.
$t_{e,S}$, $b_{e,S}$	Top and bottom ranks of e's tie group in S.
$c_{e,S\|d}$	Contribution of item e in S given depth d.
Inactive item	One that is surely below d (i.e. $d < t_e$).
Active item	One that is surely above d (i.e. $b_e \leq d$).
Crossing group	One that is crossed by d (i.e. $t_e \leq d < b_e$).

2 RANK-BIASED OVERLAP

Throughout the paper we will use the notation in Table 2 and the example in Figure 2. In particular, let S and L be two indefinite rankings of lengths s and l, where S is generally shorter than L. Webber et al. [35] defined their *overlap*, up to a depth d, as the number of items that are in common between the two rankings:

$$X_{S,L,d} = |S_{:d} \cap L_{:d}| \ , \tag{1}$$

where $S_{:d} = \{e \ : \ S_e^{-1} \leq d\}$ represents the set of items in S that are ranked at or above the evaluation depth d; throughout the paper we refer to these items as *active* in S given d. The proportion of active items that overlap is called the *agreement*:

$$A_{S,L,d} = \frac{X_{S,L,d}}{d} \ . \tag{2}$$

In the example from Figure 2, only item a overlaps at depth 3 and thus $A_3 = 1/3$. The Rank-Biased Overlap is then defined as the infinite and weighted sum of the agreements at all depths:

$$RBO_{S,L,p} = \frac{1-p}{p} \sum_{d=1}^{\infty} A_{S,L,d} \cdot p^d \ , \tag{3}$$

where p^d is the weight given to the agreement at depth d, and the $(1-p)/p$ term ensures that *RBO* is bounded in the range $[0, 1]$. Fully disjoint rankings result in $RBO = 0$, while identical rankings result in $RBO = 1$. This is because agreement would be 0 and 1, respectively, at all depths.

The parameter p is called *persistence*, and it determines how steep the decline in weights is: a small p places a very high weight at the top of the ranking compared to the bottom, while a large p flattens the decay so that the weight of deep items is not as small compared to those at the top. For a full account on the properties of *RBO*, such as metricity, the reader is referred to [35].

3 TIES IN RANK-BIASED OVERLAP

As discussed in Section 1.1, the original view on ties by Webber et al. [35] is that tied items occur at the same rank, while the view traditionally taken in the Statistics literature is that a tie represents uncertainty as to which item goes first, that is, a loss of information. In the subsequent subsections we first describe Webber et al.'s approach, which we call w-variant, and then fully develop both the a- and b-variants.

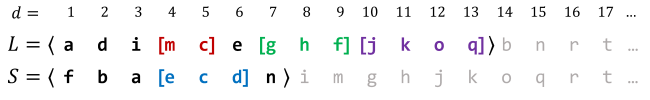


Figure 2: Main example used throughout the paper. Colored letters represent tied items, and square brackets represent tie groups. The grayed-out sections illustrate the arrangement of items that would maximize RBO_{MAX}. Exemplifying notation in Table 2: $l = 13, s = 7, L_6 = \mathsf{e}, S_{\mathsf{b}}^{-1} = 2, t_{\mathsf{g},L} = t_{\mathsf{h},L} = t_{\mathsf{f},L} = 7$, and $b_{\mathsf{e},S} = b_{\mathsf{c},S} = b_{\mathsf{d},S} = 6$.

3.1 w-variant: RBO^w

Webber et al. [35] assumed that all items tied in a group occur at the same rank, namely the top rank of the group. Rankings where tied items are assigned ranks in this way are typically known as *sports* rankings.[3] In essence, they assumed that items that are tied *really* occur at the same rank.

From this point of view, when considering a group crossed by the evaluation depth d, all its items will be active because they are assumed to occur at the top of the group, that is, $S_e^{-1} = t_{e,S}$. Thus, under the sports ranking assumption, $S_{:d}$ may contain more than d items, in particular all items that are either above the depth or tied with it. Webber et al. then modified the definition of agreement as:

$$A^w_{S,L,d} = \frac{2 \cdot X_{S,L,d}}{|S_{:d}| + |L_{:d}|} \ . \tag{4}$$

In the example from Figure 2, $S_{:5}$ includes item d because it belongs to a crossing group, which happens to increase overlap because it is also found in $L_{:5}$. The denominator is thus 11.

This new definition of agreement may just be plugged into Eq. (3) to calculate a tie-aware variant of *RBO* under the assumption of sports rankings. We will refer to this variant as RBO^w. Note that in the absence of ties A^w reduces to A, so $RBO^w = RBO$ as expected.

At this point we hint at the gap in the work of Webber et al. [35]. Indeed, A^w can be directly used for computing the full *RBO* score on infinite rankings, but it remains unclear how to use it when computing *RBO* from prefix evaluation (i.e. how to compute RBO^w_{EXT} and its bounds). This is because the relevant equations in their work, as well as their rationale, are always expressed in terms of overlap divided by a constant equal to the evaluation depth d, not in terms of agreement. The tie-aware A^w in Eq. (4) has a different functional form altogether, so one can not simply use their equations to compute RBO^w. This will become more evident next, when we introduce the a- and b-variants, as well as in Section 4 when we actually deal with prefix evaluation.

3.2 a-variant: RBO^a

For the problem of rank correlation, the first way of handling ties under the semantics of uncertainty asked this question: what is the average correlation across all possible permutations of the ties? This interpretation was followed by Woodbury [36] to define a tie-aware variant of Spearman's ρ, later by Kendall [13] to define his τ_a, and recently by Urbano and Marrero [32] to formulate a tie-aware variant of Yilmaz et al.'s τ_{ap}, namely $\tau_{ap,a}$.

[3] For example, two athletes tied at position 2 share the same rank 2. The next athlete would be at position 4, so that rank 3 is simply unassigned.

In the same spirit, we ask this question: what is the average agreement among all possible permutations of the ties? This is:

$$A^a_{S,L,d} = \frac{1}{|\mathcal{P}_S||\mathcal{P}_L|} \sum_{\tilde{S}\in\mathcal{P}_S} \sum_{\tilde{L}\in\mathcal{P}_L} \frac{X_{\tilde{S},\tilde{L},d}}{d} , \quad (5)$$

where $\tilde{S}$ refers to a single permutation of the tied items in S, and $\mathcal{P}_S$ refers to all possible such permuted rankings.[4] In our example, the number of permutations are 288 and 6 for L and S, respectively. Note that the overlap is still defined as in Eq. (1), only that it is computed between permuted rankings instead of the originals.

