# Comparing top $k$ lists

Ronald Fagin Ravi Kumar D. Sivakumar

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

{fagin, ravi, siva}@almaden.ibm.com

**Abstract**

Motivated by several applications, we introduce various distance measures between "top $k$ lists." Some of these distance measures are metrics, while others are not. For each of these latter distance measures, we show that they are "almost" a metric in the following two seemingly unrelated aspects:

(i) they satisfy a relaxed version of the polygonal (hence, triangle) inequality, and

(ii) there is a metric with positive constant multiples that bound our measure above and below.

This is not a coincidence—we show that these two notions of almost being a metric are formally identical. Based on the second notion, we define two distance measures to be *equivalent* if they are bounded above and below by constant multiples of each other. We thereby identify a large and robust equivalence class of distance measures.

Besides the applications to the task of identifying good notions of (dis-)similarity between two top $k$ lists, our results imply polynomial-time constant-factor approximation algorithms for the *rank aggregation problem* [DKNS01] with respect to a large class of distance measures.

**To appear in SIAM J. on Discrete Mathematics.**
**Extended abstract to appear in 2003 ACM-SIAM Symposium on Discrete Algorithms (SODA '03).**

# 1 Introduction

The notion of a "top $k$ list" is ubiquitous in the field of information retrieval (IR). A top 10 list, for example, is typically associated with the "first page" of results from a search engine. While there are several standard ways for *measuring* the "top $k$ quality" of an information-retrieval system (e.g., precision and recall at various values of $k$), it appears that there is no well-studied and well-understood method for *comparing* two top $k$ lists for similarity/dissimilarity. Precision and recall based methods yield a way to compare two top $k$ lists by comparing them both to "ground truth." However, there are two limitations of such an approach: First, these methods typically give absolute (unary) ratings of top $k$ lists, rather than give a relative, binary measure of distance. Second, for information retrieval in the context of the world-wide web, there is often no clear notion of what ground truth is, so precision and recall are harder to use.

These observations lead to the following question in discrete mathematics: *how do we define reasonable and meaningful distance measures between top $k$ lists*? We motivate the study of this problem by sketching some applications.

**Applications.** The first group of applications we describe is in the comparison of various search engines, or of different variations of the same search engine. What could be a more natural way to compare two search engines than by comparing their visible outputs (namely, their top $k$ lists)? It is also important to compare variations (using slightly different ranking functions) of the same search engine, as an aid in the design of ranking functions. In particular, we can use our methodology to test the effect on the top $k$ lists of adding/deleting ranking heuristics to/from the search engine. Similar issues include understanding the effect of augmenting the "crawl" data to add more documents, of indexing more data types (e.g., PDF documents), etc. For a more complex application in this group, consider a large-scale search engine. Typically, its ranking function is a composite algorithm that builds on several simpler ranking functions, and the following questions are of interest. What is the "contribution" of each component to the final ranking algorithm, or how similar is the top $k$ composite output to the top $k$ of each of its constituents, and how similar is each component to the others? A good quantitative way to measure these (which our methodology supplies) could be a valuable tool in deciding which components to retain, enhance, or delete so as to design a better ranking algorithm. Similarly, our methodology can be used to compare a "meta-search" engine with each of its component search engines, in order to understand the degree to which the metasearch engine aligns itself with each of its components. In Section 9, we report our results on the comparisons of seven popular Web search engines and on comparing a metasearch engine with its components.

The second group of the applications can be classified as "engineering optimizations." A fairly simple example is a system that draws its search results from several servers; for the sake of speed, a popular heuristic is to send the query to the servers, and return the responses as soon as, say, 75% of the servers have responded. Naturally, it is important to ensure that the quality of the results are not adversely affected by this approximation. What one needs here are meaningful and quantitative measures with which to estimate the difference in the top $k$ lists caused by the approximation. A more subtle example in the same category is the following (where, in fact, our methodology has already been successfully utilized). Carmel et al. [CCF+01] explored the effect of pruning the index information of a search engine. Their experimental hypothesis, which they verified using one of our distance measures, was that their pruning technique would have only small effects on the top $k$ list, for moderate values of $k$. Since what a user sees is essentially a top $k$ list, they concluded that they could prune the index greatly, which resulted in better space and time performance, without much effect on the search results. Another scenario in a similar vein is in the area of approximate near-neighbor searching, a very common technique for categorization problems. Here an important goal is to understand the difference between approximate and exact near-neighbor search; once again, since what matters the most are the top few results, our problem arises naturally.

Another application of comparing top $k$ lists arises from the processing of data logs to discover emerging

trends (see [CCFC02] for an example). For example, a search engine could compute the top 100 queries each day and see how they differ from day to day, from month to month, etc. Other examples include processing inventory logs and sales logs in retail stores, logs of stocks traded each day, etc. In these cases, a spike in the difference between day-to-day or hour-to-hour top $k$ lists could trigger a closer analysis and action (e.g., buy/sell shares, add inventory, etc.). For these settings, one needs good notions of difference between two given top $k$ lists.

