Average Rank-Biased Overlap between independent rankings

Revealing average benchmarks: An Empirical Investigation

Mark Dragnev[1]

Supervisor(s): Julian Urbano[1]**, Matteo Corsi**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Final project course: CSE3000 Research Project
Thesis committee: Julian Urbano, Matteo Corsi, Mathijs Molenaar

Abstract

Rankings play a crucial role in various contexts but often exhibit incompleteness, top-weightedness, and indefiniteness. Comparing rankings can reveal underlying similarities, yet traditional correlation coefficients like Kendall's tau do not adequately address these complexities. Rank-Biased Overlap (RBO) addresses these challenges by accommodating differences in rank length, appropriately weighting ranks, and minimizing data assumptions. This paper investigates the average Rank-Biased Overlap (RBO) between independent rankings, addressing the need for clearly indicated reference values similar to those of correlation coefficients. Our study explores how the expected RBO changes with varying p-parameters, prefix lengths, and degrees of conjointness between domains. To facilitate this analysis, an algorithm is developed that performs extensive simulations across different values of p, list and domain sizes. By analyzing the simulation results, trends are provided in the average RBO between independent rankings based on these varying parameters and establish relevant reference values. This study focuses on scenarios where prefixes are of the same length and there are no ties in the rankings.

1 Introduction

Rankings are pervasive in our daily lives, from sports and financial markets to academic publications and search engine results. These rankings are often incomplete, since they do not cover all elements from their origin domain. Furthermore, they have a top-weighted nature, where the head of the list has more weight than the tail. For example, the difference between the first and the second result returned by a search engine query matters substantially more than any other positioned elements further down in depth. In addition, these rankings are also indefinite. Usually only prefixes are taken into account, where one covers a small ratio of the whole ranking. The size of this prefix can be random, since listing more items offers diminishing returns compared to the increasing cost.

Comparing rankings is common practice, aiming to infer similarities between the underlying processes which produce them. For instance, when two different search engines produce results for the same query, finding the similarity between the two lists might tell us how different those engines are.

Existing rank similarity measures primarily address (non)conjoint and (un)weighted rankings. Such examples are Bar-Ilan's ρ [1] and Buckley's AnchorMAP [2]. Kendal's τ [3] requires that two rankings both come from fully-conjoint domains, meaning that elements from both domains are the same. The measure is also unweighted, since a misalignment of elements at some depth contributes equally as another misalignment positioned deeper in the list. Yilmaz et al.[5] later improved it by introducing τ_{AP}, which is top-weighted.

Those measures however do not tackle the indefinite nature of rankings. A new similarity measure has been proposed by Webber et al. [4], called Rank-Biased Overlap. It has proven to accommodate differences in length, while also assigning proper weight to ranks, and minimizing assumptions about the data.

When using Rank-Biased Overlap to measure data, we often face uncertainty about what the results truly signify due to the absence of the concept of a reference value in literature, with regards to RBO itself. For instance, if we compute RBO between two rankings that represent the favourite sport teams of two friends, then it is challenging to draw meaningful conclusions from that result. Without a reference value, we cannot determine if the measured overlap indicates a significant similarity in their preferences or if it is merely coincidental.

In order to build the intuition behind this reference value, we have to start with what factors influence one RBO evaluation. These factors include the $p-$ parameter that is used, length of the rankings, domains and their degree of conjointness. By examining them, we can conceptualize the reference value as the expected RBO between independent rankings, given these specific properties. This average value is essential for assessing the significance of any observed scores with real data. Without such a benchmark, it is difficult to discern whether the measured overlap indicates true similarity or random coincidence.

In contrast, for Kendall's τ [3], which is a correlation coefficient, the reference value is clearly defined as 0, indicating no correlation. This established benchmark allows for straightforward interpretation of the correlation measurement. Similarly, establishing a reference value for RBO is crucial for providing context and meaning to its measurements, thus enhancing their utility in various applications.

The question that this research strives to answer is *What is the average Rank-Biased Overlap between independent rankings?* In order to answer this, the following sub-questions have to be answered:

- What is the expected RBO between independent rankings when the p-parameter changes?
- What is the expected RBO between independent rankings when their prefix length changes?
- What is the expected RBO between independent rankings when their degree of conjointness changes?

In more detail, the behavior of the expected RBO between independent rankings will be investigated as a function of these variables. Subsequently, insights will be derived to determine, or at least approximate, a reference value based on the observed patterns across different variable configurations. This work only focuses on the assumption that two rankings are of equal size, and that they do not contain any ties.

Section 2 gives some background for RBO and tries to built some fundamental knowledge, so that the rest of the paper can be followed. Section 3 gives detailed insights about how a dataset is build, based on different configurations when evaluating two rankings, such as the chosen p parameter, the lengths of the lists, the sizes of the domains, and their degree of conjointness. Section 4 analyses the generated dataset, and explores the trend of the expected RBO, based on the different settings, which mainly represent the three sub-questions.

Section 5 is about Responsible Research, and how the algorithm and dataset can be used in follow-up studies. Section 6 summarizes the results of the alanisys from Section 4, and the last Section describes future work.

2 Background

This section delves into an in-depth examination of *Rank-Biased Overlap* and its further derived formulas. Moreover, this chapter includes an analysis of other correlation measures, elucidating their distinct properties and limitations within the context of ranking evaluation. Specifically, it will be how these measures establish a benchmark or a reference value that provides context and meaning to their evaluations, contrasting this with the challenge faced by RBO in lacking a similar standardized interpretation.