Computing Eq. (5) by enumerating all possible permutations would be extremely expensive because the number of permutations grows factorially with the number of ties. In order to derive a simple expression that does not require enumeration, we begin by reformulating overlap as follows: instead of calculating the size of the intersection between the items active at depth d, we will count the number of items that are active in both rankings:

$$\begin{aligned} X_{S,L,d} &= \sum_{e\in\Omega} \mathbb{I}\left[\{S_e^{-1} \le d \ \wedge \ L_e^{-1} \le d\}\right] \\ &= \sum_{e\in\Omega} \mathbb{I}\left[\{S_e^{-1} \le d\}\right] \cdot \mathbb{I}\left[\{L_e^{-1} \le d\}\right] , \end{aligned} \quad (6)$$

where $\Omega = \{S \cup L\}$ is the set of all items, and $\mathbb{I}$ is the Iverson bracket (i.e. $\mathbb{I}[P]=1$ if P is true, 0 otherwise). As expected, an item will contribute to overlap only if it appears at or above d in both rankings. In our example, the summand for item b is $1\cdot 0$ at $d=3$ (i.e. does not overlap), whereas for item a it is $1\cdot 1$ (i.e. it does overlap).

We can now plug the overlap between two permutations, as defined in Eq. (6), into the agreement averaged across permutations in Eq. (5). After minor rearranging, we obtain:

$$A^a_{S,L,d} = \frac{1}{d} \sum_{e\in\Omega} \sum_{\tilde{S}\in\mathcal{P}_S} \frac{\mathbb{I}\left[\{\tilde{S}_e^{-1} \le d\}\right]}{|\mathcal{P}_S|} \sum_{\tilde{L}\in\mathcal{P}_L} \frac{\mathbb{I}\left[\{\tilde{L}_e^{-1} \le d\}\right]}{|\mathcal{P}_L|} . \quad (7)$$

Note that the last two summations represent, respectively, the fraction of permutations of S and L where the item e is active. There are three possibilities for an arbitrary item and ranking:

(1) Inactive: an item or group that is entirely *below* the evaluation depth d (i.e. $d < t_e$) will remain below in all permutations and will never contribute to overlap. In our example, items e, c, d and n are inactive in S at $d=3$.
(2) Active: an item or group that is entirely *at or above* the evaluation depth d (i.e. $b_e \le d$) will remain above in all permutations and will always be able to contribute to overlap. In our example, items a, d, i, m and c are active in L at $d=5$.
(3) Crossing: an item within a crossing group (i.e. $t_e \le d < b_e$) will be able to contribute to overlap in as many permutations as it is placed at or above d. The item will appear in position t_e a total of $(b_e - t_e)!$ times, at position t_e+1 another $(b_e - t_e)!$ times, and so on. Only $(d - t_e + 1)$ positions will make the item active, which happens in $(d - t_e + 1)\cdot(b_e - t_e)!$ permutations. Because there are a total of $(b_e - t_e + 1)!$ permutations of the group, the item's total contribution to overlap is $(d - t_e + 1)\,/\,(b_e - t_e + 1)$. In our example, item e in S appears at each of ranks 4, 5 and 6 in 1/3 of the permutations. At depth 5 there are two slots available for the items in the crossing group (i.e. 4 and 5), so each of its items will have a contribution of 2/3.

At a given depth d, we can therefore define the total contribution of an item e across permutations as:

$$c_{e|d} = \begin{cases} 0 & d < t_e \text{ (inactive)} \\ 1 & b_e \le d \text{ (active)} \\ \frac{d-t_e+1}{b_e-t_e+1} & \text{otherwise (crossing)} \end{cases} . \quad (8)$$

In doing so, we are implicitly assuming that the first unseen item is not tied with the last one seen, as otherwise it should have been seen in the prefix. In the example, b is assumed to not be tied with the last tie group in L, so that $c_{L|d}$ can be computed from the prefix only. The contribution of item c in S would be 0 for depths 1 to 3, 1/3 and 2/3 for depths 4 and 5, and 1 for depths 6 and beyond.

Back in Eq. (7), we can now replace each of the last two summations with the corresponding contributions of e in S and L, avoiding the need for enumerating all permutations. All together, a simple and efficient formulation of the agreement for RBO^a is:

$$A^a_{S,L,d} = \frac{1}{d} \sum_{e\in\Omega} c_{e,S|d} \cdot c_{e,L|d} . \quad (9)$$

In the absence of ties, note that only the first two cases apply in Eq. (8), which means that A^a reduces to A because Eq. (6) reduces to Eq. (1), so that $RBO^a = RBO$. Also note that, by construction, RBO^a gives the right answer to the naive approach of computing bare RBO after breaking ties at random, eliminating unnecessary noise. In addition, note that Eq. (9) can be computed efficiently because the summation does not need to enumerate all items in Ω, but only those that are active or crossing in both rankings, that is, those where $c_{e,S|d} > 0$ and $c_{e,L|d} > 0$.

3.3 *b*-variant: RBO^b

Still under the semantics of uncertainty, the principle for this variant is that it should account for the amount of information actually available to measure overlap. Because a tie represents uncertainty with respect to the actual rank of items, they can not contribute fully to the measured overlap. This inability to fully contribute should be reflected in the normalization term (i.e. the denominator in the agreement function). This leads to the idea of *measurable* overlap in a ranking, or "untiedness" as called by Vigna [33]. Whilst the a-variant always expects a full measurable overlap of d regardless of the presence of ties, the b-variant should not.

This principle to handle ties was first followed by Student [29] to propose a b-variant of Spearman's ρ, later by Kendall [13] to define his τ_b, and recently by Vigna [33] and Urbano and Marrero [32]. In order to define a b-variant of the agreement for RBO, we thus draw inspiration from Kendall's τ_b which, for two arbitrary (conjoint) rankings U and V is:

$$\tau_b(U,V) = \frac{\sum_{i<j} \text{sign}(u_j - u_i)\cdot\text{sign}(v_j - v_i)}{\sqrt{\sum_{i<j}\text{sign}(u_j-u_i)^2}\sqrt{\sum_{i<j}\text{sign}(v_j-v_i)^2}} , \quad (10)$$

where the numerator quantifies *actual*, observed concordance between the rankings, and the denominator quantifies their *measurable* concordance. From this equation we can recognize how an item pair (i, j) affects τ_b when it is tied: because the tie represents

[4] For instance, the possible permutations of ranking ⟨A [B C][D E] F⟩ are ⟨A B C D E F⟩, ⟨A B C E D F⟩, ⟨A C B D E F⟩ and ⟨A C B E D F⟩.