Finally, we consider the context of synthesizing a good composite ranking function from several simpler ones. In the *rank aggregation problem* [DKNS01], given several top $k$ lists, the goal is to find a top $k$ list that is a "good" consolidation of the given lists. In [DKNS01] this problem is formulated by asking for an aggregation that has the minimum total distance with respect to the given lists, where the distance is computed according to some distance measure of interest. The choice of distance measure turns out to have a direct bearing on the complexity of computing the best solution: some distance measures lead to NP-hard optimization problems, while others admit polynomial-time solutions. A main algorithmic consequence of our work is in enabling the design of efficient constant-factor approximation algorithms for the aggregation problem with respect to a large class of distance measures. This is achieved by identifying a class of distance measures that are within constant factors of each other.

**Results.** We approach the problem of defining distance measures between top $k$ lists from many angles. We make several proposals for distance measures, based on various motivating criteria—ranging from naive, intuitive ones to ones based on rigorous mathematics. While the plethora of measures is good news (since it gives a wide choice), it also poses the challenging question of how to understand their relative merits, or how to make a sound choice among the many competing proposals.

One of our main contributions is a unified framework in which to catalog and organize various distance measures. Concretely, we propose the notion of an *equivalence class* of distance measures and, in particular, place many of the proposed distance measures into one large equivalence class (which we dub the "big equivalence class"). Our big equivalence class encompasses many measures that are intuitively appealing (but whose mathematical properties are nebulous), as well as ones that were derived via rigorous mathematics (but lacking in any natural, intuitive justification that a user can appreciate). The main message of the equivalence class concept is that up to constant factors (that do not depend on $k$), all distance measures in an equivalence class are essentially the same.

Our equivalence classes have the property that if even one distance measure in a class is a *metric* (in the usual mathematical sense), then each of the others in that class is a "near metric." To make the foregoing idea precise, we present two distinct but seemingly unrelated definitions of a near metric—satisfying a relaxed version of the "polygonal inequality" (the natural extension of the standard triangle inequality), and there existing a metric with positive constant multiples that bound our measure above and below. We prove the surprising result that these two notions of near metric are, in fact, equivalent.

Our results have the following two consequences:

(1) The task of choosing a distance measure for IR applications is now considerably simplified. The only conscious choice a user needs to make is about which equivalence class to use, rather than which distance measure to use. Our personal favorite is the big equivalence class that we have identified, mainly because of the rich variety of underlying intuition and the mathematically clean and algorithmically simple methods that it includes.

(2) We obtain constant-factor approximation algorithms for the rank aggregation problem with respect to every distance measure in our big equivalence class. This is achieved using the fact that the rank aggregation problem can be optimally solved in polynomial time (via minimum cost perfect matching) for one of the distance measures in this equivalence class.

As we noted, in Section 9 we present an illustration of the applicability of our methods in the context of search and metasearch. Based on the results for 750 user queries, we study the similarities between the top 50

lists of seven popular Web search engines, and also their similarity to the top 50 list of a metasearch engine built using the seven search engines. The quantitative comparison of the search engines' top 50 results brings some surprising qualitative facts to light. For example, our experiments reveal that AOL Search and MSN Search yield very similar results, despite the fact that these are competitors. Further analysis reveals that the crawl data for these search engines (and also for the search engine HotBot) comes in part from Inktomi. The fact that the top 50 results from HotBot are only moderately similar to that of AOL and MSN suggests that while they all use crawl data from Inktomi, HotBot probably uses a ranking function quite different from those of AOL and MSN. We believe these studies make an excellent case for the applicability of quantitative methods in comparing top $k$ lists.

**Methodology.** A special case of a top $k$ list is a "full list," that is, a permutation of all of the objects in a fixed universe. There are several standard methods for comparing two permutations, such as Kendall's tau and Spearman's footrule (see the textbooks [KG90, Dia88]). We cannot simply apply these known methods, since they deal only with comparing one permutation against another over the same elements. Our first (and most important) class of distance measures between top $k$ lists is obtained by various natural modifications of these standard notions of distances between permutations.

A less sophisticated attempt at defining a metric is to compute the intersection of the two top $k$ lists (viewing them as sets). This approach has in fact been used in several papers in information retrieval [Lee95, Lee97, CCF+01]. In order to obtain a metric, we consider the notion of the symmetric difference (union minus the intersection), appropriately scaled. This, unfortunately, is not adequate for the top $k$ distance problem, since two top 10 lists that are reverses of each other would be declared to be "very close." We propose natural extensions of this idea that leads to more robust metrics that are meaningful for top $k$ lists. Briefly, the idea is to truncate the top $k$ lists at various points $i \leq k$, compute the symmetric difference metric between the resulting top $i$ lists, and take a suitable combination of them. This gives a second type of notion of the distance between top $k$ lists.

As we noted, our distance measure based on the intersection gives a metric. What about our distance measures that are generalizations of metrics on permutations? Some of these turn out to be metrics, but others do not. For each of these distance measures $d$ that is not a metric, we show that $d$ is a "near metric" in two seemingly different senses. Namely, $d$ satisfies each of the following two properties.

*Metric boundedness property:* There is a metric $d'$ and positive constants $c_1$ and $c_2$ such that for all $x, y$ in the domain, $c_1 d'(x, y) \leq d(x, y) \leq c_2 d'(x, y)$ for all $x, y$ in the domain.

Thus, metric boundedness says that $d$ and some metric $d'$ are within constant multiples of each other.