2.1 Kendall's τ

Kendall's τ [3] provides an intuitive method to compare two rankings by counting concordant, P, and discordant, Q, pairs among their items. A concordant pair of items is one where the two elements have the same order in both rankings. The formula

$$\tau = \frac{P - Q}{P + Q} \tag{1}$$

quantifies the degree of similarity between rankings, ranging from -1 (complete inverse order) to 1 (identical order), with 0 indicating equal likelihood of concordance and discordance. This measure has a probabilistic interpretation, where τ reflects the correlation between rankings, and can be used for inference on population similarity through methods like confidence intervals and hypothesis testing.

Correlation concepts are inadequate for incomplete rankings where only the top portion is visible, as random and negatively correlated rankings appear similar without common elements in the observed top. Valid similarity measures for incomplete rankings must assume strong correlation, particularly at the top ranks.

2.2 Rank-Biased Overlap

Let S and T be two infinite lists. For any $i \in \mathbb{N}$, S_i and T_i represent the elements at position i. Webbet et al. [4] gives the concept of an *overlap*:

$$X_{S,T,d} = |S_{:d} \cap T_{:d}| \tag{2}$$

which is the size of the intersection of the first d elements between S and T. Their *agreement* is then

$$A_{S,T,d} = \frac{X_{S,T,d}}{d} \tag{3}$$

and represents the proportion of the active items up to depth d. For example, let $S = \{b, c, a, e, f\}$ and $T = \{a, c, d, z, x\}$. Their agreement at depth 3 is $A_{S,T,3} = 2/3$.

Finally, *Rank-Biased Overlap* between S and T is defined as the infinite and weighted sum of the agreements at all depths:

$$RBO(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} A_d \tag{4}$$

The parameter p, termed as persistence, governs the rate of weight decay in rankings: smaller values assign significantly higher weights to top-ranked elements relative to lower-ranked ones, whereas larger p values attenuate this decay, resulting in more uniform weighting across ranks. For detailed treatments of ties and uneven lists, refer to [4].

2.3 Weight of a prefix

In this section, we derive the weight assigned to each rank. Every agreement up to depth d has a weight. It also

$$W_{RBO}(d) = \frac{1-p}{p} \sum_{i=d}^{\infty} \frac{p^i}{i} \tag{5}$$

The weight of a prefix of size d, W_{RBP} is then the sum of the weights of the ranks to that depth

$$W_{RBO}(1:d) = \sum_{j=1}^{d} W_{RBO}(d) = \frac{1-p}{p} \sum_{j=1}^{d} \sum_{i=j}^{\infty} \frac{p^i}{i} \tag{6}$$

After some rearangement, and by using this equality to convert the infinite sum to a finite form:

$$\sum_{i=1}^{\infty} \frac{p^i}{i} = \ln \frac{1}{1-p}, \quad 0 < p < 1 \tag{7}$$

, we get:

$$W_{RBO}(1:d) = 1 - p^{d-1} + \frac{1-p}{p} \times d \times (\ln \frac{1}{1-p} - \sum_{i=1}^{d-1} \frac{p^i}{i}) \tag{8}$$

Generally it can be noted that for a fixed prefix size, its weight, $W_{RBO}(1:d)$, is decreasing when p increases. This comes natural, since as p approaches arbitrarily close to 1, the weights of the elements become arbitrarily flat, and the evaluation becomes arbitrarily deep [4]. As p increases, the elements contained only in the prefix will have a decreasingly smaller importance, compared to the elements that are positioned after the truncated depth. Since the weight of the latter elements, $W_{RBO}(d+1:\infty)$, is defined as $1 - W_{RBO}(1:d)$, the weight of the tail is therefore increasing.

This can be shown by Figure 1. It plots a contour map of $1 - W_{RBO}(1:d)$ for a different values of p and d (depth or also prefix size). For example, for a prefix of size 15 and $p = 0.8$, the tail would have a weight of ~ 0.01. It jumps to ~ 0.07, when $p = 0.9$. For $p = 0.95$, the same weight increases to ~ 0.22. Finally, for $p = 0.99$ and 0.999, $1 - W_{RBO}(1:d)$ is equal to ~ 0.64 and ~ 0.93 respectively.

Weights of prefixes, based on their sizes, and also different p values chosen for evaluation, are provided in Appendix A.

2.4 Extrapolated RBO

The original RBO equation can be used when we want to evaluate the known similarity between two rankings, up until the depth equal to the size of those rankings. However, this score depends on the length and sets lower bound on the full evaluation. Webber et. al [4] define RBO_{MIN} and RBO_{RES}, the tight lower bound and residual uncertainty,

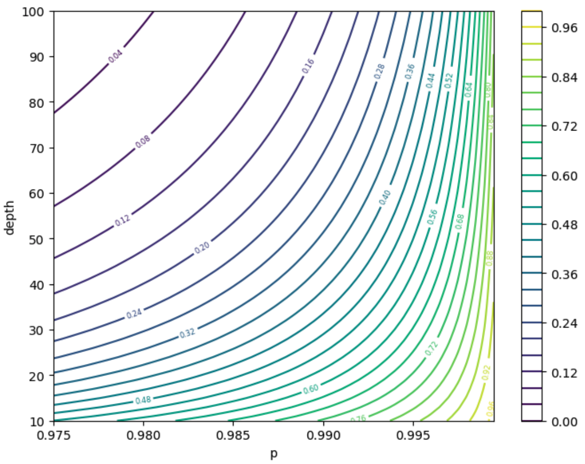


Figure 1: A contour map of different weights of prefixes based on N and P

which is the range between the tight lower and tight upper bound. For more details, refer to [4].

