A Probabilistic Account of the Uncertainty Due to Ties in Rank-Biased Overlap

Efficient Estimation of the Uncertainty Distribution for Tied Data

Lukáš Chládek[†]

Supervised by:
Prof. Julián Urbano[†]

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

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

June 22, 2025

Name of the student:	Lukáš Chládek
Final project course:	CSE3000 Research Project
Thesis committee:	Prof. Julián Urbano and Prof. Lilika Markatou

A Probabilistic Account of the Uncertainty Due to Ties in Rank-Biased Overlap

Lukáš Chládek

Delft University of Technology
Delft, The Netherlands

Abstract

Rank similarity quantifies the difference between two ordered sets of items. Rank-Biased Overlap (RBO) is a top-weighted measure of rank similarity that can be used for a pair of indefinite rankings, such that only a prefix is known and that items need not be present in both rankings. This method is frequently used in Information Retrieval (IR), such as to compare search engine results. RBO defines tight lower and upper bounds, $\mathrm{RBO}_{\mathrm{min}}$ and $\mathrm{RBO}_{\mathrm{max}}$, which give the uncertainty due to items in the unseen suffix. Another source of uncertainty are ties: two items are tied in a ranking if their true order is not known. Recent work on the treatment of ties in RBO has made it a tie-aware measure. However, unlike the uncertainty due to unseen items, uncertainty due to ties does not disappear for longer prefixes. Determining the distribution of possible scores is $O((n!)^2)$ if all arrangements of ties are considered, and existing methods only find the lower and upper bound for RBO with respect to ties. We investigate whether a probabilistic estimator for the uncertainty distribution can be constructed. We use an iterative convolution method to compose the marginal PMFs of each item. By evaluating against synthetic data, we show that this estimate distribution can be used to reliably compute confidence intervals, mean, and variance. We conclude that a probabilistic method is a viable solution when seeking deterministic results with fast computation.

KEYWORDS

Information retrieval, rank similarity, rank-biased overlap, uncertainty, ties

ACM Reference Format:
Lukáš Chládek. 2025. A Probabilistic Account of the Uncertainty in Rank-Biased Overlap. Bachelor's thesis. Delft University of Technology, Delft, The Netherlands. Retrieved from https://repository.tudelft.nl/.

1 INTRODUCTION

Sports results, Web searches, and triage lists are all examples of rankings [1]. Coefficients of rank similarity are methods that can compare two such rankings and indicate the degree to which they agree or differ—comparing different search engines, for example, is of particular interest in the field of Information Retrieval (IR). Is a faster search engine better even though it ranks the most relevant result lower? Rank similarity provides objective values that enable such comparisons.

A subclass of the problem of rank similarity is rank correlation, for which two notable coefficients are Kendall's τ [6] and Spearman's ρ [11]. These are defined for a pair of definite rankings, which requires finite length and conjointness (each ranking is complete along a common domain).

1.1 Similarity of Indefinite Rankings

Unweighted correlation coefficients, such as τ, are not well-suited to applications in Information Retrieval (IR) and Recommender Systems (RecSys). In IR, a top-weighed measure is of interest due to the higher importance of earlier results [13]. (In comparing two Web search engines on one query, the similarity of Page 1 is more telling than that of Page 100). Such results are also nonconjoint, as IR systems may not have the same access or criteria for candidate results, and can have a length that approaches infinity [13]. Webber et al. define the properties of an indefinite ranking, in that it is nonconjoint, top-weighted, and can be truncated arbitrarily [13].

The authors also propose Rank-Biased Overlap (RBO), a new rank similarity coefficient for indefinite rankings [13]. T RBO allows for evaluating the prefix of a pair of nonconjoint rankings; by its definition, an infinitely long tail of unknown values does not dominate the known prefix.

1.2 Uncertainty in Rank-Biased Overlap

In Rank-Biased Overlap, Webber et al. include uncertainty due to prefix evaluation [13]. For a nonconjoint pair of infinite rankings S and T, RBO can give results even if only fixed-length prefixes are known and an infinite tail remains unknown. In such a case, the bounds $\mathrm{RBO}_{\mathrm{min}}$ and $\mathrm{RBO}_{\mathrm{max}}$ represent the lower and upper bound on the complete ranking's RBO, where $\mathrm{RBO}_{\mathrm{min}}$ assumes that the unknown tails are entirely disjoint while $\mathrm{RBO}_{\mathrm{max}}$ assumes full unseen agreement. Extrapolating the agreement seen up to d, the point estimate $\mathrm{RBO}_{\mathrm{ext}}$ can also be used.

Due to RBO's top-weightedness, the resulting uncertainty decreases monotonically with evaluation depth d. Any RBO value at a greater depth will necessarily be between $\mathrm{RBO}_{\mathrm{min}}$ and $\mathrm{RBO}_{\mathrm{max}}$ as evaluated at d. This principle can be seen in Figure 1, which visualizes RBO on an arbitrarily chosen pair

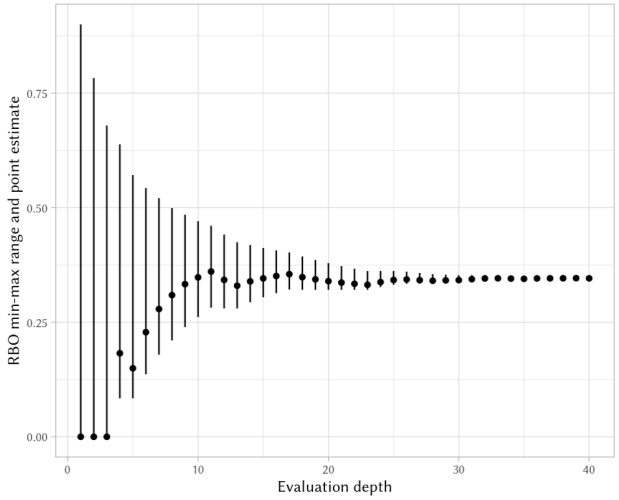

Figure 1: **RBO uncertainty due to unseen items converges.** S and T are arbitrarily chosen and without ties. $p = 0.9$.