$P = \langle$ [a b c] d $\rangle$
$Q = \langle$ [a b] c d $\rangle$
$R = \langle$ a [b c d e f …] $\rangle$
$T = \langle$ z [b c d e f …] $\rangle$
$W = \langle$ a [2 3 4 5 6 …] $\rangle$

Figure 3: Sample rankings that would lead to unexpected *RBO* scores with strops rankings. Colored letters represent tied items, and square brackets represent tie groups.

uncertainty, it does not contribute to the numerator in a positive or negative direction. As for the denominator:

(1) If the item pair is tied in both rankings, it does not contribute to the denominator either. In this case, the item pair is essentially ignored because it does not bear any information about measurable concordance. As such, τ_b *can still* be 1 if both rankings tie exactly the same items.
(2) If the item pair is tied in only one ranking, it will still contribute to the denominator on behalf of the other ranking. In this case, the item pair is not completely ignored because it still contributes to the measurable concordance. As such, τ_b *can not* reach 1 any more.

Applying the same rationale to *RBO*'s agreement, we recognize that tied items should be inactive until the evaluation depth reaches the bottom rank of the group. This is because, until then, their actual ranks are unknown and it should therefore not be possible for them to contribute to overlap at earlier depths. In our example, at depth 5 we can not know which items are actually at ranks 4 and 5 in S; it could be any two of e, c and d. Thus, only items f, b and a can contribute to overlap. In contrast, both m and c can contribute in L because their group is entirely active already at depth 5.

Items in a crossing group should therefore not contribute to the numerator. As for the denominator:

(1) If both rankings have crossing groups at the same ranks, the amount of untiedness is the same in both, say n (in the example, $n = 3$ at depth 4). Therefore, the measurable overlap is at most n. As such, agreement *can still* reach 1 if both rankings tie exactly the same ranks, regardless of which items they tie.
(2) If the rankings do not have crossing groups at the same ranks, the amount of active items they contribute to the measurable overlap is different (in the example, 6 from L and 7 from S at depth 7). As such, agreement *can not* reach 1 any more.

Dealing with ties in this way, we can recognize an approach somewhat opposite to that of A^w. Indeed, while the w-variant assigns to tied items the top rank of their group (i.e. $S_e^{-1} = t_{e,S}$), here we assign them the bottom rank instead (i.e. $S_e^{-1} = b_{e,S}$). By analogy, let us refer to the resulting ranking with the anadrome "strops" ranking, to clearly reflect the reverse of "sports" ranking.

Unfortunately, a naive application of the strops ranking would lead to unexpected *RBO* results, as illustrated in Figure 3:

- R vs W: the rankings only have the top item in common, so we should intuitively expect an *RBO* score close to 0. However, agreement is 1 at *every* depth except at the end, where it becomes nearly 0 because all items would contribute to measurable overlap while the actual overlap remains as 1. The final *RBO* would thus be close to 1 instead of close to 0.
- R vs T: the top item is different but all the tied items are the same, so we should intuitively expect and *RBO* score close to 1. However, agreement is 0 at *every* depth except at the end, where it becomes nearly 1. The final *RBO* would therefore be close to 0 instead of close to 1.
- P vs Q: the top items are tied in both rankings, which means that at the earliest depths there is no actual overlap, but no measurable overlap either. This would naturally lead to an undefined agreement at those depths, ultimately resulting in an undefined *RBO*.

From these examples, we can make two observations:

- O1: we need to "look inside" the tie groups to distinguish between simply tying the same items (e.g. R vs T) and tying different items altogether (e.g. R vs W).
- O2: the measurable overlap should always be non-zero to guarantee that agreement is always defined (e.g. P vs Q).

Note that O1 is also required if we want to ensure that the *RBO* of a ranking with itself is always 1 regardless of the ties. On the other hand, O2 contrasts with Kendall's τ_b because it *is* undefined if one ranking is fully tied, reflecting the absence of information to measure concordance in the first place. But for *RBO* this can not be the case because at least at the very end of the rankings there is no more uncertainty due to ties, so all items are active and contribute to measurable overlap.

In the path towards a solution, we recognize in the actual concordance of τ_b (i.e. the numerator in Eq. (10)) a similar structure to that of *RBO*'s actual overlap in Eq. (6): they are both defined as the accumulation of the product of two individual contribution terms, one from each ranking. To define these individual contribution terms, we note that observation O1 above told us to "look inside" the tie groups so that their items have a chance to contribute to actual overlap. This is precisely what we achieved in Eq. (8) when formulating A^a, so if we similarly define actual overlap as $\sum_{e\in\Omega} c_{e,S|d} \cdot c_{e,L|d}$, then by analogy to Kendall's τ_b we can define the measurable overlap of a single ranking as $\sqrt{\sum_{e\in\Omega} c^2_{e|d}}$. Finally, we note that $c_{e|d}$ is always non-zero for every item in Ω, which ensures a non-zero measurable overlap as required by observation O2 above.

All this considered, we propose the following formulation of agreement for RBO^b:

$$A^b_{S,L,d} = \frac{\sum_{e\in\Omega} c_{e,S|d} \cdot c_{e,L|d}}{\sqrt{\sum_{e\in\Omega} c^2_{e,S|d}}\sqrt{\sum_{e\in\Omega} c^2_{e,L|d}}}, \tag{11}$$

which is bounded between 0 and 1 due to the Cauchy–Schwarz inequality. In the absence of ties, note again that only the first two cases apply in Eq. (8) and the denominator equals d, which means that A^b reduces to A and therefore $RBO^b = RBO$.

Note that A^b and A^a measure actual overlap at the numerator in the same way, but differ in the measurable overlap at the denominator. In our main example, A^a_4 has a measurable overlap of 4, while for A^b_4 ranking S contributes $\sqrt{3 + 1/3}$ and ranking L contributes $\sqrt{3 + 1/2}$. As a side product then, A^a is always less than or equal to A^b. This relation between the a- and b-variants is a result of how Kendall [13] connected the work of Student [29] and Woodbury

[36] for his definitions of τ_a and τ_b. The numerator is the same, nicely representing the expected amount of concordance across all permutations of the tied items, but the denominators then differ in whether they correct for untiedness or not. This is exactly the relation we have between A^a and A^b, and by extension between RBO^a and RBO^b.