Webber et al.[4] define a formulation for a single RBO score, where the seen agreement in the rankings is assumed to continue on indefinitely. That is, for all $d' > d$, $A_{S,T,d'} = X_{S,T,d}/d$. By assigning it the appropriate weight, and addint it to the base $RBO@k$ score, we get:

$$RBO_{EXT}(S,T,p,k) = \frac{X_k}{k} \times p^k + \frac{1-p}{p} \sum_{d=1}^{k} \frac{X_d}{d} \times p^d \quad (9)$$

It must be noted that Extrapolated RBO depends on the evaluation depth and can either increase or decrease when the prefix gets larger, i.e. it is *not* monotonic. However, larger agreements lead to an increase, and smaller agreements will lead to a decrease in RBO_{EXT} respectively.

3 Dataset

This section describes the algorithm, that is used to run a lot of simulations for different configurations, such as prefix size, p parameter, degree of conjointness between two domains, and their respective sizes.

A single simulation between two independent rankings can be described in the following algorithm. The procedure takes 5 parameters: p, size of the prefix n, desired degree of conjointness between their domains, and their sizes.

To begin with, a concept of a *step* variable must be discussed. The two domains, from which elements will be sampled, have to be constructed in such a way, that their cojnointness matches a required conjointness, passed in the method. The term conjointness gives a value that shows how similar two lists are. If they are fully conjoint, i.e. they have the same elements, $cojnoitness = 1$. When $conjointness = 0$, the two lists have nothing in common.

A good index that can be used to measure the conjointness between two lists is the *Jaccard similarity coefficient*, given by:

$$J(A,B) = \frac{A \cap B}{A \cup B} \quad (10)$$

Let's take two domains $S = \{1, 2, ..., p\}$ and $T = \{p+1, p+2, ..., q\}$, for some numbers $p, q \in \mathbb{N}$, such that $1 < p < q$. Right now, they are disjoint, so their degree of conjointness is 0. We can *shift* the elements of the second domain to the left by some number. We can call this number *step*. Now, the second domain is $T' = \{x \in T \mid x - step\}$. We can further model their conjointness by using Eq. 10, and by using appropriate substitutions, we get the following:

$$J(S,T') = \frac{step}{q - step} \quad (11)$$

After some rearengement, and expessing eq. 11 in terms of *step*, we get:

$$step = \frac{J(S,T') \times q}{J(S,T') + 1} \quad (12)$$

By using this formula, one can discover the desired *step* given some conjointness and two domain sizes. With that, two domains can be constructed, which reflect their similarity properly.

This algorithm is then calculated for different values of p, n, $conjointness$, $size1$ and $size2$. Every simulation is executed $10,000$ in order to provide a good approximation for $E[RBO]$.

Algorithm 1 Average RBO_{Ext}

```
Require: $p$, $n$, $c$, $s1$, $s2$
Ensure: Average RBO value
function AVERAGE($p$, $n$, $c$, $s1$, $s2$)
    $step \leftarrow$ getStep($c$, $s1$, $s2$)
    $D_b = 1 + s1 - step$
    $D_e = s1 + s2 - step$
    S ← {}
    for $i \leftarrow 1$ to $s1$ do
        S ← S ∪ {$i$}
    end for
    T ← {}
    for $i \leftarrow D_b$ to $D_e$ do
        T ← T ∪ {$i$}
    end for
    $sum \leftarrow 0$
    for $i \leftarrow$ 1 to 10000 do
        $s \leftarrow$ sample(S, n)
        $t \leftarrow$ sample(T, n)
        $sum \leftarrow sum + RBO_{EXT}(s,t,p,n)$
    end for
    return $\frac{sum}{10000}$
end function
```

4 Experimental Setup and Results

Section 3 gives the methodology for obtaining the expected RBO_{EXT}, based on different variables, namely p-parameter, prefix size, conjointness and domain size. This section analyzes the obtained results and tries to answer each of the individual subquestions independently.

4.1 Expected RBO when *p*-parameter changes

In order to answer this question, the dataset has to be analyzed in such a way, that one can observe the behavior of the average RBO, with respect to p.

First, let's take for example a single simulation, where two independent rankings of size 10 are continuously sampled from a domain of size 500. The parameter of choice is $p = 0.8$. Over *10,000* iterations, 10 random elements are sampled, and their RBO_{EXT} is computed. The final outcome is the average over all iterations, and the result $E[RBO] \sim 0.0091$ is produced.

The number of iterations is big enough to produce a somewhat accurate number, however, it still gives room for some variance, if executed multiple times. In order to examine this variance, this simulation is independently ran over 500 additional times. A mean of 0.0089 and a standard deviation of 000.31 are produced. The coefficient of variation, i.e. the relative amount of sampling error associated with the estimate, is 3.5%. Figure 2 provides the results of this simulation. Indeed, the histogram resembles a normal distribution.

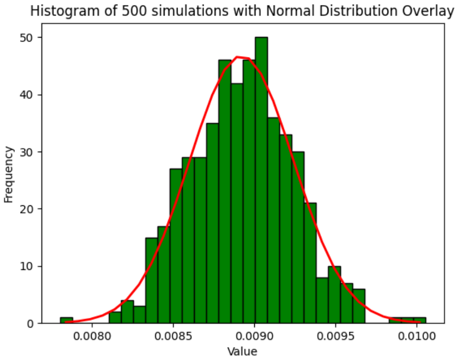


Figure 2: Different values for E[RBO]. $p = 0.8, N = 10, D1 = D2 = 500$.

Now we have to see how the expected RBO changes when p is varied. Table 1 provides this insight. For more expresiveness, the table includes also simulations for prefix sizes of *5, 15 and 20*. The domains that they are sampled from have a fixed size of $D1 = D2 = 500$. For the sake of consistency, the degree of conjointness is fixed at 1.0, which leads to the fact that the two rankings are sampled from the same domain. And again, each produced average is tested *500* times, in order for us to analyze the standard deviation and the coefficient of variance.