*Relaxed polygonal inequality:* There is a constant $c$ such that for all $n > 0$ and $x, z, x_1, \ldots, x_{n-1}$ in the domain, $d(x, z) \leq c(d(x, x_1) + d(x_1, x_2) + \cdots + d(x_{n-1}, z))$

As remarked earlier, we show the surprising fact that these two seemingly unrelated notions of being a "near metric" are the same. Note that the relaxed polygonal inequality immediately implies the relaxed triangle inequality [FS98], which says that there is a constant $c$ such that $d(x, z) \leq c(d(x, y) + d(y, z))$ for all $x, y, z$ in the domain. Relaxed triangle and polygonal inequalities suggest that the notion of "closeness" under these measures are "reasonably transitive." Interestingly enough, the equivalence of our two notions of "near metric" requires that we consider the relaxed polygonal inequality, rather than simply the relaxed triangle inequality; the relaxed triangle inequality is not sufficient to imply the metric boundedness property.

**Organization.** In Section 2, we review two metrics on permutations, which form the basis for various distance measures that we define and study. In Section 3, we develop our new distance measures between top $k$ lists. In Section 4, we present various notions of near metric, and show the equivalence between metric boundedness and the relaxed polygonal inequality. In Section 5 we define the notion of equivalence of distance measures, and show that all of our distance measures are in one large and robust equivalence

class, called the "big equivalence class." Thus each of the distance measures between top $k$ lists introduced in Section 3 is a metric or a near metric. In Section 6, we give an algorithmic application that exploits distance measures being in the same equivalence class.

## 2 Metrics on permutations

The study of metrics on permutations is classical. The book by Kendall and Gibbons [KG90] provides a detailed account of various methods. Diaconis [Dia88] gives a formal treatment of metrics on permutations. We now review two well-known notions of metrics on permutations.

A *permutation* $\sigma$ is a bijection from a set $D = D_\sigma$ (which we call the *domain*, or *universe*), onto the set $[n] = \{1, \ldots, n\}$, where $n$ is the size $|D|$ of $D$. Let $S_D$ denote the set of all permutations of $D$. For a permutation $\sigma$, we interpret $\sigma(i)$ as the position (or rank) of element $i$. We say that $i$ *is ahead of* $j$ *in* $\sigma$ if $\sigma(i) < \sigma(j)$. Let $\mathcal{P} = \mathcal{P}_D = \{\{i,j\} \mid i \neq j \text{ and } i,j \in D\}$ be the set of unordered pairs of distinct elements. Let $\sigma_1, \sigma_2$ be two members of $S_D$.

Kendall's tau metric between permutations is defined as follows. For each pair $\{i,j\} \in \mathcal{P}$ of distinct members of $D$, if $i$ and $j$ are in the same order in $\sigma_1$ and $\sigma_2$, then let $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 0$; and if $i$ and $j$ are in the opposite order (such as $i$ being ahead of $j$ in $\sigma_1$ and $j$ being ahead of $i$ in $\sigma_2$), then let $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 1$. Kendall's tau is given by $K(\sigma_1, \sigma_2) = \sum_{\{i,j\} \in \mathcal{P}} \bar{K}_{i,j}(\sigma_1, \sigma_2)$. The maximum value of $K(\sigma_1, \sigma_2)$ is $n(n-1)/2$, which occurs when $\sigma_1$ is the reverse of $\sigma_2$ (that is, when $\sigma_1(i) + \sigma_2(i) = n+1$ for each $i$). Kendall's tau turns out to be equal to the number of exchanges needed in a bubble sort to convert one permutation to the other.

Spearman's footrule metric is the $L_1$ distance between two permutations. Formally, it is defined by $F(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|$. The maximum value of $F(\sigma_1, \sigma_2)$ is $2\lfloor (n+1)/2 \rfloor \lceil (n-1)/2 \rceil$, in the case when $\sigma_1$ is the $n^2/2$ when $n$ is even, and $(n+1)(n-1)/2$ when $n$ is odd. As with Kendall's tau, the maximum occurs when $\sigma_1$ is the reverse of $\sigma_2$. Later, we shall discuss a variation of Spearman's footrule called "Spearman's rho."

## 3 Measures for comparing top $k$ lists

We now discuss modifications of these metrics for the case when we only have the top $k$ members of the ordering. Formally, a *top* $k$ *list* $\tau$ is a mapping from a domain $D_\tau$ (intuitively, the members of the top $k$ list) to $[k]$. We say that $i$ *appears in* the top $k$ list $\tau$ if $i \in D_\tau$. Similar to our convention for permutations, we interpret $\tau(i)$ (for $i$ in $D_\tau$) as the rank of $i$ in $\tau$. If $\tau(i) < \tau(j)$, then we say that $i$ *is ahead of* $j$ or $i$ *precedes* $j$ in $\tau$. If $\tau$ is a top $k$ list and $\sigma$ is a permutation, then we say that $\sigma$ is an *extension* of $\tau$, which we denote $\sigma \succeq \tau$, if $\sigma(i) = \tau(i)$ for all $i \in D_\tau$.

Assume that $\tau_1$ and $\tau_2$ are top $k$ lists. In this section, we give several measures for the distance between $\tau_1$ and $\tau_2$. We begin by recalling the definition of a metric, and formally define a distance measure. A binary function $d$ is called *symmetric* if $d(x,y) = d(y,x)$ for all $x$, $y$ in the domain and is called *regular* if $d(x,y) = 0$ if and only if $x = y$. We define a *distance measure* to be a nonnegative, symmetric, regular binary function. A *metric* is a distance measure $d$ that satisfies the *triangle inequality* $d(x,z) \leq d(x,y) + d(y,z)$ for all $x, y, z$ in the domain. All of the measures of closeness between top $k$ lists that we have considered are distance measures.