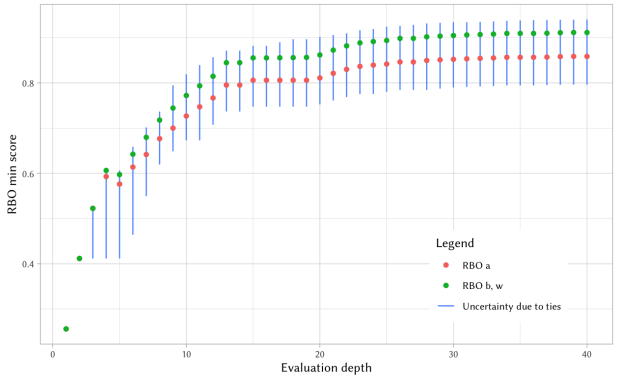

Figure 2: **RBO uncertainty due to ties does not converge.** S and T are arbitrarily chosen with ties. RBO_{min}, $p = 0.9$.

of correlated rankings: as more of S and T is revealed with increasing depth d, the uncertainty narrows and RBO_{ext} converges to its true value for S and T.

This convergence allows control over the uncertainty; for an infinite ranking, increasing the evaluation depth will always cause lower uncertainty, allowing for compromise between desired precision and computation time.

1.3 Uncertainty Due to Ties

While uncertainty due to unseen items can be mitigated by evaluating at greater depth, this is not the case for ties. First, the bounds RBO_{min} and RBO_{max} only consider unseen items and do not take uncertainty due to ties into account. A tie is when two or more items occupy the same rank in a ranking. Each tie of length k can be broken in $k!$ ways, and selecting any one permutation would be arbitrary. Figure 3 shows a tied ranking S and its $2! = 2$ combinations.

$$
\begin{aligned}
S &= < A \ (B \ C) \ D \ E > \\
S^a &= < A \ B \ C \ D \ E > \\
S^b &= < A \ C \ B \ D \ E >
\end{aligned}
$$

Figure 3: **Breaking a pair of tied items.** S is a tied ranking. S^a and S^b are the ways of breaking it.

A tie can represent two underlying features of the ranked data [3, 6]: equality or uncertainty. In the case of equality, there is no true order between two items - for instance, two ice hockey teams could both score the same number of points and occupy the same rank in a tournament. In the case of uncertainty, a true ordering exists but is not known. For example, two different search results could have the same retrieval score as a result of rounding. Breaking such ties would be arbitrary.

In any case, ties affect rank similarity measures: Kendall [7] first adapted the τ correlation coefficient to the tie-aware variants τ_a and τ_b, and Corsi and Urbano [3] similarly proposed tie-aware RBO: RBO^a and RBO^b for use when a true order exists, but is unknown, and RBO^w for use when ties represent exact equality. The tie-aware variants of RBO allow its use with tied data, such as by producing the expectation for all tie permutations (RBO^a), but do not give the researcher an accurate uncertainty to quote. The uncertainty is, however, necessary, because each permutation of the ties can have a different RBO value.

In practice, the number of possible values is an issue. To define it, let us establish that S and T are indefinite rankings with ties and that the notation $|S_k|$ denotes the number of items tied for rank k of ranking S. By the principles of combinatorics, the number of possible permutations of ties in a prefix of length d of rankings S and T is given by:

$$\prod_{i=1}^{n} |S_i|! \times |T_i|! \tag{1}$$

The factorial growth quickly outpaces computational limits, with a worst-case complexity of $O((d!)^2)$ if the entire rankings are tied. In the words of Corsi and Urbano [4],

> A brute-force approach that calculates RBO for all possible permutations is off the table, for the number of permutations grows factorially with the number of ties. To put this into perspective, we note that rankings by a typical TREC Web run have more permutations than atoms in the observable universe.

1.4 Related Works

Obtaining the exact uncertainty distribution by examining all combinations is clearly out of the question for practical applications. However, it cannot be ignored, as evidenced by a number of works in the field of IR. Yang et al. [14] discuss the effect of uncertainty due to ties in Rank-Biased Precision (RBP), an RBO-adjacent measure of retrieval precision [9], concluding that there are cases where this uncertainty can disrupt otherwise sound conclusions. Ounis et al. [10] note that the choice of tie-breaking heuristic affected comparisons between retrieval precision of groups participating in the TREC 2011 Microblog track. Lin and Yang [8] similarly note that arbitrary tie-breaking challenges experiment reproducibility and suggest to consistently break by document ID. Most recently, Corsi and Urbano contribute in [3] that ties in Rank-Biased Overlap should be broken using their tie-aware variants, discarding all arbitrary orderings such as document ID. This viewpoint is

shared by Cabanac et al. [2], who note that such orderings must be acknowledged an uncontrolled parameter.

Summarily, prior research generally agrees on the fact that uncertainty due to ties affects conclusions made using measures such as RBO. Authors' views generally differ on the heuristics or variants that are to be used, each of which yield different values for Rank-Biased Overlap. A correct means of ordering tied items cannot exist; consequently, valid conclusions can only be drawn by quantifying the uncertainty created by ties.

In their second paper, Corsi and Urbano [4] proposed an algorithm to efficiently find RBO^{low} and RBO^{high}: the bounds for RBO for all possible ties. Bounds are also used for the uncertainty due to unseen items: RBO_{min} and RBO_{max}. However, there is a clear difference between unseen item uncertainty and tie uncertainty: the former converges to zero with greater depth, while the latter never converges. This is illustrated by Figure 2. Uncertainty due to ties cannot be eliminated except by breaking the ties, and bounds alone discard any information about what is between them, including the skewness and spread of the underlying distribution.

1.5 Contribution

This paper proposes to fill the research gap of efficient estimation of the uncertainty distribution due to ties. Section 1.4 has outlined the importance of uncertainty due to ties while showing that its distribution is computationally intractable to compute. Further, we note that the approach of Corsi and Urbano [4] is only suitable for finding the distribution bounds, and that the authors have commented that determining confidence intervals would be a useful future approach. Consequently, we propose to answer the following research question:

> How can the uncertainty of Rank-Biased Overlap for tied rankings be represented probabilistically?

A useful answer to this research question would entail a probabilistic model that can efficiently estimate the distribution of uncertainty due to ties in Rank-Biased Overlap.