It can be observed, in fact, that for any configuration, *10,000* iterations are enough to produce a result, which has a relative error of 5% or less. There is no reason this to be different for the rest of the simulations from the dataset, since they only compute $E[RBO]$ based on different settings, but if tested also for relative error, the latter would be around the same magnitude.

It is important to note that the results in Table 1 are put there to provide well-approximated reference values, however some of them should not be looked at in practise. This is due to the relationship between the size of the prefix and p, which can be examined through the weight of the prefix, defined in Section 2.3.

P	N	Conj	D1	D2	E[RBO]		
					mean	*sd*	*cv*
0.8	5	1	500	500	*0.006721*	*0.00034*	*0.051*
	10				*0.008944*	*0.00031*	*0.035*
	15				*0.009643*	*0.00032*	*0.033*
	20				*0.009878*	*0.0003*	*0.0304*
0.9	5	1	500	500	*0.008169*	*0.00037*	*0.0455*
	10				*0.013023*	*0.00034*	*0.0268*
	15				*0.015864*	*0.00032*	*0.0204*
	20				*0.01758*	*0.0003*	*0.0173*
0.95	5	1	500	500	*0.009058*	*0.00041*	*0.0454*
	10				*0.016047*	*0.00036*	*0.0225*
	15				*0.021479*	*0.00033*	*0.0153*
	20				*0.025669*	*0.00031*	*0.0122*
0.99	5	1	500	500	*0.009814*	*0.00044*	*0.045*
	10				*0.019064*	*0.00041*	*0.0214*
	15				*0.028016*	*0.0004*	*0.0143*
	20				*0.036455*	*0.00038*	*0.0103*

Table 1: E[RBO] for varying p.

For instance, if a researcher wants to evaluate some data with $p = 0.99$ and decided to truncate it to 10 elements for each list, the weight of those 10 elements is only 17%. This is substantially small for a prefix in order to use it for calculating RBO_{EXT}, and thus lower values for p should be used. Specifically for $p = 0.99$, a prefix should be at least 150 in depth to have 90% weight. When $p = 0.95$ or $p = 0.9$, a good enough size is 30 and 15 respectively.

Now the trend of $E[RBO]$ can be examined. Figure 3 displays the growth of $E[RBO]$ for $p = [0.8, 0.9, 0.95, 0.99]$. The size of the prefixes is again 10, and the domain contains 500 elements.

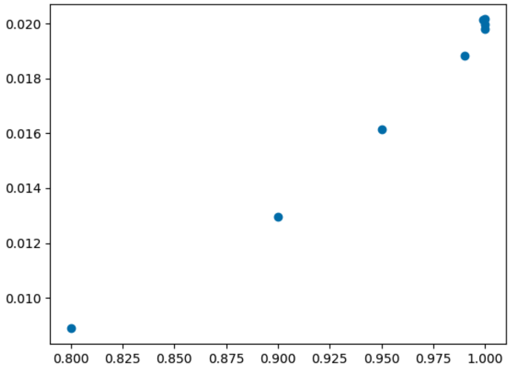


Figure 3: Growth of E[RBO] when p is increasing. $N = 10, D1 = D2 = 500$.

Typically, as p approaches 1, the weights become arbitrarily flat, and the evaluation extends deeper into the rankings. An increase in p leads to higher RBO scores due to the increased emphasis on overlap at lower ranks.

This can be explained by the following. RBO_{EXT} is composed of two parts - the base score at depth k *plus* weighted agreement of the tail. The former is shown to decrease rapidly with increasing p in Figure 4. However, the latter increases with increasing p, shown in Figure 5.

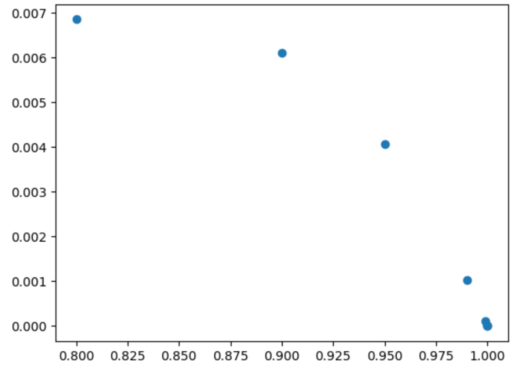


Figure 4: $RBO@k$ when p increases. $N = 10, D1 = D2 = 500$.

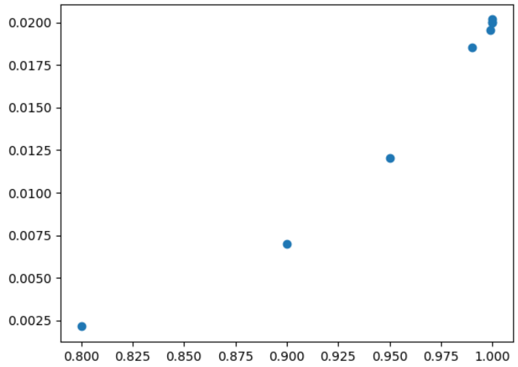


Figure 5: Weighted agreement of tail when p increases. $N = 10, D1 = D2 = 500$.

As p increases, the assumption, that the agreement seen up to a certain rank continues indefinitely, has a more pronounced effect. This is because the contribution of the top-ranked elements, which might differ significantly due to randomness, is diminished, thereby decreasing the influence of differences in the top ranks. In other words, high p values reduce the influence of discrepancies in the top ranks; whatever the differences between the top ranks of the lists are, these differences are down-weighted with higher p, leading to higher overall RBO scores.