**Global notation.** Here we set up some global notation that we use throughout the paper. When two top $k$ lists $\tau_1$ and $\tau_2$ are understood, we write $D = D_{\tau_1} \cup D_{\tau_2}$; $Z = D_{\tau_1} \cap D_{\tau_2}$; $S = D_{\tau_1} \setminus D_{\tau_2}$; $T = D_{\tau_2} \setminus D_{\tau_1}$. Let $z = |Z|$. Note that $|S| = |T| = k - z$, and $|D| = 2k - z$.

### 3.1 Kendall's tau

There are various natural ways to generalize Kendall's tau to measure distances between top $k$ lists. We now consider some of them. We begin by generalizing the definition of the set $\mathcal{P}$. Given two top $k$ lists $\tau_1$ and $\tau_2$, we define $\mathcal{P}(\tau_1, \tau_2) = \mathcal{P}_{D_{\tau_1} \cup D_{\tau_2}}$ to be the set of all pairs of distinct elements in $D_{\tau_1} \cup D_{\tau_2}$.

For top $k$ lists $\tau_1$ and $\tau_2$, the *minimizing Kendall distance* $K_{\min}(\tau_1, \tau_2)$ between $\tau_1$ and $\tau_2$ is defined to be the minimum value of $K(\sigma_1, \sigma_2)$, where $\sigma_1$ and $\sigma_2$ are each permutations of $D_{\tau_1} \cup D_{\tau_2}$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$.

For top $k$ lists $\tau_1$ and $\tau_2$, the *averaging Kendall distance* $K_{\mathrm{avg}}(\tau_1, \tau_2)$ between $\tau_1$ and $\tau_2$ is defined to be the expected value $\mathrm{E}(K(\sigma_1, \sigma_2))$, where $\sigma_1$ and $\sigma_2$ are each permutations of $D_{\tau_1} \cup D_{\tau_2}$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$. Here $\mathrm{E}(\cdot)$ gives the expected value where all extensions are taken to be equally likely.

Next we consider an approach that we will show gives both the minimizing Kendall distance and the averaging Kendall distance as special cases. Let $p$ be a fixed parameter with $0 \leq p \leq 1$. Similarly to our definition of $\bar{K}_{i,j}(\sigma_1, \sigma_2)$ for permutations $\sigma_1, \sigma_2$, we define a penalty $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2)$ for top $k$ lists $\tau_1, \tau_2$ for $\{i, j\} \in \mathcal{P}(\tau_1, \tau_2)$. There are four cases.

*Case 1: $i$ and $j$ appear in both top $k$ lists.* If $i$ and $j$ are in the same order (such as $i$ being ahead of $j$ in both top $k$ lists), then let $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2) = 0$; this corresponds to "no penalty" for $\{i, j\}$. If $i$ and $j$ are in the opposite order (such as $i$ being ahead of $j$ in $\tau_1$, and $j$ being ahead of $i$ in $\tau_2$), then let the penalty $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2) = 1$.

*Case 2: $i$ and $j$ both appear in one top $k$ list (say $\tau_1$), and exactly one of $i$ or $j$, say $i$, appears in the other top $k$ list ($\tau_2$).* If $i$ is ahead of $j$ in $\tau_1$, then let the penalty $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2) = 0$, and otherwise let $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2) = 1$. Intuitively, we know that $i$ is ahead of $j$ as far as $\tau_2$ is concerned, since $i$ appears in $\tau_2$ but $j$ does not.

*Case 3: $i$, but not $j$, appears in one top $k$ list (say $\tau_1$), and $j$, but not $i$, appears in the other top $k$ list ($\tau_2$).* Then let the penalty $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2) = 1$. Intuitively, we know that $i$ is ahead of $j$ as far as $\tau_1$ is concerned, and $j$ is ahead of $i$ as far as $\tau_2$ is concerned.

*Case 4: $i$ and $j$ both appear in one top $k$ list (say $\tau_1$), but neither $i$ nor $j$ appears in the other top $k$ list ($\tau_2$).* This is the interesting case (the only case where there is really an option as to what the penalty should be). We call such pairs $\{i, j\}$ *special pairs*. In this case, we let the penalty $\bar{K}_{i,j}^{(p)}(\tau_1, \tau_2) = p$.

Based on these cases, we now define $K^{(p)}$, the *Kendall distance with penalty parameter $p$*, as follows:

$$K^{(p)}(\tau_1, \tau_2) = \sum_{\{i,j\} \in \mathcal{P}(\tau_1, \tau_2)} \bar{K}_{i,j}^{(p)}(\tau_1, \tau_2).$$

When $p = 0$, this gives an "optimistic approach." It corresponds to the intuition that we assign a nonzero penalty score to the pair $\{i, j\}$ only if we have enough information to know that $i$ and $j$ are in the opposite order according to the two top $k$ lists. When $p = 1/2$, this gives a "neutral approach." It corresponds to the intuition that we do not have enough information to know whether the penalty score should be 0 or 1, so we assign a neutral nonzero penalty score of $1/2$. Later, we shall show that the optimistic approach gives precisely $K_{\min}$, and the neutral approach gives precisely $K_{\mathrm{avg}}$.