The method must be resistant to outliers, should produce variable confidence interval-like bounds, and must also be more efficient than searching the tie permutation space, which is $O((n!)^2)$ in the input size. This paper proposes a method of estimating the uncertainty distribution of RBO for tied data that satisfies these requirements.

2 RANK-BIASED OVERLAP

RBO, as defined by Webber et al. [13], is defined as a sum across depth d of two rankings S and T:

$$\text{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} A_d \tag{2}$$

In the equation above, RBO uses the persistence parameter $p \in (0,1)$, to tune the top-weightedness of the measure. A_d is the agreement of rankings S and T up to depth d such that:

$$A_d = \frac{|S_{:d} \cap T_{:d}|}{d} \tag{3}$$

3 UNCERTAINTY IN ITEM CONTRIBUTION

This section will outline the transformation of RBO, which iteratively sums agreements up to depth d, into a formulation that allows us to estimate its distribution due to ties. First, we will separate RBO into the contributions of individual items in the intersection of S and T. We will then derive the marginal probability mass function (PMF) for an item's contribution to RBO. Next, Section 4 will describe how a modified convolution can efficiently combine these PMFs to obtain the uncertainty distribution of RBO due to ties.

3.1 Item Contribution to RBO

We will define item contribution for RBO for infinite rankings. Starting with the sum-of-agreements formulation of Rank-Biased Overlap, which is equivalent to Equation 2:

$$\begin{aligned}\text{RBO}(S,T,p) &= (1-p)\sum_{d=1}^{\infty} \frac{p^{d-1}|S_{:d} \cap T_{:d}|}{d} \\ &= (1-p)\sum_{d=1}^{\infty} \sum_{i \in (S_{:d} \cap T_{:d})} \frac{p^{d-1}}{d}\end{aligned} \tag{4}$$

We use $\mathbb{I}$ to denote the Iverson bracket, where $\mathbb{I}[P]$ is 1 if the logical proposition P is true and 0 otherwise. We convert the summation from the ranking slices $S_{:d}$ and $T_{:d}$ to *all* items $S \cap T$. Note that the contribution of each item $i \notin S \cap T$ is zero, as these items never add to the agreement.

$$\begin{aligned}&= (1-p)\sum_{d=1}^{\infty} \sum_{i \in (S \cap T)} \mathbb{I}[i \in S_{:d} \wedge i \in T_{:d}] \frac{p^{d-1}}{d} \\ &= (1-p)\sum_{i \in (S \cap T)} \sum_{d=1}^{\infty} \mathbb{I}[i \in S_{:d} \wedge i \in T_{:d}] \frac{p^{d-1}}{d}\end{aligned} \tag{5}$$

We introduce the value M_i, which represents the *effective rank* of item i. It is defined as the earliest position d such that item i intersects both slices $S_{:d}$ and $T_{:d}$. Writing the rank of item i in ranking S as $S_{(i)}$, we define M as follows:

$$M_i = \max\left(S_{(i)}, T_{(i)}\right) \tag{6}$$

By the definition of M_i, we replace the Iverson bracket:

$$\text{RBO}(S,T,p) = (1-p)\sum_{i \in (S \cap T)} \sum_{d=M_i}^{\infty} \frac{p^{d-1}}{d} \tag{7}$$

This can further be rewritten using the closed-form contribution, using the following identity from Webber et al. [13]:

$$\sum_{d=1}^{\infty} \frac{p^d}{d} = \ln \frac{1}{1-p} \tag{8}$$

This operation converts the infinite sum to a finite sum for easier computation.

$$\begin{aligned}\text{RBO}(S,T,p) &= (1-p)\sum_{i \in (S \cap T)} \left[\sum_{d=M_i}^{\infty} \frac{p^{d-1}}{d}\right] \\ &= \frac{1-p}{p}\sum_{i \in (S \cap T)} \left[\sum_{d=1}^{\infty} \frac{p^d}{d} - \sum_{d=1}^{M_i-1} \frac{p^d}{d}\right]\end{aligned} \tag{9}$$

$$= \frac{1-p}{p} \sum_{i \in (S \cap T)} \left[\ln \frac{1}{1-p} - \sum_{d=1}^{M_i - 1} \frac{p^d}{d} \right] \tag{10}$$

The result is a sum over all the items in the intersection of S and T. An individual item contribution C_i for item i can be expressed as follows:

$$C_i(S, T, p) = \frac{1-p}{p} \left[\ln \frac{1}{1-p} - \sum_{d=1}^{M_i - 1} \frac{p^d}{d} \right] \tag{11}$$

Conversely, RBO is the sum of its item contributions.

$$\mathrm{RBO}(S, T, p) = \sum_{i \in (S \cap T)} C_i \tag{12}$$

Excluding the p constant, C_i depends on one value only: the effective rank, $M_i \in \mathbb{Z}^+$. We will define a set of constants K_n for $n \in \mathbb{Z}^+$ which give the contribution of any item with effective rank $M_i = n$.

$$K_n = \frac{1-p}{p} \left[\ln \frac{1}{1-p} - \sum_{d=1}^{n-1} \frac{p^d}{d} \right] \tag{13}$$

$$\forall i \quad M_i = n \;\leftrightarrow\; C_i = K_n \tag{14}$$

To illustrate that each K_n is a constant given p, Figure 4 shows K_n for three common values of the p parameter.

3.2 Marginal Probability of Item Contribution

We will determine the probability mass function of C_i. As we have shown that the contribution C_i is determined by the effective rank M_i, we will begin by deriving the PMF of the latter.