4 PREFIX EVALUATION WITH TIES

As presented in Eq. (3), *RBO* is defined on infinite rankings, but they are usually truncated, as mentioned in Section 1. This means that rankings actually consist of a seen part or prefix, and an unseen part that extends up to infinity. Therefore, *RBO* scores have to be computed from a prefix only, ideally accompanied by some quantification of the uncertainty due to the unseen items. Webber et al. [35] presented the rationale and equations to compute upper and lower bounds on *RBO*. For the lower bound RBO_{MIN}, it is assumed that all items in the unseen parts are disjoint, thus minimizing the agreement. For the upper bound RBO_{MAX}, it is assumed that every item in the unseen part of one ranking matches an item in the other one, thus maximizing the agreement. The difference, $RBO_{RES} = RBO_{MAX} - RBO_{MIN}$, directly quantifies the magnitude of the residual. Lastly, they also introduced a point estimate named RBO_{EXT}, calculated by extrapolating the agreement measured in the prefixes, assuming the same agreement would be observed throughout the unseen parts.

However, as mentioned already in Section 3.1, the relevant equations they present (i.e. (11), (30) and (32)), as well as the rationale behind them, are always expressed in terms of overlap, not in terms of agreement. It is clear already with their A^w in Eq. (4) that a definition for agreement may have, not only a custom denominator other than simply d, but a different functional form altogether. This is now even more evident from A^a in Eq. (9) and A^b in Eq. (11). As a consequence, their equations can *not* be used for prefix evaluation of *RBO* in the presence of ties and, as will be shown in Section 4.1, the rationales behind RBO_{MAX} and RBO_{EXT} are actually a bit more involved than it seemed with bare *RBO* and no ties.

In order to fill this gap and derive a general formulation for *RBO* and prefix evaluation, we first rewrite the full *RBO* from Eq. (3) by explicitly separating three sections: 1) from depth 1 up to s, where both rankings are seen, 2) from $s+1$ up to l, where only L is seen, and 3) from $l+1$ up to infinity, where both rankings are unseen:

$$RBO_{S,L,p} = \frac{1-p}{p}\left(\underbrace{\sum_{d=1}^{s} A_d p^d}_{1} + \underbrace{\sum_{d=s+1}^{l} A_d p^d}_{2} + \underbrace{\sum_{d=l+1}^{\infty} A_d p^d}_{3}\right). \quad (12)$$

For simplicity, in the remainder of the paper we will use exclusively the formulations that compute overlap based on the products of individual contributions, such as in Eq. (6) and Eq. (9). Note that the w-variant can be easily expressed in this way, for example as:

$$A^w_{S,L,d} = \frac{2 \cdot \sum_{e\in\Omega} c_{e,S|d} \cdot c_{e,L|d}}{\sum_{e\in\Omega} c_{e,S|d} + \sum_{e\in\Omega} c_{e,L|d}}, \quad (13)$$

where $c_{e,S|d} = \mathbb{I}\left[t_{e,S} \le d\right]$, and likewise for $c_{e,L|d}$.

Note that agreement can be readily measured in the first section because both rankings are seen, but for the second section we need to make an assumption about the unseen items in S and their overlap with L. Likewise, in the third section we need an assumption about unseen items in both S and L. What assumptions are made, depends on whether we compute RBO_{MIN}, RBO_{MAX} or RBO_{EXT}.

In addition, we assume there are no ties in the unseen parts. This is necessary for the w-variant because prior information about conjointness would otherwise be needed to compute bounds. Indeed, tying all the unseen items when rankings are mostly conjoint would maximize RBO^w because they would contribute to both the numerator and denominator in Eq. (4) at earlier depths. On the other hand, tying all unseen items in a mostly non-conjoint case would actually minimize RBO^w because they would only contribute to the denominator. For the a- and b-variants, it just makes sense to assume no ties, as they assume the existence of fully untied rankings in the first place; recall that in these variants a tie just reflects an inability to distinguish items that are too close together.

4.1 Second Section: from $s+1$ to l

The agreement in this second section depends on how the unseen items in S are assumed to overlap with L, and how the agreement function combines this unseen overlap with both the seen and measurable overlaps, which ultimately depends on the tie-variant.

Let us define this assumed overlap $\tilde{X}_d$ by separating two components: one measuring the actual overlap among the seen items, and another one incorporating the assumed contribution of the unseen items in S:

$$\tilde{X}_d = \underbrace{\sum_{e\in\Omega} c_{e,S|d} \cdot c_{e,L|d}}_{\text{seen}} + \underbrace{\sum_{k=s+1}^{d} \tilde{c}_{k,S|d} \cdot \tilde{c}_{k,L|d}}_{\text{unseen}}. \quad (14)$$

Note that the first summation is simply the regular overlap, and that an item that only appears in L will *not* contribute here because its $c_{S|d}$ is 0. These unmatched items from L are the ones that have a chance to overlap with the $d-s$ unseen items in S through the second summation. For each of those unseen items in S we assume a contribution $\tilde{c}_{k,S|d}$, and a corresponding non-constant contribution $\tilde{c}_{k,L|d}$ that depends on what item is actually matched in L. This is the point where the derivations by Webber et al. [35] are not sufficiently general to accommodate ties, because when an item in L is matched in the unseen part of S, they give it a unitary contribution to overlap, which is not necessarily correct in the presence of ties and fractional contributions.

In order to compute agreement in each of the three variants, we need to combine the assumed overlap in Eq. (14) with the measurable overlap. In this respect, we note that the measurable overlap contributed by S is always equal to d because unseen items are assumed to be untied. The second summation in Eq. (12) becomes:

$$\left(\sum_{d=s+1}^{l} A_d p^d\right)^w = \sum_{d=s+1}^{l} \frac{2 \cdot \tilde{X}_d}{d + |L_{:d}|} p^d, \quad (15)$$

$$\left(\sum_{d=s+1}^{l} A_d p^d\right)^a = \sum_{d=s+1}^{l} \frac{\tilde{X}_d}{d} p^d, \text{ and} \quad (16)$$

$$\left(\sum_{d=s+1}^{l} A_d p^d\right)^b = \sum_{d=s+1}^{l} \frac{\tilde{X}_d}{\sqrt{d}\sqrt{\sum_{e\in\Omega} c^2_{e,L|d}}} p^d. \quad (17)$$

4.1.1 RBO_{MIN}. For the lower bound it is assumed that all unseen items are disjoint. This means that their individual contributions are 0, so they do not contribute to the unseen overlap in any way:[5]

$$\big(\tilde{c}_{k,S|d}\big)_{MIN} = 0\,, \quad \big(\tilde{c}_{k,L|d}\big)_{MIN} = 0\,. \tag{18}$$

4.1.2 RBO_{MAX}. Every unseen item in S has a unitary contribution because it is untied and it matches an item in L. However, the corresponding contribution in L must take into account the order of the unmatched items. Indeed, *RBO* is maximized when the k-th unseen item at rank $s+k$ matches the k-th still unmatched item in L. These are the grayed-out items in Figure 2. For instance, `m` is the item that maximizes agreement at depth 9, with a contribution $\tilde{c}_{L|d} = 1$. Note that the item maximizing agreement at depth 12 can be any of `j`, `k`, `o` and `q`, for they are in a crossing group at that depth. In the *w*-variant they would have an individual contribution $\tilde{c}_{L|d} = 1$, or $\tilde{c}_{L|d} = 3/4$ in the *a*- and *b*-variants.