The next lemma gives a formula, which we shall find useful later, for $K^{(p)}$.

**Lemma 3.1.** $K^{(p)}(\tau_1, \tau_2) = (k - z)((2 + p)k - pz + 1 - p) + \sum_{i,j \in Z} \bar{K}_{i,j}^{(0)}(\tau_1, \tau_2) - \sum_{j \in S} \tau_1(j) - \sum_{j \in T} \tau_2(j)$.

*Proof.* We analyze the four cases in the definition of $K^{(p)}(\tau_1, \tau_2)$ and obtain formulas for each of them in terms of our global notation. Case 1 is the situation when for a pair $\{i, j\}$, we have $i, j \in Z$. In this case,

the contribution of this pair to $K^{(p)}(\tau_1, \tau_2)$ is

$$\sum_{i,j \in Z} \bar{K}^{(0)}_{i,j}(\tau_1, \tau_2). \tag{1}$$

Case 2 is the situation when for a pair $\{i, j\}$, one of $i$ or $j$ is in $Z$ and the other is in either $S$ or $T$. Let us denote by $i$ the element in $Z$, and by $j$ the element in $S$ or $T$. Let us now consider the case when $i \in Z, j \in S$. Let $j_1 < \cdots < j_{k-z}$ be the elements in $S$. Fix an $\ell \in \{1, \ldots, k-z\}$ and consider the element $j_\ell$ and its rank $\tau_1(j_\ell)$ in the first top $k$ list $\tau_1$. There will be a contribution of 1 to $K^{(p)}(\tau_1, \tau_2)$ for all $i \in Z$ such that $\tau_1(i) > \tau_1(j_\ell)$, that is, all the elements $i \in Z$ such that $j_\ell$ is ahead of $i$ in $\tau_1$; denote this net contribution of $\ell$ to $K^{(p)}(\tau_1, \tau_2)$ by $\gamma(\ell)$. We now obtain an expression for $\gamma(\ell)$. The total number of elements that $j_\ell$ is ahead of in $\tau_1$ is $k - \tau_1(j_\ell)$ and of these elements, $\ell - 1$ of them belong to $S$ and the rest belong to $Z$. This gives $\gamma(\ell) = k - \tau_1(j_\ell) - (\ell - 1)$. Now, summing over all $\ell$, the contribution to $K^{(p)}(\tau_1, \tau_2)$ is $\sum_{\ell=1}^{k-z} \gamma(\ell) = (k-z)(k+z+1)/2 - \sum_{j \in S} \tau_1(j)$. Similarly, for the case when $i \in Z, j \in T$, the contribution to $K^{(p)}(\tau_1, \tau_2)$ is $(k-z)(k+z+1)/2 - \sum_{j \in T} \tau_2(j)$. Summing these, the term corresponding to Case 2 contributing to $K^{(p)}(\tau_1, \tau_2)$ is

$$(k-z)(k+z+1) - \sum_{j \in S} \tau_1(j) - \sum_{j \in T} \tau_2(j). \tag{2}$$

Case 3 is the situation when for a pair $\{i, j\}$, we have $i \in S$ and $j \in T$. The total contribution to $K^{(p)}(\tau_1, \tau_2)$ from this case is

$$|S| \times |T| = (k-z)^2. \tag{3}$$

Finally, Case 4 is the situation when for a pair $\{i, j\}$, we have either $i, j \in S$ or $i, j \in T$. The total contribution to $K^{(p)}(\tau_1, \tau_2)$ from this case is

$$p\binom{|S|}{2} + p\binom{|T|}{2} = 2p\binom{k-z}{2}. \tag{4}$$

Adding Equations (1)–(4), we obtain

$$K^{(p)}(\tau_1, \tau_2) = (k-z)((2+p)k - pz + 1 - p) + \sum_{i,j \in Z} \bar{K}^{(0)}_{i,j}(\tau_1, \tau_2) - \sum_{j \in S} \tau_1(j) - \sum_{j \in T} \tau_2(j).$$

□

Let $A$ and $B$ be finite sets of objects (in our case of interest, these objects are permutations). Let $d$ be a metric of distances between objects (at the moment, we are interested in the case where $d$ is the Kendall distance between permutations). The *Hausdorff distance* between $A$ and $B$ is given by

$$d_{\text{Haus}}(A, B) = \max \left\{ \max_{\sigma_1 \in A} \min_{\sigma_2 \in B} d(\sigma_1, \sigma_2), \max_{\sigma_2 \in B} \min_{\sigma_1 \in A} d(\sigma_1, \sigma_2) \right\}.$$

The Hausdorff distance is well known to be a metric. Although this looks fairly nonintuitive, it is actually quite natural, as we now explain. The quantity $\min_{\sigma_2 \in B} d(\sigma_1, \sigma_2)$ is the distance between $\sigma_1$ and the set $B$. Therefore, the quantity $\max_{\sigma_1 \in A} \min_{\sigma_2 \in B} d(\sigma_1, \sigma_2)$ is the maximal distance of a member of $A$ from the set $B$. Similarly, the quantity $\max_{\sigma_2 \in B} \min_{\sigma_1 \in A} d(\sigma_1, \sigma_2)$ is the maximal distance of a member of $B$ from the set $A$. Therefore, the Hausdorff distance between $A$ and $B$ is the maximal distance of a member of $A$ or $B$ from the other set. Thus, $A$ and $B$ are within Hausdorff distance $s$ of each other precisely if every member

of $A$ and $B$ is within distance $s$ of some member of the other set. The Hausdorff distance is well known to be a metric.