The behavior of $E[RBO]$ can be examined when domain sizes *tend to infinity*. For the following simulations, two rankings of size 15 are sampled. Figure 6 provides an overview of the average RBO for four different domains. The blue line is bound the a configuration, where the domains are equal to 500. The orange one shows that when the domains are equal to 1000, the growth is similar, but more steady, and the range of values now is decrease by some factor. The green and red lines are based on a domain sizes of 1500 and 2000 respectively, and again provide the insight that as the domain gets even larger, the growth becomes slower, and also for some fixed p, $E[RBO]$ is lower.

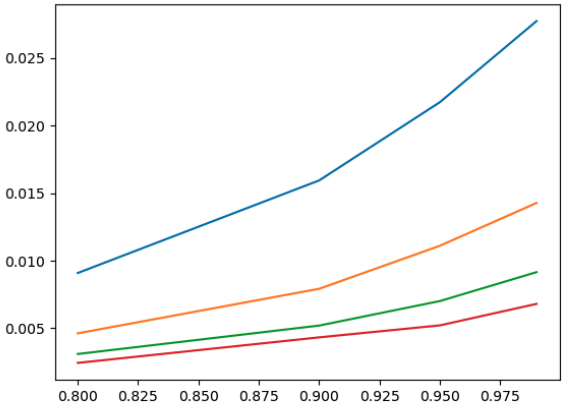


Figure 6: Growth of E[RBO] when p is increasing. $N = 15, D1 = D2 = [500, 1000, 1500, 2000]$.

Figure 7 provides the same insight, but now the domain sizes are $D1 = D2 = [5000, 10000, 50000, 100000]$, for $blue$, $orange$, $green$ and red respectively.

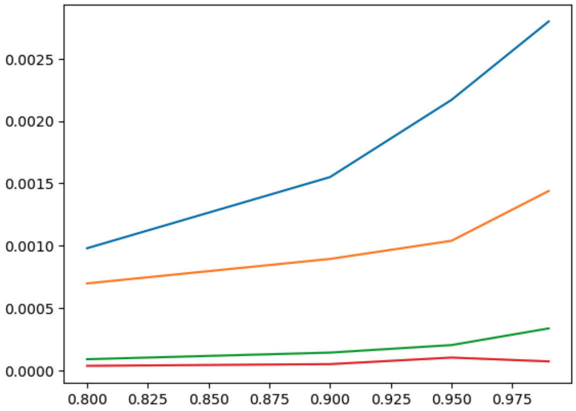


Figure 7: Growth of E[RBO] when p is increasing. $N = 15, D1 = D2 = [500, 1000, 1500, 2000]$.

This is expected, since when a fixed amount of elements are sampled from larger domains, less elements will overlap on average. Thus, it can be confidently said that as $D \to \infty$, $E[RBO] \to 0$.

4.2 Expected RBO when prefix size changes

This section analyzes the trend of $E[RBO]$ when the size of the prefixes change. Figure 8 provides results about four simulations for different prefix sizes, sampled from a domain of 1000. Every simulation is bound to a different p value.

When $p = 0.8$, $E[RBO]$ starts growing rapidly when $N = [1, 2, ..., 15]$. Around $N = 20$, the average value starts converging to some number, with some variances present.

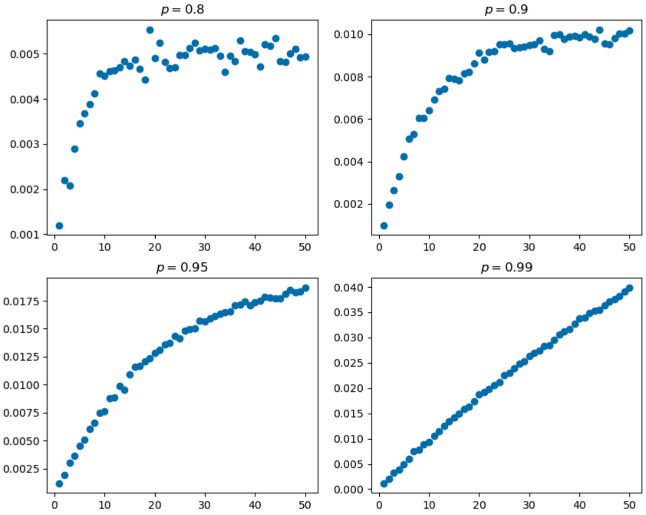


Figure 8: Four different plots for $p = [0.8, 0.9, 0.95, 0.99]$, when prefix size changes. $D1 = D2 = 1000$.

This is due to the fact that a prefix of size has 0.998 weight, and this weight will only increase when N increases.

When $p = 0.9$, the trend becomes a bit more stable. Around $N = 25$, the weight of the prefix is already at 99%, which again will only increase with positive changes in p. Convergence is observed when $N >= 30$.

When $p = 0.95$, the average value is seen to slow down, when N is around 50. This size has a weight of 98%. Only when $N >= 60$, the weight is at least 99%, so $E[RBO]$ will start converging around that size.

And finally, when $p = 0.99$, the trend looks *almost* linear. Now it can be said that it grows more steadily. No convergence is observed, because the weight of the prefix of size $N = 50$ is only 67%. This weight will climb to 99% when $N = 304$. Thus, if $p = 0.99$, for a practical evaluation, where it has been decided to cut the ranking at depth of less than 50, further deeper evaluation will only increase the value of RBO_{EXT}, and no convergence will be observed at all until depth of somewhere around 304.

It is interesting to observe that in all cases, $E[RBO]$ grows up. When more elements are sampled randomly from the same domain, the overlap and thus RBO will increase on average. The level of p determines the slope of that growth. The lower the p, the faster a convergence is observed, if any. This is relative to the maximum cut-off depth chosen for an evaluation. Convergence is always observed though, but this happens for larger sizes, when p is large. More concretely, the magnitude of a prefix size that first reaches some threshold value of weight, for example, 99%, represents the boundary of when RBO will start converging. This comes natural from the properties of RBO, since when p tends to 1, the weights become arbitrarily flat, and the weight overlap seen in the top ranks decreases.