We therefore need to know the sequence of items unique to L that are potentially active, and arranged in the same order; let us refer to these as $U_d = \langle u_i \,:\, c_{u_i,L|d} > 0 \wedge c_{u_i,S|d} = 0 \wedge c_{u_i,L|d} \geq c_{u_{i+1},L|d} \rangle$. The individual contributions to the unseen overlap are therefore:

$$\big(\tilde{c}_{k,S|d}\big)_{MAX} = 1\,, \quad \big(\tilde{c}_{k,L|d}\big)_{MAX} = c_{u_k,L|d}\,. \tag{19}$$

In the absence of ties, note that all $\tilde{c}_{k,L|d}$ are equal to 1.

4.1.3 RBO_{EXT}. In this case, no specific items are assumed for the unseen ranks of S. Instead, the assumption is about their individual contributions to unseen overlap. Webber et al. [35] decided to set these contributions equal to A_s, albeit the corresponding contribution from ranking L was still assumed to be 1. However, and similarly to the case of RBO_{MAX}, these contributions are not necessarily 1 when dealing with ties and crossing groups. We take a slightly different approach and assume that an unseen item in S may match, with equal probability, any of the potentially active but still unmatched items in L, that is, any of the items in U_d. We then ask for the expected value of the joint contribution at rank k:

$$\begin{aligned}\mathrm{E}\big[c_{S|d} \cdot c_{L|d} \,\big|\, k\big] &= P(\text{unmatch} \mid k) \cdot 0 + \\ &+ P(\text{match } u_1 \mid k) \cdot 1 \cdot c_{u_1,L|d} + \\ &+ P(\text{match } u_2 \mid k) \cdot 1 \cdot c_{u_2,L|d} + \cdots = \\ &= P(\text{match} \mid k) \cdot \mathrm{E}\big[c_{L|d} \,\big|\, U_d\big]\,,\end{aligned} \tag{20}$$

We note here that $P(\text{match} \mid k)$ is precisely where extrapolation happens via A_s; indeed, agreement can be interpreted as the probability that an item chosen at random appears in both rankings. From this view, A_s is therefore not the assumed contribution of the unseen item in S, which is always 1 because we assume it to be untied, but rather the probability that it matches something in L. Nonetheless, the individual contributions may be defined as follows to incorporate in Eq. (14):

$$\big(\tilde{c}_{k,S|d}\big)_{EXT} = A^*_s\,, \quad \big(\tilde{c}_{k,L|d}\big)_{EXT} = \frac{1}{|U_d|}\sum\nolimits_{e \in U_d} c_{e,L|d}\,. \tag{21}$$

Note that the agreement at s depends on what tie-aware variant is being used, which we indicate with the star $*$. In the absence of ties, all contributions from L are also unitary, which means that all $\tilde{c}_{L|d}$ are ultimately equal to 1, making the total contribution of an unseen item equal to A_s, as formulated for bare *RBO*.

[5] Setting $\tilde{c}_{k,L|d} = 0$ is arbitrary, but this is irrelevant because $\tilde{c}_{k,S|d}$ *must* be 0 anyway.

Finally, we must note that the agreement in Eq. (17) is still bounded by 1. Loosening notation, $\sum \tilde{c}_{S|d}\tilde{c}_{L|d}$ is bounded by $\sqrt{\sum \tilde{c}^2_{S|d}}\sqrt{\sum \tilde{c}^2_{L|d}}$ due to the Cauchy–Schwarz inequality. The first term is itself bounded by $\sqrt{k}$ because $A^*_s \leq 1$, and the second term is bounded by $\sqrt{\sum c^2_{L|d}}$ due to Jensen's inequality.

4.2 Third Section: from $l+1$ to ∞

Regarding the third summation in Eq. (12), we first note that all the items seen in the prefixes are active at depths l and beyond, so the seen overlap at those depths is independent of the tie variant.

4.2.1 RBO_{MIN}. All unseen items in the third section are assumed to be disjoint, so the overlap will remain constant and equal to X_l:

$$\left(\sum_{d=l+1}^{\infty} A_d p^d\right)_{MIN} = \sum_{d=l+1}^{\infty} \frac{X_l}{d} p^d = X_l \left[\ln\left(\frac{1}{1-p}\right) - \sum_{d=1}^{l} \frac{p^d}{d}\right]\,. \tag{22}$$

4.2.2 RBO_{MAX}. The $l-s$ unseen items in S from the second section are assumed to match an item in L, so the assumed overlap at depth l becomes $X_l + l - s$. After l, every unseen item in L is assumed to match an unmatched item in S and vice-versa, thus contributing +2 to the overlap (in our example, this happens at depths 14 and 15). This continues until all the remaining unmatched items are placed after the prefixes, which happens at depth $f = l + s - X_l$ (in our example, $f = 15$). After f, it is assumed that the same item would appear in both rankings, thus continuing the full agreement indefinitely. The third summation for RBO_{MAX} is therefore split in two subsections: from $l+1$ up to f, where overlap increases by 2 at each step, and from $f+1$ up to ∞, where full agreement is assumed:

$$\begin{aligned}\left(\sum_{d=l+1}^{\infty} A_d p^d\right)_{MAX} &= \sum_{d=l+1}^{f} \frac{2d - l - s + X_l}{d} p^d + \sum_{d=f+1}^{\infty} p^d \\ &= \sum_{d=l+1}^{f} \frac{2d - l - s + X_l}{d} p^d + \frac{p^{f+1}}{1-p}\,.\end{aligned} \tag{23}$$