In Equation 6, we have defined $M_i = \max\left(S_{(i)}, T_{(i)}\right)$. For rankings S and T with no ties, $S_{(i)}$ and $T_{(i)}$ give a constant rank for each item i.

p	K_1	K_2	K_3	K_4	K_5	K_6	K_7	...
0.8	.402	.202	.122	.080	.054	.038	.027	
0.85	.335	.185	.121	.085	.062	.046	.035	
0.9	.256	.156	.111	.084	.066	.052	.043	

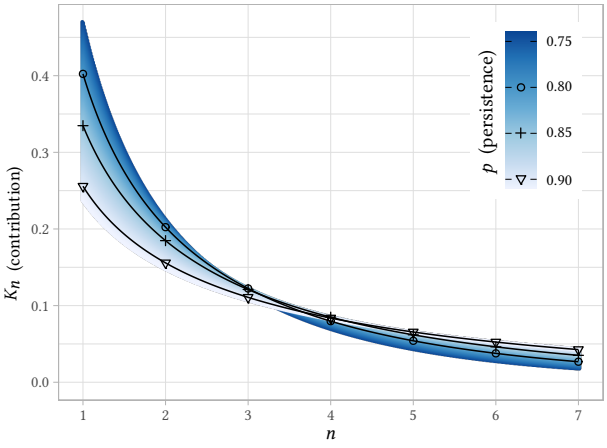

Figure 4: **Item contribution K_n for common p values.** For any item i, $C_i = K_n$ iff the effective rank $M_i = n$.

If we introduce ties, we can model $S_{(i)}$ and $T_{(i)}$ as discrete random variables. The sample space corresponds to all permutations of breaking ties in S and T, as in Figure 3. The arrangement of ties in S and T is independent and all permutations are equally likely. It follows that for a single item i, the RVs $S_{(i)}$ and $T_{(i)}$ denoting its position in S and T are independent of each other. Each follows a discrete uniform PMF covering the ranks i can take (the positions covered by its tie). We use notation where $S_{(i)}^{\text{lower}}$ corresponds to the lowest possible rank i could take when ties are broken in S.

$$\begin{aligned} S_{(i)} &\sim \mathcal{U}\left(S_{(i)}^{\text{lower}}, S_{(i)}^{\text{upper}}\right) \\ T_{(i)} &\sim \mathcal{U}\left(T_{(i)}^{\text{lower}}, T_{(i)}^{\text{upper}}\right) \end{aligned} \tag{15}$$

Figure 5 shows the intuition for one example ranking.

Computing the PMF for M_i is possible from the independent uniform distributions S_i and T_i. Combining the two random variables in this way is shown in Figure 6:

$$\begin{aligned} M_i &= \max\left(S_{(i)}, T_{(i)}\right) \\ &\sim \max\left[\mathcal{U}\left(S_{(i)}^{\text{lower}}, S_{(i)}^{\text{upper}}\right), \mathcal{U}\left(T_{(i)}^{\text{lower}}, T_{(i)}^{\text{upper}}\right)\right] \end{aligned} \tag{16}$$

The result can also be written as a standalone probability mass function using the definition of max. The derivation of this PMF is Equation 18, with an example in Figure 7. The function is expressed as a summation for clarity, but can also be rewritten for $O(1)$ computation.

$$\begin{aligned} \mathrm{P}[M_i = d] &= \mathrm{P}\left[\max\left(S_{(i)}, T_{(i)}\right) = d\right] \\ &= \mathrm{P}\left[S_{(i)} < d \wedge T_{(i)} = d\right] + \\ &\quad \mathrm{P}\left[S_{(i)} = d \wedge T_{(i)} < d\right] + \\ &\quad \mathrm{P}\left[S_{(i)} = d \wedge T_{(i)} = d\right] \end{aligned} \tag{17}$$

S	$=$	$<$	A	$(B$	C	$D)$	E	F	$>$
$\mathrm{P}\left[S_{(A)}{=}d\right]$	$=$		1	0	0	0	0	0	...
$\mathrm{P}\left[S_{(B)}{=}d\right]$	$=$		0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	...
$\mathrm{P}\left[S_{(C)}{=}d\right]$	$=$		0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	...

Figure 5: **$S_{(i)}$ is a discrete uniform random variable.** Item i can take all positions in its tie with equal probability.

S	$=$	$<$	A	$(B$	C	$D)$	E	F	$>$
$\mathrm{P}\left[S_{(B)}{=}d\right]$	$=$		0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	...
T	$=$	$<$	$(B$	E	C	$F)$	A	D	$>$
$\mathrm{P}\left[T_{(B)}{=}d\right]$	$=$		$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	...
$\mathrm{P}[M_B{=}d]$	$=$		0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	0	0	...

Figure 6: **$S_{(i)}$ and $T_{(i)}$ are indep. and determine M_i.** M_i represents item the 'effective rank' of i in $S \cap T$.

$$S = < A \ (B \ C \ D) \ E \ F >$$
$$T = < (B \ E \ C \ F) \ A \ D >$$

inside: M_B

$S_{(B)}$ \ $T_{(B)}$	**1**	**2**	**3**	**4**
2	2	2	3	4
3	3	3	3	4
4	4	4	4	4

$$\mathrm{P}\left[M_{(B)} = 2\right] = \tfrac{1}{6}$$
$$\mathrm{P}\left[M_{(B)} = 3\right] = \tfrac{2}{6}$$
$$\mathrm{P}\left[M_{(B)} = 4\right] = \tfrac{1}{2}$$

Figure 7: M_i **is the maximum of** $S_{(i)}$ **and** $T_{(i)}$**.** S and T are the same rankings as in Figure 6.

$$= \sum_{j=1}^{d-1} \left(\mathrm{P}\left[S_{(i)} = d\right] \mathrm{P}\left[T_{(i)} = j\right] + \mathrm{P}\left[T_{(i)} = d\right] \mathrm{P}\left[S_{(i)} = j\right] \right) + \mathrm{P}\left[S_{(i)} = d\right] \times \mathrm{P}\left[T_{(i)} = d\right] \tag{18}$$

In Equation 18, we have obtained the probability mass function of M_i. Recalling the definition in Equation 14, $C_i = K_n$ if and only if $M_i = n$. This one-to-one mapping gives a PMF for C_i.

$$\mathrm{P}[C_i = K_n] = P[M_i = n] \tag{19}$$

3.3 Dependence of the Item Contributions

It may seem that the uncertainty distribution due to ties for RBO can trivially be obtained from the contribution distributions C_i for $i \in S \cap T$. This would be the case if the RVs were independent.