Critchlow [Cri80] used the Hausdorff distance to define a distance measure between top $k$ lists. Specifically, given a metric $d$ that gives the distance between permutations, Critchlow defined the distance between top $k$ lists $\tau_1$ and $\tau_2$ to be

$$\max\left\{\max_{\sigma_1\succeq\tau_1}\min_{\sigma_2\succeq\tau_2} d(\sigma_1,\sigma_2), \max_{\sigma_2\succeq\tau_2}\min_{\sigma_1\succeq\tau_1} d(\sigma_1,\sigma_2)\right\}. \tag{5}$$

Critchlow assumed that there is a fixed domain $D$, and so $\sigma_1$ and $\sigma_2$ range over all permutations with domain $D$. This distance measure is a metric, since it is a special case of a Hausdorff metric.

We, too, are interested in considering a version of the Hausdorff distance. However, in this paper we do not assume a fixed domain. Therefore, we define $K_{\mathrm{Haus}}$, the Hausdorff version of the Kendall distance between top $k$ lists, to be given by Equation (5) with $d(\sigma_1,\sigma_2)$ as the Kendall distance $K(\sigma_1,\sigma_2)$, but where, unlike Critchlow, we take $\sigma_1$ and $\sigma_2$ to be permutations of $D_{\tau_1}\cup D_{\tau_2}$.

Critchlow obtains a closed form for his version of Equation (5) when $d(\sigma_1,\sigma_2)$ is the Kendall distance $K(\sigma_1,\sigma_2)$. Specifically, if $n$ is the size of the underlying domain $D$, and $d(\sigma_1,\sigma_2)=K(\sigma_1,\sigma_2)$, he shows that Equation (5) is given by

$$(k-z)\left(n+k-\frac{k-z-1}{2}\right)+\sum_{i,j\in Z}\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)-\sum_{i\in S}\tau_1(i)-\sum_{i\in T}\tau_2(i). \tag{6}$$

By replacing $n$ by $2k-z$, we obtain a closed form for $K_{\mathrm{Haus}}$:

**Lemma 3.2.**

$$K_{\mathrm{Haus}}(\tau_1,\tau_2)=\frac{1}{2}(k-z)(5k-z+1)+\sum_{i,j\in Z}\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)-\sum_{i\in S}\tau_1(i)-\sum_{i\in T}\tau_2(i).$$

We show that the "optimistic approach" given by $K^{(0)}$ and the "neutral approach" given by $K^{(1/2)}$ are exactly $K_{\min}$ and $K_{\mathrm{avg}}$, respectively. Furthermore, we show the somewhat surprising result that the Hausdorff distance $K_{\mathrm{Haus}}$ also equals $K^{(1/2)}$.

**Proposition 3.3.** $K_{\min}=K^{(0)}$.

*Proof.* Let $\tau_1$ and $\tau_2$ be top $k$ lists. We must show that $K_{\min}(\tau_1,\tau_2)=K^{(0)}(\tau_1,\tau_2)$. Define $\sigma_1$ to be the extension of $\tau_1$ over $D$ where the elements are, in order, the elements of $D_{\tau_1}$ in the same order as they are in $\tau_1$, followed by the elements of $T$ in the same order as they are in $\tau_2$. For example, if $k=4$, the top 4 elements of $\tau_1$ are, in order, 1, 2, 3, 4, and the top 4 elements of $\tau_2$ are, in order, 5, 4, 2, 6, then the ordering of the elements for $\sigma_1$ is 1, 2, 3, 4, 5, 6. We similarly define the extension $\sigma_2$ of $\tau_2$ by reversing the roles of $\tau_1$ and $\tau_2$. First, we show that $K_{\min}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$ and next, show that $K(\sigma_1,\sigma_2)=K^{(0)}(\tau_1,\tau_2)$.

It is clearly sufficient to show that if $\sigma_1'$ is an arbitrary extension of $\tau_1$ (over $D$) and $\sigma_2'$ is an arbitrary extension of $\tau_2$ (over $D$), and if $\{i,j\}$ is an arbitrary member of $\mathcal{P}(\tau_1,\tau_2)$, then

$$\bar{K}_{i,j}(\sigma_1,\sigma_2)\le\bar{K}_{i,j}(\sigma_1',\sigma_2'). \tag{7}$$

When $\{i,j\}$ is not a special pair (that is, when $\{i,j\}$ falls into the first three cases of the definition of $\bar{K}^{(p)}_{i,j}(\tau_1,\tau_2)$), we have equality in (7), since the ordering of $i$ and $j$ according to $\sigma_1,\sigma_2,\sigma_1',\sigma_2'$ are forced by $\tau_1,\tau_2$. When $\{i,j\}$ is a special pair, we have $\bar{K}_{i,j}(\sigma_1,\sigma_2)=0$, and so again (7) holds.