4.2.3 RBO_{EXT}. Recall that we have to extrapolate the agreement at l to all subsequent depths up to infinity. Because all items are finally active in this third section, we have that $\tilde{c}_{L|d}$ is always 1. This means that the assumed overlap $\tilde{X}_l$ that we extrapolate equals $X_l + A^*_s(l-s)$. In addition, both S and L have the same full contribution to measurable overlap at the denominators, which means that the extrapolated agreement at l takes the same form for all tie-variants: $\big(X_l + A^*_s(l-s)\big)/l$. Extrapolating this agreement up to infinity, we have the third summation for RBO_{EXT}:

$$\begin{aligned}\left(\sum_{d=l+1}^{\infty} A_d p^d\right)^*_{EXT} &= \sum_{d=l+1}^{\infty} \frac{X_l + A^*_s(l-s)}{l} p^d \\ &= \frac{X_l + A^*_s(l-s)}{l} \cdot \frac{p^{l+1}}{1-p}\,.\end{aligned} \tag{24}$$

In the absence of ties, all these formulations reduce to bare *RBO*.

5 EXPERIMENTAL DEMONSTRATIONS

We now illustrate the use of *RBO* in the presence of ties, emphasizing the differences between each of the three tie-aware variants

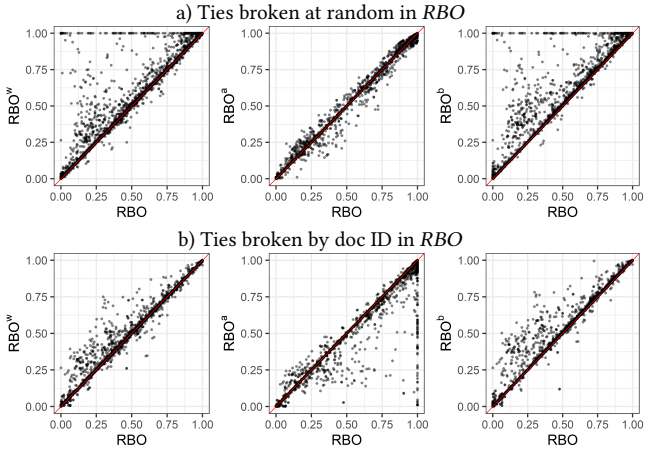


Figure 4: Differences between bare *RBO* and each of the three tie-aware variants. TREC data, $p = 0.9$.

and bare *RBO*, thus showing the importance of handling ties explicitly. Because using *RBO* normally involves the calculation of the extrapolated score, we focus only on RBO_{EXT}. Revisiting Section 4, we see that all variants are essentially the same with respect to the bounds, where the only difference is the definition of $\left(\tilde{c}_{k,L|d}\right)_{MAX}$. These differences are negligible and make RBO_{RES} depend mostly on the evaluation depth, not on the handling of ties.

5.1 TREC Data

A common use case for *RBO* is comparing the document rankings returned by an experimental system with those of a baseline. From all adhoc runs in the TREC 2009–2014 Web track (see Table 1), we compared the rankings between all 255 pairs of runs by the same group and over the 50 topics, for a grand total of 12,750 pairs of rankings. Of those, 9,256 contained ties, which are the ones we report about. The average ranking length was 978 documents, with a maximum of 1,000. For every pair of rankings, we computed the tie-unaware *RBO*, breaking ties at random and by doc ID, as well as RBO^w, RBO^a and RBO^b. The chosen values for p were 0.8, 0.9 and 0.95, thus setting the expected number of results compared by the p-persistent user to 5, 10 and 20 documents, respectively.

Figure 4 compares all *RBO* scores for $p = 0.9$. All variants are of course highly correlated, but we can observe some striking differences. Focusing first on RBO^w and RBO^b, we see that they are closer to bare *RBO* when breaking ties by doc ID. As mentioned earlier, this is expected because both rankings will have these tied items artificially sorted in the same order when presented to bare *RBO*, inflating the result. On the other hand, RBO^w and RBO^b are specifically designed to deal with these items, and if they happen to be similarly distributed in both rankings they will actually contribute positively. In contrast, RBO^a is closer to bare *RBO* when breaking ties at random. Again, this is expected, as they are by construction equal on expectation (see Section 3.2). As a consequence, RBO^a is generally lower than *RBO* breaking ties by doc ID because these documents contribute in different directions across permutations. The key takeaway is that there can be *very* large differences among

Table 3: Summary of differences between bare *RBO* and each of the three tie-aware variants. M for medium differences in $(0.01, 0.1]$, and L for large in $(0.1, 1]$. TREC data.

a) Ties broken at random in *RBO*

	$\lvert RBO - RBO^w \rvert$				$\lvert RBO - RBO^a \rvert$				$\lvert RBO - RBO^b \rvert$			
p	Avg.	Max.	M	L	Avg.	Max.	M	L	Avg.	Max.	M	L
0.8	0.02	1	8%	5%	0.01	0.43	7%	2%	0.02	1	7%	5%
0.9	0.02	1	9%	4%	<.01	0.26	7%	1%	0.02	1	8%	4%
0.95	0.01	1	9%	3%	<.01	0.15	6%	<1%	0.01	1	8%	3%

b) Ties broken by doc ID in *RBO*

	$\lvert RBO - RBO^w \rvert$				$\lvert RBO - RBO^a \rvert$				$\lvert RBO - RBO^b \rvert$			
p	Avg.	Max.	M	L	Avg.	Max.	M	L	Avg.	Max.	M	L
0.8	0.01	0.60	6%	2%	0.01	1	6%	4%	0.01	0.63	5%	3%
0.9	0.01	0.45	6%	2%	0.01	0.99	8%	2%	0.01	0.48	5%	2%
0.95	<.01	0.31	5%	1%	0.01	0.98	8%	2%	0.01	0.45	5%	2%

variants, so a sensible decision should be made as to which one should be computed depending on the specific meaning of ties.

Table 3 provides a summary of these differences among variants. In particular, we can see that, while differences are small on average, there are *very* large cases. Overall, the table confirms that RBO^w and RBO^b are most different from bare *RBO* when breaking ties at random, whilst that makes it closer to RBO^a. Recall here that the table reports *absolute* differences; signed differences with RBO^a are actually 0, as expected. Another way to look at deviations is by classifying them in large (more than 0.1), medium (between 0.01 and 0.1) or small (less than 0.01), which roughly translates into differences in the first, second, or third decimal digit of a reported *RBO* score, respectively; the first two are identified as M and L in Table 3. As can be seen, about 4% of the observed differences are large, while about 8% are of medium size. This means that the strategy followed to deal with ties brings a substantial difference in about 12% of the comparisons between rankings.