If discrete RVs X and Y are independent, this law holds:

$$\forall x, y \in \mathbb{Z}^2 : \quad \mathrm{P}[X = x, Y = y] = \mathrm{P}[X = x]\mathrm{P}[Y = y] \tag{20}$$

We can use the following counterexample to show that M_i and M_j for two items i and j are not independent RVs.

$$S = T = < (A, B) >$$
$$\mathrm{P}[M_A = 1] = \mathrm{P}[M_B = 1] = \frac{1}{4} \tag{21}$$
$$\mathrm{P}[M_A = 1]\mathrm{P}[M_B = 1] = \frac{1}{16}$$

Note that by definition, $M_i = 1$ only if $S_{(i)} \leq 1 \wedge T_{(i)} \leq 1$. However, this would require items A and B to be first in both rankings. There is no untied pair S' such that $S'_{(A)} \leq 1 \wedge S'_{(B)} \leq 1$. Thus:

$$\mathrm{P}[M_A = 1 \wedge M_B = 1] = 0$$
$$\mathrm{P}[M_A = 1 \wedge M_B = 1] \neq \mathrm{P}[M_A = 1]P[M_B = 1] \tag{22}$$

By the one-to-one mapping in Equation 19, the item contributions C_i and C_j for $i \neq j$ are thus also not independent. We have obtained their marginal PMFs.

4 UNCERTAINTY IN THE CONTRIBUTION SUM

For this step, we treat each PMF of item contribution as independent. While this assumption does not hold, as shown in Section 3.3, this paper will demonstrate that reconstructing RBO with the independence assumption can be an *efficient estimate* of the distribution of uncertainty due to ties.

First, we will write a discrete PMF as a list of pairs $(x, P[x])$, each pair being an outcome and a nonzero probability such that $x \in \mathbb{R}$, $\mathrm{P}[x] \in [0, 1]$. The total probability is 1.

$$X \text{ is a PMF } \rightarrow \sum_{(x, P[x]) \in X} P[x] = 1 \tag{23}$$

Next, we will use the $\otimes$ symbol to represent the summing convolution. This convolution has the function signature shown in Equation 24: it takes two probability mass functions X and Y, each represented as a pair-list, and outputs all possibilities of the sum of their values under the assumption that X and Y represent independent RVs. This is defined in Equation 25.

$$\otimes : (\mathbb{R}, \mathbb{R})^m, (\mathbb{R}, \mathbb{R})^n \rightarrow (\mathbb{R}, \mathbb{R})^q \tag{24}$$

$$\otimes (X, Y) = \forall x \in X : \forall y \in Y : (x + y, \ P[x]P[y]) \tag{25}$$

We will now apply $\otimes$ to iteratively evaluate the uncertainty distribution of RBO. In Algorithm 1, we use the $\otimes$ operation to add the marginal PMF of a single item contribution i to a PMF representing the intermediate sum of evaluated items. An example step in this procedure is shown in Figure 8. We obtain an approximate PMF for the RBO itself by sequentially adding the marginal contributions of items $\{A, B, ...\} \in S \cap T$.

$$C_A \otimes C_B \otimes ... \approx \mathrm{RBO}(S, T, p) \tag{26}$$

Marginal item PMF C_C

c	$\mathrm{P}[C_C] = c$
0.02	2/3
0.03	1/3

Approx. PMF of $C_A + C_B$

c	$\mathrm{P}[C_A + C_B] \approx c$
0.4	2/5
0.5	3/5

$\otimes$

Approx. PMF of $C_A + C_B + C_C$

c	$\mathrm{P}[C_A + C_B + C_C] \approx c$
0.42	4/15
0.43	2/15
0.52	6/15
0.53	3/15

Figure 8: **The summing convolution, written as** $\otimes$**.** Note that $X + Y = X \otimes Y$ only holds for indep. PMFs.

4.1 An Approach to Reducing Covariance

Unfortunately, estimating RBO using Algorithm 1 and the convolution function from Equation 25 is not accurate. The problem with this approach lies in the independence assumption. Section 3.3 shows that the assumption can trivially be broken, and a basic convolution makes no effort to deal with the covariance.

On the other hand, eliminating covariance by finding the joint distribution of item contributions C_i is equivalent to searching the entire permutation space. The problem is related to tensor estimation of joint distributions, a well-known dimensionality reduction problem in Machine Learning, however, this is a complex approach that would require two- or three-dimensional pairwise marginal probabilities [5].

procedure ESTIMATERBO(S, T, p)

▷ Initialize the PMF X for length 0.
▷ This stores a list of RBO-probability pairs.
▷ For an empty ranking pair, $\mathrm{P}[X = 0] = 1$.
$X \leftarrow \{(0, 1)\}$

for each item i in $S \cap T$ **do**
  ▷ Obtain the contribution PMF C_i for item i.
  ▷ The function is given in Equation 19.
  $C_i \leftarrow$ CONTRIBUTION(i, S, T, p)

  ▷ Perform the convolution.
  $X \leftarrow X \otimes C_i$
end

▷ Return the resultant PMF.
▷ X approximates the uncertainty distribution.
return X

end

Algorithm 1: **Estimation by convolution**

This part will describe a modular approach of solving this problem. We will continue using iterative convolution, emulating the covariance by adding 'rules' into the $\otimes$ function. First, we will modify the algorithm with a new data structure, a 'bitstring', to store intermediate values instead of raw numeric contribution. Next, we will show that this scheme makes it easy to apply rules and achieve better results.

Algorithm 1 maps the PMF for the first effective index of an item M_i to contribution value C_i, then convolves the values. We will instead convolve the indices and map them to a value later. From the one-to-one mapping of M_i to C_i (Equation 19), we can show that for an arbitrary set of items I, the sum of contributions C_i can be calculated from the set of their effective ranks M_i. We will store this instead of the value.

$$\sum_{i \in I} C_i = \sum_{i \in I} K_{M_i} \tag{27}$$

Now, let us define the 'bitstring', which will encode the set of M_i for an item set I as a frequency table of infinite length. The kth digit of a bitstring refers to the number of occurrences of k. Equation 28 shows an example of this.

$$\begin{gathered}(a \times \{1\}) \cup (b \times \{2\}) \cup (c \times \{3\}) \cup \ldots = \text{"}abc\ldots\text{"} \\ \{1, 2, 4, 4\} = \text{"}110200\ldots\text{"}\end{gathered} \tag{28}$$