It is important to note that for a fixed p and increasing n, at some point the weight of the prefix reaches 100%, which renders testing for larger n unnecessary. One should have an informed decision on what weight their lists would have, based on the value of p that is chosen. This can be calculated by W_{RBO}, defined in Section 2.3.

P	N	Conj	D1	D2	E[RBO]		
					mean	*sd*	*cv*
0.8	5	1	1000		*0.003364*	*7.968e-05*	*0.0237*
	10				*0.004455*	*7.375e-05*	*0.0166*
	15				*0.00482*	*7.013e-05*	*0.0145*
	20				*0.00495*	*6.506e-05*	*0.0131*
	30				*0.005030*	*0.00022*	*0.044*
0.9	5	1	1000		*0.004155*	*0.00026*	*0.064*
	15				*0.007969*	*0.00025*	*0.0314*
	20				*0.008782*	*0.00019*	*0.0226*
	40				*0.009839*	*0.00019*	*0.0198*
	100				*0.010006*	*0.00019*	*0.0198*
0.95	10	1	1000		*0.008025*	*0.00026*	*0.0327*
	20				*0.012823*	*0.00020*	*0.0162*
	40				*0.017427*	*0.00017*	*0.0102*
	100				*0.019876*	*0.00021*	*0.0106*
0.99	10	1	1000		*0.009533*	*0.00029*	*0.0306*
	20				*0.018280*	*0.00027*	*0.0151*
	40				*0.033102*	*0.00023*	*0.0072*
	50				*0.039467*	*0.00023*	*0.0058*
	100				*0.063377*	*0.0002*	*0.0031*
	200				*0.086579*	*0.0002*	*0.0023*
	350				*0.097012*	*0.00017*	*0.0018*

Table 2: Different E[RBO] when p changes. At some point, it converges around a value, when N is large enough.

Table 2 gives some example reference values for lists that are sampled from a domain of 1000, and different values for p are used for the evaluation. It can be observed that, with $p = 0.8$, large prefix sizes converge to a value of 0.005, around when $N = 20$. For $p = [0.9, 0.95, 0.99]$, $E[RBO]$ converges at 0.01, 0.02, 0.1, at around $N = [50, 60, 500]$ respectively.

4.3 Expected RBO when degree of conjointness changes

This section analyzes the trend of the RBO when the degree of conjointness between the domains, where elements are sampled from, is varied. It also examines configurations for simulations, where the domains' sizes differ.

Figure 9 presents different values for $E[RBO]$ when the degree of conjointness varies. It this example, $p = 0.8$, the rankings are each of length 15, and the domains are equal to 500. It is noticeable that as the conjointness grows, $E[RBO]$ grows slower. An observation was made in Section 4.2 that when p tends to 1, $E[RBO]$ converges for larger prefixes.

The same growth is observed when p is increased to 0.95, and when the prefixes get to larger sizes, in Figure 10. From table [insert table in appendix of different weights of prefixes for n and p], one can see that the weights of the four prefixes are 0.85, 0.93, 0.96 and 0.98.

The analysis of the trend of $E[RBO]$ of rankings, sampled from larger domains, is illustrated in Figure 11. As the domain sizes increases, the magnitude of *all* point estimates decreases and tends to *zero*. The overall rate of growth remains consistent, and decelerates as the degree of conjointness gets

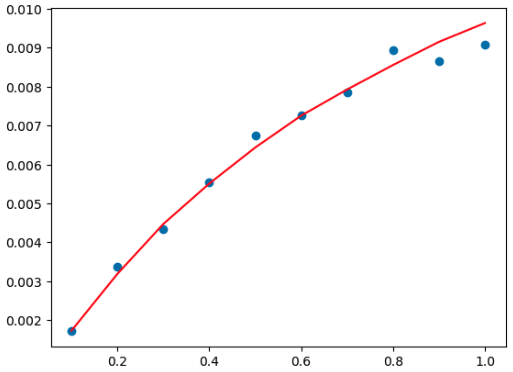


Figure 9: Growth of E[RBO] for different values of degree of conjointness. $p = 0.8, N = 15, D1 = D2 = 500$.

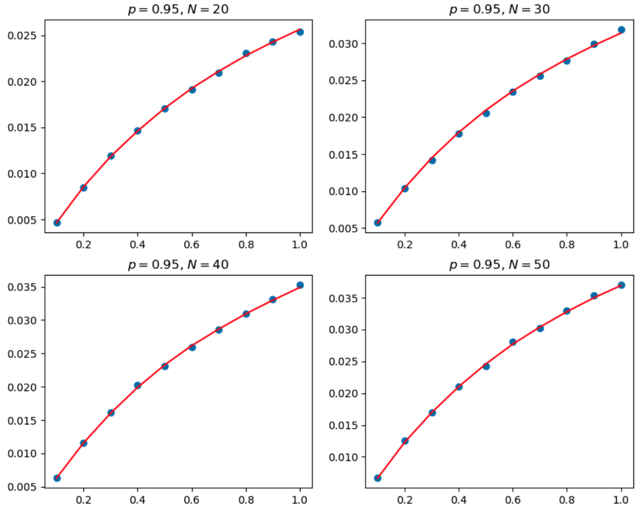


Figure 10: Trend of E[RBO] for different degrees of conjointness. $p = 0.95, N = [20, 30, 40, 50], D1 = D2 = 500$.

closer to 1. However, Figure 11 does not clearly depict the precise approximation of growth, when $D >= 2,000,000$. This ambiguity is expected due to the margin of error inherent in the simulations, which utilize only $10,000$ iterations. The substantial variance in individual results is significant compared to the miniscule magnitude of the actual results. A simulation employing $1,000,000$ iterations, for the same configuraion, is provided in Appendix B for further evidence of a more well-approximated growth.