While these summary statistics give us a broad idea of the contrast between dealing with ties or not, we must note that the largest differences are mostly found between rankings with an extreme structure in one or two specific aspects. The first aspect is the *amount* of ties: RBO^w and RBO^b achieve maximum overlap when the tie groups are the same in both rankings, and the chance of this happening increases when all or most items are actually tied (e.g., 9% of the rankings have at least 90% of their items tied). For instance, Figure 5-top illustrates the case of bare *RBO* vs RBO^b, faceting by the amount of ties found in the rankings. The most extreme differences indeed appear when most items are tied, although even with a small number of ties we can observe differences larger than 0.2. The second aspect is the *position* of the tied items: it is not enough to have many ties; they need to appear towards the top of the ranking in order to have an impact in the score. We may roughly quantify the potential impact of ties by simply summing the p-dependent weight of their ranks: $\left(\sum_d \mathbb{I}[d \text{ is tied}]\, p^d\right) / \sum_d p^d$. Figure 5-bottom confirms that differences between bare *RBO* and RBO^b are more pronounced when the potential impact of ties is large. Therefore, rankings with a moderate number of ties may still exhibit large variations if those ties appear towards the top.

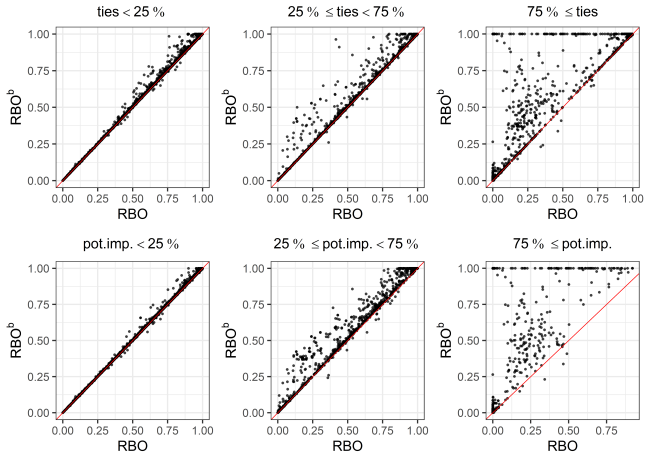


Figure 5: Differences between bare *RBO* (ties broken at random) and RBO^b, faceted by the amount of tied items (top) and by their potential impact (bottom). TREC data, $p = 0.9$.

In general, Figure 5 shows that the presence of ties does not immediately lead to a difference with respect to bare *RBO*, as this ultimately depends on the arrangement of those tied items. What ties give is room for these variations to be large. In other words, ties affect mostly the variance between measures, not the bias.

5.2 Synthetic Data

The results from the previous section should not generalize well to non-IR settings, as they involved quite long rankings, mostly even, and where the underlying domains (i.e. the document collections) are several orders of magnitude larger than the rankings, thus leading to a high degree of non-conjointness. In order to provide more general results, in this section we consider a synthetic dataset, generated as follows. Two rankings are generated with a Kendall τ between 0.5 and 1 over the same 1,000 items. Ties are introduced at random in each ranking and independent of the other, for a target tiedness between 10% and 100%, after which it is truncated to a length between 10 and 100. This was repeated 100,000 times, resulting in rankings with an average of 55 items, an average length difference of 30 items, and an average of 54% items tied. In this case, we only break ties at random when computing bare *RBO*.

Figure 6 similarly compares all *RBO* scores for $p = 0.9$. In clear contrast with Figure 4, we can first see that there are far fewer extreme deviations. This is because the simulated dataset does not contain rankings with such extreme structures as displayed in the TREC data. In general, we see that RBO^w, and specially RBO^b, tend to produce higher scores than bare *RBO*, because at a given depth they allow all items in a group to contribute to the final score, not only those that are active at that depth. On the other hand, RBO^a calculates the expected *RBO* over permutations of the tied items, so differences are again nicely distributed around the diagonal.

Even if there are no extreme cases, deviations are generally larger than observed with the TREC data. As summarized in Table 4, average deviations are 3 to 4 times larger, but most importantly, the amount of medium and large deviations is even an order of magnitude larger. In this case, the strategy followed to deal with ties brings a substantial difference in about 70% of the comparisons.

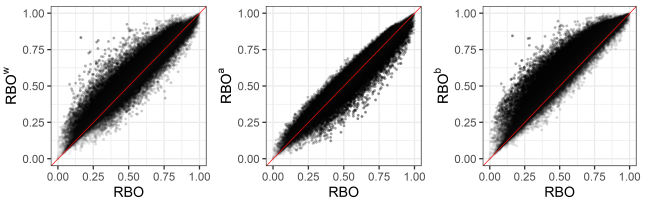


Figure 6: Differences between bare *RBO* (ties broken at random) and the three tie-aware variants. Synthetic data, $p = 0.9$.

Table 4: Summary of differences between bare *RBO* and each of the three tie-aware variants. M for medium differences in $(0.01, 0.1]$, and L for large in $(0.1, 1]$. Synthetic data.

	$\lvert RBO - RBO^w \rvert$				$\lvert RBO - RBO^a \rvert$				$\lvert RBO - RBO^b \rvert$			
p	Avg.	Max.	M	L	Avg.	Max.	M	L	Avg.	Max.	M	L
0.8	0.07	0.76	50%	26%	0.05	0.51	52%	17%	0.08	0.77	46%	31%
0.9	0.04	0.67	64%	10%	0.03	0.34	63%	4%	0.06	0.68	56%	20%
0.95	0.03	0.53	63%	3%	0.02	0.25	56%	<.01%	0.04	0.54	62%	8%

6 CONCLUSION AND RECOMMENDATIONS

In this work we delved into the problem of tied items in Rank-Biased Overlap. First, we argued that the existing approach is incomplete, for it is unclear how to apply it to compute RBO_{EXT} and its bounds. More importantly, we showed that the notion of ties behind this approach (i.e. the sports ranking) is very different from the one traditionally used in the Statistics literature (i.e. uncertainty as to the actual order), most notably in Kendall's τ. We therefore developed two other variants of *RBO* to accommodate this traditional view on ties. Through a general formulation for prefix evaluation of *RBO*, we also showed how to fully compute all three variants.