Let us extend our definition of the contribution constants K_n to a bitstring. Using function notation, $K_n : n \in \mathbb{Z}^+ \to \mathbb{R}$ gives the contribution of a single item where $M_i = n$. Similarly, we extend this to multiple items. Let $K_b : b \in \mathbb{B}^n \to \mathbb{R}$, where $\mathbb{B}$ is the set of all bitstrings. K_b is the contribution sum of all M_i values encoded in the bitstring: the dth digit represents the frequency of items where $M_i = d$. An example is shown in Equation 30.

$$b \text{ is a bitstring} \quad \rightarrow \quad K_b = \sum_{d}^{\infty} b_d K_d \tag{29}$$

$$K_{\text{"}110200\ldots\text{"}} = K_1 + K_2 + 2K_4 \tag{30}$$

The new formulation of the algorithm will perform this step at the very end while returning the value. Bitstrings will be used to represent a sum of item contributions. The convolution function will be written accordingly (Equation 31).

$$\otimes : (\mathbb{B}, \mathbb{R})^m, (\mathbb{B}, \mathbb{R})^n \to (\mathbb{B}, \mathbb{R})^q \tag{31}$$

A single item contribution, in bitstring form, is a a '1' at digit d, where $d = M_i$, and '0' otherwise.

$$M_i = n \ \rightarrow \ \text{bitstring of item } i = \begin{cases} 1 \text{ at digit } n \\ 0 \text{ otherwise} \end{cases} \tag{32}$$

With a basic $\otimes$ function such as described by Equation 25, using bitstrings will not change the final estimation value. With the PMF in the bitstring formulation, however, we can apply 'rules' by modifying the $\otimes$ function.

4.2 Applying Rules

We will define a set of rules that our bitstring must satisfy in order to reduce the impact of covariance on our estimation. The three rules that follow are limitations on possible permutations from the definition of a ranking.

Rule 1. The sum of digits in a bitstring is equal to the number of contributing items.

Rule 1 states that in a valid permutation, all items contributing to RBO must have exactly one value of M_i. Since the bitstring is a frequency table representation of contributing M_i values, the frequencies must count to the number of contributing items.

Rule 2. The cumulative sum of digits up to index d is less than or equal to d.

Rule 2 states that the agreement of a ranking at depth d cannot be greater than d. The agreement is equivalent to the frequency of items where $M_i \leq d$, or the cumulative sum of the bitstring.

Rule 3. No digit is greater than 2.

Rule 3 is derived from the $\mathrm{RBO}^{\mathrm{high}}$ algorithm from Corsi and Urbano [4]. It states that there is no valid permutation such that depth d adds $n > 2$ items to the agreement: this would require 3 or more items to be at rank d, which is not possible.

Let us define a modified 'culling convolution', $\otimes_{\mathrm{cull}}$, which works analogously to $\otimes$, only with the additional step of removing bitstrings from the PMF that do not satisfy Rules 1, 2, and 3. The total probability is then renormalised[1] to 1. This algorithm is shown in Algorithm 2. The effects of "culling" are twofold: first, the estimate PMF is a more accurate representation of the actual distribution. Second, the resultant PMF has significantly fewer cases and thus takes up less memory.

Figure 9 shows an example ranking pair where the difference can be seen: while the basic convolution produces two values outside of the actual distribution of permutations, the culling convolution removes these 'impossible' cases and more accurately represents the true distribution.

While the depicted example of $S = <(A \ \ B \ \ C)>$, $T = <(A \ \ B) \ \ C>$ is clearly in favor of the culling convolution, quantifying this improvement is of interest. Kullback-Leibler divergence and similar commonly used statistical distances are unsuitable for this comparison due to the disjoint support: cases

[1] This is a naive approach that does not preserve the marginal probabilities. Its implications will be discussed in Section 6.

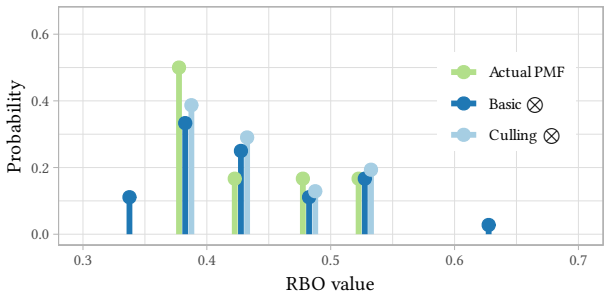

Figure 9: **Basic and culling ⊗ on an example ranking pair**. Example: $S = <(A\ \ B\ \ C)>,\ T = <(A\ \ B)\ \ C>$

procedure $\otimes_{\text{cull}}(\boldsymbol{X}, \boldsymbol{Y})$

▷ Initialize the result PMF Z.
▷ This is a list of bitstring-probability pairs.
$Z \leftarrow \{(\text{"000..."}, 1)\}$

for each pair $(b_X, \mathrm{P}[b_X])$ in X **do**
  for each pair $(b_Y, \mathrm{P}[b_Y])$ in Y **do**
    if $b_X + b_Y$ satisfies rules **then**
      $Z \leftarrow Z + (b_X + b_Y, \mathrm{P}[b_X]\mathrm{P}[b_Y])$
    end
  end
end

▷ Normalize probabilities in Z to 1.
$Z \leftarrow$ **Normalize**(Z)
return RBO, p_{RBO}

end

Algorithm 2: **The modified convolution,** $\otimes_{\text{cull}}$

in the estimate PMF may not be present in the real distribution. This is also the case for the culling convolution, as not all impossible permutations have been eliminated by our three simple rules. The measure suitable for this comparison is the Earth Mover's Distance (Wasserstein metric W_1) [12], which allows for effective comparison of weighted discrete distributions with disjoint support. In the example in Figure 9, $\otimes_{\text{cull}}$ results in an improvement of the EM distance from 0.0131 to 0.0069.

5 EXPERIMENTAL EVALUATION

A systematic evaluation on synthetic and real-world data was conducted to evaluate the method's performance. To be a successful implementation, the estimator must perform well for a diverse range of cases; to that end, synthetic TREC-like data was used.

5.1 Simulation Conditions

We conducted large-scale empirical evaluation of the estimator on synthetic data. The aim of this evaluation is to test the method on a large number of randomized rankings, establish a picture of its baseline efficiency, and reason about situations where it performs better or worse. This section will explain the simulation methodology and introduce the results. A discussion will follow in Section 6.