We have shown that $K_{\min}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$. Hence, we need only show that $K^{(0)}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$. To show this, we need only show that $\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)=\bar{K}_{i,j}(\sigma_1,\sigma_2)$ for every pair $\{i,j\}$. As before, this is automatic when $\{i,j\}$ is not a special pair. When $\{i,j\}$ is a special pair, we have $\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)=0=\bar{K}_{i,j}(\sigma_1,\sigma_2)$. This concludes the proof. □

**Proposition 3.4.** $K_{\mathrm{avg}} = K^{(1/2)} = K_{\mathrm{Haus}}$.

*Proof.* Let $\tau_1, \tau_2$ be top $k$ lists. Then

$$\begin{aligned} K_{\mathrm{avg}}(\tau_1, \tau_2) &= \mathrm{E}(K(\sigma_1, \sigma_2)) \\ &= \mathrm{E}\left( \sum_{\{i,j\} \in \mathcal{P}(\tau_1, \tau_2)} \bar{K}_{i,j}(\sigma_1, \sigma_2) \right) \\ &= \sum_{\{i,j\} \in \mathcal{P}(\tau_1, \tau_2)} \mathrm{E}\left(\bar{K}_{i,j}(\sigma_1, \sigma_2)\right) \end{aligned} \tag{8}$$

We shall show that

$$\mathrm{E}\left(\bar{K}_{i,j}(\sigma_1, \sigma_2)\right) = \bar{K}_{i,j}^{(1/2)}(\tau_1, \tau_2). \tag{9}$$

This proves that $K_{\mathrm{avg}} = K^{(1/2)}$, since the result of substituting $\bar{K}_{i,j}^{(1/2)}(\tau_1, \tau_2)$ for $\mathrm{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2))$ in (8) gives $K^{(1/2)}(\tau_1, \tau_2)$. Similarly to before, when $\{i, j\}$ is not a special pair, we have $\bar{K}_{i,j}(\sigma_1, \sigma_2) = \bar{K}^{(1/2)}(\tau_1, \tau_2)$, and so (9) holds. When $\{i, j\}$ is a special pair, then $\bar{K}_{i,j}^{(1/2)}(\tau_1, \tau_2) = 1/2$. So we are done with showing that $K_{\mathrm{avg}} = K^{(1/2)}$ if we show that when $\{i, j\}$ is a special pair, then $\mathrm{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2)) = 1/2$. Assume without loss of generality that $i, j$ are both in $D_{\tau_1}$ but neither is in $D_{\tau_2}$. The ordering of $i, j$ in $\sigma_1$ is forced by $\tau_1$. Further, there is a one-one correspondence between those permutations $\sigma_2$ that extend $\tau_2$ with $i$ preceding $j$ and those that extend $\tau_2$ with $j$ preceding $i$ (the correspondence is determined by simply switching $i$ and $j$). Therefore, for each choice of $\sigma_1$, exactly half of the choices for $\sigma_2$ have $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 0$, and for the other half, $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 1$. So $\mathrm{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2)) = 1/2$, as desired.

We now show that $K_{\mathrm{Haus}} = K^{(1/2)}$. If we set $p = 1/2$ in our formula for $K^{(p)}$ given in Lemma 3.1, we obtain the right-hand side of the equation in Lemma 3.2. Thus, $K_{\mathrm{Haus}} = K^{(1/2)}$. We now give a direct proof, that does not require the use of Lemma 3.2, and hence does not require the use of Critchlow's formula given by Equation (6).

Let $\tau_1, \tau_2$ be top $k$ lists. Then $K_{\mathrm{Haus}}(\tau_1, \tau_2)$ is given by

$$\max\left\{ \max_{\sigma_1 \succeq \tau_1} \min_{\sigma_2 \succeq \tau_2} K(\sigma_1, \sigma_2), \max_{\sigma_2 \succeq \tau_2} \min_{\sigma_1 \succeq \tau_1} K(\sigma_1, \sigma_2) \right\}.$$

Let $\sigma_1^*$ be the permutation over $D_{\tau_1} \cup D_{\tau_2}$ where $\sigma_1^* \succeq \tau_1$ and where $\sigma_1^*(k+1), \ldots, \sigma_1^*(2k-z)$ are, respectively, the members of $T$ in reverse order. It is easy to see that

$$\max_{\sigma_1 \succeq \tau_1} \min_{\sigma_2 \succeq \tau_2} K(\sigma_1, \sigma_2) = \min_{\sigma_2 \succeq \tau_2} K(\sigma_1^*, \sigma_2),$$

and that in fact

$$K_{\mathrm{Haus}}(\tau_1, \tau_2) = \min_{\sigma_2 \succeq \tau_2} K(\sigma_1^*, \sigma_2).$$

Let $\sigma_2^*$ be the permutation over $D_{\tau_1} \cup D_{\tau_2}$ where $\sigma_2^* \succeq \tau_2$ and where $\sigma_2^*(k+1), \ldots, \sigma_2^*(2k-z)$ are, respectively, the members of $S$ in order (not in reverse order). It is easy to see that $\min_{\sigma_2 \succeq \tau_2} K(\sigma_1^*, \sigma_2) = K(\sigma_1^*, \sigma_2^*)$. Therefore, $K_{\mathrm{Haus}}(\tau_1, \tau_2) = K(\sigma_1^*, \sigma_2^*)$. So we need only show that $K(\sigma_1^*, \sigma_2^*) = K^{(1/2)}(\tau_1, \tau_2)$.