Filling this gap, researchers can now make a conscious and sensible decision when dealing with ties. Our **recommendations** are:

- When a tie represents equality, so that tied items *really* occur at the same rank, one should compute $\boldsymbol{RBO^w}$.
- When a tie represents uncertainty, so that it is not known which item appears first:
  - Ties should not be broken deterministically, such as by doc ID, because it inflates *RBO* scores.
  - Ties should not be broken at random because it introduces noise. $\boldsymbol{RBO^a}$ should be used instead, as it precisely computes the expected *RBO* when breaking ties at random.
  - If the measured overlap should be corrected by the amount of information lost due to ties, $\boldsymbol{RBO^b}$ should be used. This ensures $RBO^b(X, X) = 1$, and implies $RBO^a \leq RBO^b$.

As future work, we will bound the uncertainty introduced by ties, similarly to how bounds are used to quantify the uncertainty due to unseen items.

ACKNOWLEDGMENTS

Work facilitated by computational resources of the Delft AI Cluster at TU Delft. We dedicate this work to Akira Toriyama.

REFERENCES

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan Yousef. 2022. Topic Modeling Algorithms and Applications: A Survey. *Information Systems* 112 (2022).

[2] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *PInternational ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–404.

[3] Judit Bar-Ilan, Mazlita Mat-Hassan, and Mark Levene. 2006. Methods for Comparing Rankings of Search Engine Results. *Computer Networks* 50, 10 (2006), 1448–1463.

[4] Amanda Buch, Petra Vértes, Jakob Seidlitz, So Kim, Logan Grosenick, and Conor Liston. 2023. Molecular and Network-level Mechanisms Explaining Individual Differences in Autism Spectrum Disorder. *Nature Neuroscience* 26 (2023), 1–14.

[5] Chris Buckley. 2004. Topic Prediction Based on Comparative Retrieval Rankings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 506–507.

[6] Guillaume Cabanac, Gilles Hubert, Mohand Boughanem, and Claude Chrisment. 2010. Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. 112–123.

[7] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline Evaluation Options for Recommender Systems. *Information Retrieval Journal* 23, 4 (2020), 387–410.

[8] Bruno Cardoso and João Magalhães. 2011. Google, Bing and a New Perspective on Ranking Similarity. *ACM International Conference on Information and Knowledge Management*, 1933–1936.

[9] Ben Carterette. 2009. On Rank Correlation and the Distance Between Rankings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 436–443.

[10] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. 2020. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *ACM International Conference on Information and Knowledge Management*. 225–234.

[11] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. 2003. Comparing Top k Lists. *SIAM Journal on Discrete Mathematics* 17, 1 (2003), 134–160.

[12] Maurice G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1 (1938), 81–93.

[13] Maurice G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251.

[14] Maurice G. Kendall. 1948. *Rank Correlation Methods* (4th ed.). Charles Griffin and Company Limited.

[15] A. Kolmogorov. 1933. Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari* 4 (1933), 83–91.

[16] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized Distances Between Rankings. In *International Conference on World Wide Web*. 571–580.

[17] Jimmy Lin and Peilin Yang. 2019. The Impact of Score Ties on Repeatability in Document Ranking. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1125–1128.

[18] Tinghuai Ma, Qin Liu, Jie Cao, Yuan Tian, Abdullah Al-Dhelaan, and Mznah Al-Rodhaan. 2020. LGIEM: Global and Local Node Influence based Community Detection. *Future Generation Computer Systems* 105 (2020), 533–546.

[19] Mika V. Mantyla, Maelick Claes, and Umar Farooq. 2018. Measuring LDA Topic Stability from Clusters of Replicated Runs. In *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*.

[20] Frank McSherry and Marc Najork. 2008. Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In *European Conference on Information Retrieval*. 414–421.

[21] Massimo Melucci. 2007. On Rank Correlation in Information Retrieval Evaluation. *ACM SIGIR Forum* 41, 1 (jun 2007), 18–33.

[22] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Transactions on Information Systems* 35, 3 (2017).

[23] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Annual Network and Distributed System Security Symposium*.

[24] Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. 1989. A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Transactions on Information Systems* 7, 3 (1989), 205—-229.

[25] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. 2020. Interplay between Hierarchy and Centrality in Complex Networks. *IEEE Access* 8 (2020), 129717–129742.

[26] Chiman Salavati, Alireza Abdollahpouri, and Zhaleh Manbari. 2018. BridgeRank: A Novel Fast Centrality Measure based on Local Structure of the Network. *Physica A: Statistical Mechanics and its Applications* 496 (2018), 635–653.

[27] Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72–101.

[28] Lukas Steuernagel, Brian Lam, Paul Klemm, Georgina Dowsett, Corinna Bauder, John Tadross, Tamara Sotelo-Hitschfeld, Almudena Martin, Weiyi Chen, Alain De Solis, Henning Fenselau, Peter Davidsen, Irene Cimino, Sara Kohnke, Debra Rimmington, Anthony Coll, Andreas Beyer, Giles Yeo, and Jens Brüning. 2022. HypoMap: A Unified Single-cell Gene Expression Atlas of the Murine Hypothalamus. *Nature Metabolism* 4 (2022), 1402–1419.

[29] Student. 1921. An Experimental Determination of the Probable Error of Dr. Spearman's Correlation Coefficients. *Biometrika* 13, 2/3 (1921), 263–282.

[30] Mingxuan Sun, Guy Lebanon, and Kevyn Collins-Thompson. 2010. Visualizing Differences in Web Search Algorithms Using the Expected Weighted Hoeffding Distance. In *International Conference on World Wide Web*. 931–940.

[31] Luchen Tan and Charles L. A. Clarke. 2015. A Family of Rank Similarity Measures Based on Maximized Effectiveness Difference. *IEEE Transactions on Knowledge and Data Engineering. Knowl. Data Eng.* 27, 11 (2015), 2865–2877.

[32] Julián Urbano and Mónica Marrero. 2017. The Treatment of Ties in AP Correlation. In *ACM SIGIR International Conference on the Theory of Information Retrieval*. 321–324.

[33] Sebastiano Vigna. 2015. A Weighted Correlation Index for Rankings with Ties. In *International Conference on World Wide Web*. 1166–1176.

[34] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 173–183.

[35] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Transactions on Information Systems* 28, 4 (2010), 1–38.

[36] Max A. Woodbury. 1940. Rank Correlation When There are Equal Variates. *The Annals of Mathematical Statistics* 11, 3 (1940), 358–362.

[37] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 587–594.

[38] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 395–404.