The test data was constructed using the `simulate.R` program by Corsi and Urbano [3]. This program creates a pair of tied rankings S and T with parameters $|S|, |T|, N$, and τ, where $N > |S|, |T|$ is the number of items in the sample space and τ is Kendall's correlation of the randomly generated scores used to construct both rankings. The simulation code selects a uniformly distributed random τ for every ranking pair, producing a variety of correlated and uncorrelated rankings.

We generated rankings in four categories of ranking "size", ranging from 6 to 29 items. Table 2 shows the values of $|S|$, $|T|$, and N, and the number of rankings generated by category. Additionally, a permutation cap of 10^5 was applied: ranking pairs where the number of possible tie arrangements was greater than 10^5 was not included. This threshold was set to exclude those pairs for which the true distribution of possible values could not be computed in a reasonable time; the impact of this decision is discussed in Section 6.

The probability mass function of the estimation method described in this paper was generated for each ranking alongside the actual PMF given by evaluating all permutations. The parameter p was set to $p = 0.9$. Quantile data was computed for each PMF without interpolation. Mean and variance data was computed for each PMF. Finally, Earth Mover's Distance (EMD) was computed to compare the two PMFs. The inclusion of this statistical distance is discussed in Section 4.2.

5.2 Evaluation Results

Table 1 is a summary of data collected during the empirical evaluation. The mean Earth Mover's Distance for all ranking sizes was found to be 1.98×10^{-3} with some variation between ranking sizes. We consider this to be an indicator that the distributional similarity is high and that the estimate distribution will yield reliable results in for a diverse set of rankings.

The "S" size rankings were observed to be more frequently subject to estimation errors, with a notably above-average mean Earth Mover's Distance at 4.69×10^{-3}, or ≈ 2.4 times the mean for all. This is also visible in the mean squared error of the variance estimator, which is 1.86×10^{-7} as compared to 2.60×10^{-8} for rankings of size "XL".

Figure 10 shows error in estimating the mean of the distribution. The plot shows that the estimate PMF is well-centered around the true distribution for all values of RBO, with no noticeable correlation with the value. An enlarged plot is included with a linear color scale to show the alignment of the x and y axes in the plot.

Figure 11 shows the error in estimating the 2.5th and 97.5th percentile of the data. The data is adjusted for equal representation of each ranking size. We note that smaller rankings are more frequently represented on the tails of the histograms, indicating once again that the error is higher for this category.

Finally, all probability mass functions generated exceeded the most extreme values for the true PMF; no estimate for the 0th or 100th quantile was closer to the mean than the true value. Therefore, the estimate distribution can be said to always overestimate the uncertainty.

6 DISCUSSION

We introduced the evaluation results, which validate that the iterative convolution method is a viable method for estimation of uncertainty due to ties in Rank-Biased Overlap. The data shows that the estimate probability mass function obtained exhibits consistent performance for a range of ranking lengths, item counts, and RBO values.

size	S	M	L	XL	All sizes
Mean EMD†	4.69×10^{-3}	2.82×10^{-3}	1.75×10^{-3}	1.22×10^{-3}	1.98×10^{-3}
MSE Mean	3.66×10^{-5}	1.35×10^{-5}	6.73×10^{-6}	4.12×10^{-6}	8.66×10^{-6}
MSE Var	1.86×10^{-7}	7.52×10^{-8}	3.89×10^{-8}	2.60×10^{-8}	4.91×10^{-8}
MSE q0	1.50×10^{-4}	1.13×10^{-4}	6.50×10^{-5}	4.40×10^{-5}	7.43×10^{-5}
MSE q0.025	2.41×10^{-4}	9.30×10^{-5}	5.15×10^{-5}	3.45×10^{-5}	6.33×10^{-5}
MSE q0.05	5.17×10^{-5}	3.57×10^{-5}	2.50×10^{-5}	1.97×10^{-5}	2.72×10^{-5}
MSE q0.95	1.93×10^{-4}	7.65×10^{-5}	4.35×10^{-5}	3.05×10^{-5}	5.29×10^{-5}
MSE q1	4.06×10^{-4}	1.52×10^{-4}	7.63×10^{-5}	4.68×10^{-5}	9.77×10^{-5}

Table 1: **Summary of evaluation data.**

Ranking size	$\lvert S\rvert = \lvert T\rvert$	N	count
S	$\mathcal{U}(6, 11)$	12	5 000
M	$\mathcal{U}(12, 17)$	18	35 000
L	$\mathcal{U}(18, 23)$	24	75 000
XL	$\mathcal{U}(24, 29)$	30	35 000

Table 2: **Rankings sizes for the experimental evaluation.**

Further, the results identify key characteristics of the obtained distribution: namely, that its minimum and maximum values are at least as extreme as the true distribution, and similarly that the an arbitrary quantile of the estimate distribution is biased away from its center.

These findings do not compare the accuracy of the estimate to any baseline, and as such we consider this study to be an exploratory evaluation. A more rigorous comparison including real TREC or other IR data would be suitable if the method were to be integrated in an R package and used in the field.

6.1 Method of the Empirical Evaluation

First, we consider the impact of the 100k permutation limit imposed on the rankings evaluated against. We justify the inclusion of this limit, because it is consistent with what similar works have used in their experiments [3, 4]. Additionally, with Rank-Biased Overlap being a top-weighed measure, a high permutation count is not necessary to obtain higher uncertainty, and the results could therefore test across the spectrum of uncertainty magnitude.

The experimental evaluation raises questions about other possibilities of sampling the distribution of all tied rankings.

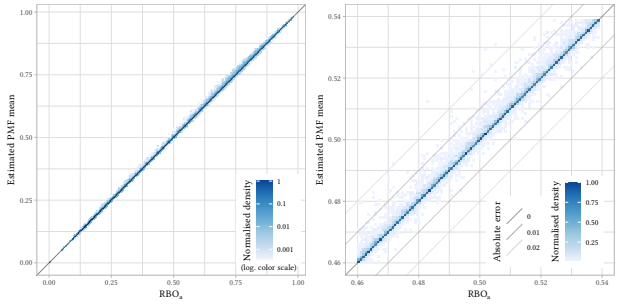

Figure 10: **Estimator performance on mean (enlarged)** Binned data. Density is logarithmic on the left graph.