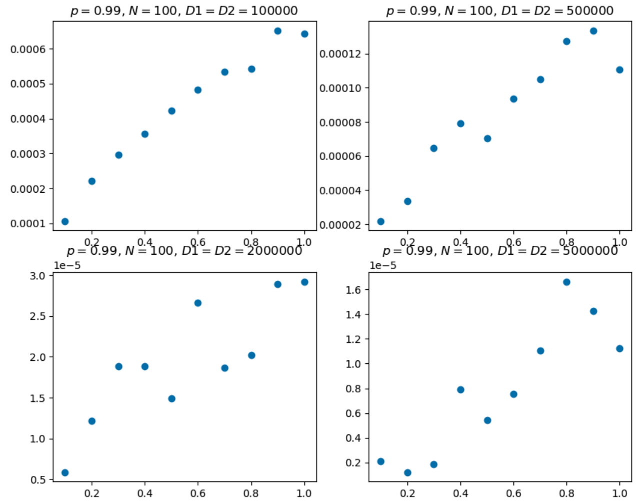


Figure 11: Trend of E[RBO] for different degrees of conjointness. $p = 0.99, N = 100$.

The impact of differing domain sizes should be closely analyzed. Figure 12 illustrates the growth of $E[RBO]$, when the larger domain size is fixed at $10,000$, and the other one approaches it. When $D1 = 1000$, $E[RBO]$ initially starts low, when conjointness is 0.1. As conjointness increases beyond 0.1, $E[RBO]$ jumps and then fluctuates until the domains are fully conjoint. This variance is attributed to the margin error of the simulations, however this is considered as convergence. It is important to note that with these domain sizes, the maximum degree of conjointness cannot exceed 0.1, since only 1000 elements will overlap in the worst case.

In the upper-right plot, $E[RBO]$ increases until the degree of conjointness reaches 0.2. In the lower-left graph, with $D1 = 5000$, the maximum actual conjointness is 0.5, after which convergence is observed. Finally, the lower-right plot, where both domain sizes are $10,000$, visualizes the expected growth previously discussed.

This indicates that the average RBO increases only when the conjointness grows to $\frac{min(D1,D2)}{max(D1,D2)}$. Therefore, it is not worthwhile to consider simulations where the conjointness is set higher than the ratio of the two domains.

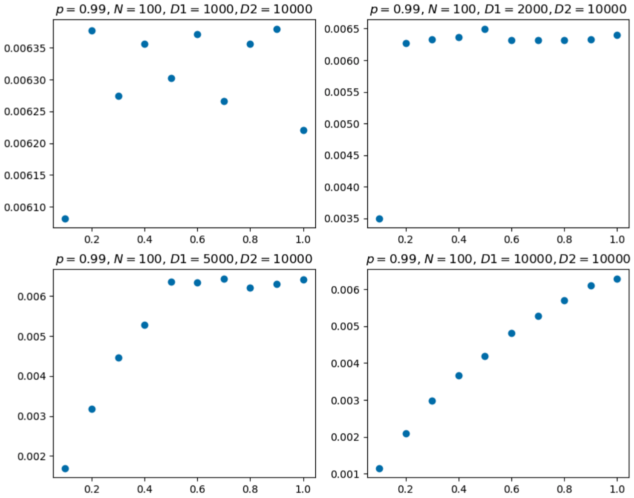


Figure 12: Trend of E[RBO] for different degrees of conjointness. $p = 0.99, N = 100$.

A further simulation is executed for $1,000,000$ iterations. There, $p = 0.9$, conjointness is fixed at 1, $D2 = 10,000$ and $D1$ ranges over $[100, 200, 500, 1,000, 2,000, 5,000, 7,500, 10,000]$. All the values calculated sit around 0.000985, which shows that for a fixed conjointness, the larger domain matters more.

5 Responsible Research

It is essential to explicitly state that all findings and conclusions drawn in this work are based on strict assumptions. The reference values derived and discussed in this paper might not be applicable or useful for some real-world data, where *Rank-Biased Overlap* is the desired measure to use.

For example, a lot of rankings might not exhibit the same properties as those used for conducting this study. Specifically, our simulations and dataset constructions are predicated on certain characteristics such as evenness (i.e., rankings having exactly the same size), uniform list and domain sizes, and the absence of ties. These controlled conditions are critical for the theoretical framework and experiments presented herein, but they might not reflect the variability and complexity found in real-world data.

For instance, the rankings can differ in size, contain ties, and lack clearly defined domains and therefore conjointness. Consequently, the applicability of our reference values and assumptions to such data is limited. It is crucial to acknowledge that the newly-proposed assumptions, while useful for theoretical exploration and controlled experiments, might not hold in diverse and unpredictable real-world contexts. Therefore, these assumptions should not be used indiscriminately without careful consideration of their relevance to the specific data at hand.

Before applying the new RBO formulations and insights derived from this research, readers must critically evaluate the assumptions underlying our approach. It is important to assess whether these assumptions are applicable to their specific use-case and data characteristics. The approach taken in this paper for understanding and drawing insights about RBO is contingent upon the controlled conditions and properties defined in our simulations. Therefore, researchers and practitioners need to ensure that their data and research questions align with these conditions before adopting the new RBO formulations. This careful consideration will help maintain the integrity and relevance of the research, ensuring that the application of these methods is both responsible and appropriate for the intended context.