†Earth Mover's Distance

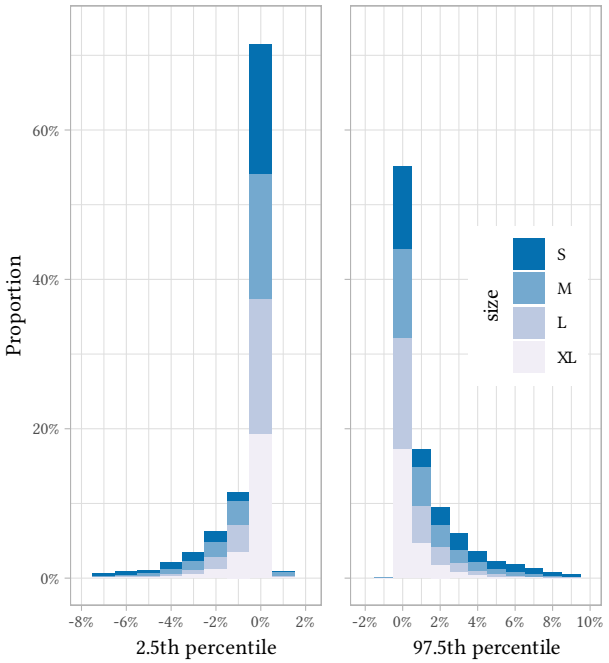

Figure 11: **CI estimation error for different ranking sizes.** Smaller rankings tend to have larger relative error.

The possibility of uniformly sampling this space was discarded due to the scope of this project, and we chose to use ranking simulation code instead. While the results showed a relation between ranking size and the error of the uncertainty estimator, we considered that this could be due to the sampling scheme, which assigned rankings based on scores for which the correlation distribution was uniform (Section 5). For the purposes of this research, it would be more useful to sample against a uniform distribution of the RBO uncertainty range.

7 CONCLUSION AND FUTURE WORK

Overall, the evaluation data suggests that this estimation technique is a strong candidate for efficient estimation of the uncertainty distribution due to ties. We suggest that future work continues to pursue efficient estimation via the modified convolution method. We believe that it is a viable solution to the dimensionality problem posed by the size of the tie distribution and should be developed further.

Our recommendation to the field of IR and RecSys with regards to uncertainty in tied data is to continue using the method of Corsi and Urbano [4] for determining bounds unless arbitrary quantile estimation would give a significant advantage. In that case, this method would provide finer-grained results.

The method we proposed further satisfies the requirements of being able to produce variable confidence interval-like bounds, and is more efficient than searching the tie permutation space. For future work, it is crucial that the impact of simplifying the estimate be explored further. We have defined an estimation method that is iterative and allows for truncation for faster evaluation. However, the speed advantage was not demonstrated in the experimental evaluation due to limitations of scope. This should be considered when applying methods like iterative convolution.

ACKNOWLEDGEMENTS

Work facilitated by computational resources of the DelftBlue cluster at TU Delft. In(de)finite thanks to the feedback and support of Prof. Urbano.

RESPONSIBLE RESEARCH

This section lists the efforts made to respect the academic principles of responsibility and integrity.

The effect of this research and its application was considered. Caution should be taken when deploying any statistical analysis to preserve the privacy and dignity of individuals; this work does not affect these obligations. This paper advocates for accurate knowledge of uncertainty in statistics; the inclusion of uncertainty using the methods described here can prevent harmful misrepresentation.

The experimental evaluation was made possible using the public R sources of Corsi and Urbano [3] and [4]. The repositories for the simulation code are in the respective publications.

The results of this experiment are fully reproducible using R code available in the following Git repository: https://github.com/lchladek/RBO. All operations involving random values are reproducible, as the random seeds have been preserved.

Finally, Generative AI was not used in any capacity.

BIBLIOGRAPHY

[1] Mayer Alvo and Philip L.H. Yu. 2014. *Statistical Methods for Ranking Data*. Springer New York. https://doi.org/10.1007/978-1-4939-1471-5

[2] Guillaume Cabanac, Gilles Hubert, Mohand Boughanem, and Claude Chrisment. 2010. Tie-Breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. In *Multilingual and Multimodal Information Access Evaluation*. Springer Berlin Heidelberg, 112–123. https://doi.org/10.1007/978-3-642-15998-5_13

[3] 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 2024)*, July 2024. Association for Computing Machinery, 251–260. https://doi.org/10.1145/3626772.3657700

[4] Matteo Corsi and Julián Urbano. 2024. How do Ties Affect the Uncertainty in Rank-Biased Overlap?. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024)*, December 2024. Association for Computing Machinery, 125–134. https://doi.org/10.1145/3673791.3698422

[5] Shahana Ibrahim and Xiao Fu. 2020. Recovering Joint Probability of Discrete Random Variables from Pairwise Marginals. (2020). https://doi.org/10.48550/ARXIV.2006.16912

[6] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1–2 (June 1938), 81–93. https://doi.org/10.1093/biomet/30.1-2.81

[7] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251. https://doi.org/10.1093/biomet/33.3.239

[8] Jimmy Lin and Peilin Yang. 2019. The Impact of Score Ties on Repeatability in Document Ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*, July 2019. ACM, 1125–1128. https://doi.org/10.1145/3331184.3331339

[9] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* 27, 1 (December 2008), 1–27. https://doi.org/10.1145/1416950.1416952

[10] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. 2013. Overview of the TREC 2011 Microblog Track. Retrieved from https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=914332

[11] C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72–101. Retrieved June 2, 2025 from http://www.jstor.org/stable/1412159

[12] Simon Urbanek and Yossi Rubner. 2023. emdist: Earth Mover's Distance. R package. Retrieved from https://cran.r-project.org/package=emdist

[13] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *Transactions on Information Systems* 28, 4 (November 2010), 1–38. https://doi.org/10.1145/1852102.1852106

[14] Ziying Yang, Alistair Moffat, and Andrew Turpin. 2016. How Precise Does Document Scoring Need to Be?. In *Information Retrieval Technology*. Springer International Publishing, 279–291. https://doi.org/10.1007/978-3-319-48051-0_21