The dataset and the algorithms utilized for its generation are available in a publicly accessible Github [1] repository. The dataset is stored in a *CSV* format, and all code is written in *Python*. This initiative ensures that researchers have full access to the data, enabling them to replicate and extend the simulations with inputs that closely match their own data. Moreover, the algorithms can be modified and employed in subsequent studies, facilitating the evaluation of RBO under varying assumptions and expanding the scope of the research.

6 Discussion

This paper empirically examines the trend of $E[RBO]$ and establishes reference values as initial benchmarks for expedient orientation. It is shown that it depends on all factors, used for one evaluation, such as prefix size, chosen p, degree of conjointness between domains, and their respective sizes. A simulation of $10,000$ iterations produces a value with small enough margin of error of 5% or less.

When the size of the prefix increases, the average value also increases. However, for big enough N, the value converges, due to the relationship between N and p, the latter of which determines the weight of the first N elements. With an increase of p, $RBO@k$, or base RBO, defined in Section 2.2, decreases. The differences of the top ranks are more downweighted. When p is low, $E[RBO]$ converges faster, relative to a same-sized prefix. Furthermore, when N increases, and prefixes of size N are sampled from larger sizes, the growth remains similar, but all of the values get closer to zero.

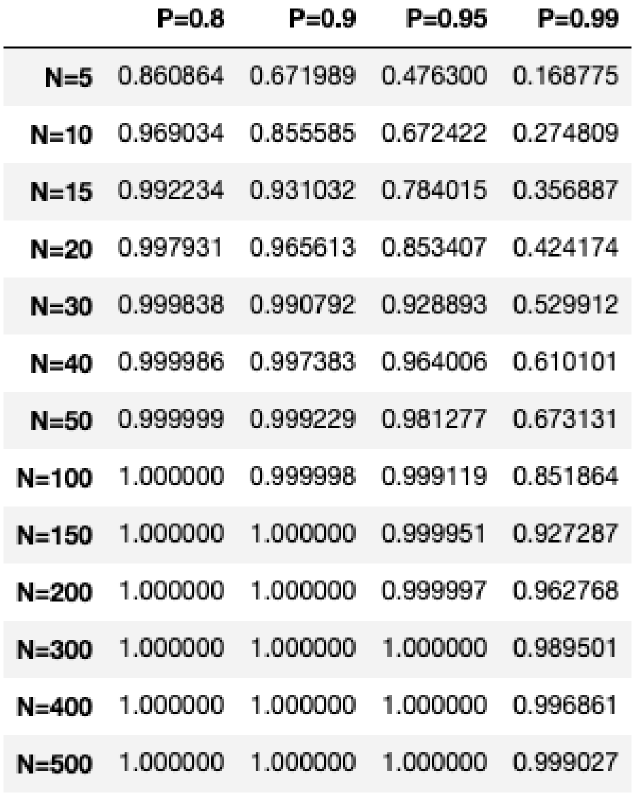
When the degree of conjointness varies, a similar growth is observed for different combinations of N and P. It decelerates when conjointness tends to 1. Even more, it has been shown that the difference between domain sizes is important only when conjointness changes, but when the latter is fixed, only the bigger domain determines the outcome.

7 Future Work

This study restricts its analysis to equal-sized prefixes, a scenario uncommon in practical applications. Additionally, it assumes the absence of ties in rankings, a condition contingent on the source generating the data. Future research endeavors may address these limitations by empirically examining diverse scenarios or developing more comprehensive mathematical frameworks to accommodate varying prefix sizes and the presence of ties in rankings.

[1] https://github.com/mark200/expected-rbo

A Weights of prefixes



	P=0.8	P=0.9	P=0.95	P=0.99
N=5	0.860864	0.671989	0.476300	0.168775
N=10	0.969034	0.855585	0.672422	0.274809
N=15	0.992234	0.931032	0.784015	0.356887
N=20	0.997931	0.965613	0.853407	0.424174
N=30	0.999838	0.990792	0.928893	0.529912
N=40	0.999986	0.997383	0.964006	0.610101
N=50	0.999999	0.999229	0.981277	0.673131
N=100	1.000000	0.999998	0.999119	0.851864
N=150	1.000000	1.000000	0.999951	0.927287
N=200	1.000000	1.000000	0.999997	0.962768
N=300	1.000000	1.000000	1.000000	0.989501
N=400	1.000000	1.000000	1.000000	0.996861
N=500	1.000000	1.000000	1.000000	0.999027

Figure 13: Weights of prefixes for different sizes and different p, chosen for evaluation.

B Plot of average RBO when domain is large

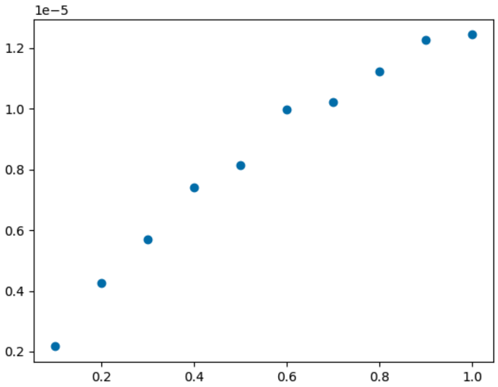


Figure 14: Trend of E[RBO] for different degrees of conjointness. $p = 0.99, N = 100, D1 = D2 = 5,000,000$.

References

[1] Judit Bar-Ilan. Comparing rankings of search results on the web. *Information Processing Management*, 41(6):1511–1519, 2005. Special Issue on Infometrics.

[2] Chris Buckley. Topic prediction based on comparative retrieval rankings. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 506–507, New York, NY, USA, 2004. Association for Computing Machinery.

[3] Maurice George Kendall. Rank correlation methods. 1948.

[4] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), nov 2010.

[5] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 587–594, New York, NY, USA, 2008. Association for Computing Machinery.