In the definition of $K^{(p)}$, let us consider the contribution of each pair $\{i, j\}$ to $K^{(1/2)}(\tau_1, \tau_2)$, as compared to its contribution to $K(\sigma_1^*, \sigma_2^*)$. In the first three cases in the definition of $K^{(p)}$, it is easy to see that $\{i, j\}$ contributes exactly the same to $K^{(1/2)}(\tau_1, \tau_2)$ as to $K(\sigma_1^*, \sigma_2^*)$. Let us now consider Case 4, where $\{i, j\}$ is a special pair, that is, where both $i$ and $j$ appear in one of the top $k$ lists $\tau_1$ or $\tau_2$, but neither appears in the other top $k$ list. If both $i$ and $j$ appear in $\tau_1$ but neither appears in $\tau_2$, then the contribution to

$K^{(1/2)}(\tau_1, \tau_2)$ is $1/2$, and the contribution to $K(\sigma_1^*, \sigma_2^*)$ is 0. If both $i$ and $j$ appear in $\tau_2$ but neither appears in $\tau_1$, then the contribution to $K^{(1/2)}(\tau_1, \tau_2)$ is $1/2$, and the contribution to $K(\sigma_1^*, \sigma_2^*)$ is 1. Since there are just as many pairs $\{i, j\}$ of the first type (where both $i$ and $j$ appear in $\tau_1$ but neither appears in $\tau_2$) as there are of the second type (where both $i$ and $j$ appear in $\tau_2$ but neither appears in $\tau_1$), the total contribution of all pairs $\{i, j\}$ of Case 4 to $K^{(1/2)}(\tau_1, \tau_2)$ and $K(\sigma_1^*, \sigma_2^*)$ is the same. This proves that $K_{\text{Haus}} = K^{(1/2)}$. □

### 3.2 Spearman's footrule

We now generalize Spearman's footrule to several methods for determining distances between top $k$ lists, just as we did for Kendall's tau.

For top $k$ lists $\tau_1$ and $\tau_2$, the *minimizing footrule distance* $F_{\min}(\tau_1, \tau_2)$ between $\tau_1$ and $\tau_2$ is defined to be the minimum value of $F(\sigma_1, \sigma_2)$, where $\sigma_1$ and $\sigma_2$ are each permutations of $D$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$.

For top $k$ lists $\tau_1$ and $\tau_2$, the *averaging footrule distance* $F_{\text{avg}}(\tau_1, \tau_2)$ between $\tau_1$ and $\tau_2$ is defined to be the expected value $\text{E}(F(\sigma_1, \sigma_2))$, where $\sigma_1$ and $\sigma_2$ are each permutations of $D_{\tau_1} \cup D_{\tau_2}$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$. Again, $\text{E}(\cdot)$ gives the expected value where all extensions are taken to be equally likely.

Let $\ell$ be a real number greater than $k$. The *footrule distance with location parameter* $\ell$, denoted $F^{(\ell)}$, is obtained, intuitively, by placing all missing elements in each of the lists at position $\ell$ and computing the usual footrule distance between them. More formally, given top $k$ lists $\tau_1$ and $\tau_2$, define functions $\tau_1'$ and $\tau_2'$ with domain $D_{\tau_1} \cup D_{\tau_2}$ by letting $\tau_1'(i) = \tau_1(i)$ for $i \in D_{\tau_1}$, and $\tau_1'(i) = \ell$ otherwise, and similarly defining $\tau_2'$. We then define $F^{(\ell)}$ by setting $F^{(\ell)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1} \cup D_{\tau_2}} |\tau_1'(i) - \tau_2'(i)|$.

A natural choice for $\ell$ is $k + 1$, and we make this choice in our experiments (Section 9). We denote $F^{(k+1)}$ simply by $F^*$.

The next lemma gives a formula, which we shall find useful later, for $F^{(\ell)}$.

**Lemma 3.5.** $F^{(\ell)}(\tau_1, \tau_2) = 2(k - z)\ell + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i)$.

*Proof.*

$$\begin{aligned} F^{(\ell)}(\tau_1, \tau_2) &= \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| + \sum_{i \in S} |\tau_1(i) - \tau_2(i)| + \sum_{i \in T} |\tau_1(i) - \tau_2(i)| \\ &= \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| + \sum_{i \in S} (\ell - \tau_1(i)) + \sum_{i \in T} (\ell - \tau_2(i)) \\ &= 2(k - z)\ell + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i). \end{aligned}$$

□

Similarly to our definition of $K_{\text{Haus}}$, we define $F_{\text{Haus}}$, the Hausdorff version of the footrule distance between top $k$ lists, to be given by Equation (5) with $d(\sigma_1, \sigma_2)$ as the footrule distance $F(\sigma_1, \sigma_2)$, where, as before, we take $\sigma_1$ and $\sigma_2$ to be permutations of $D_{\tau_1} \cup D_{\tau_2}$.

Just as he did with the Kendall distance, Critchlow considered his version of Equation (5) when $d(\sigma_1, \sigma_2)$ is the footrule distance $F(\sigma_1, \sigma_2)$, and where there is a fixed domain of size $n$. Again, he obtained a closed formula, given by

$$(k - z)(2n + 1 - (k - z)) + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i).$$

By replacing $n$ by $2k - z$, we obtain a closed form for $F_{\text{Haus}}$:

**Lemma 3.6.**

$$\begin{aligned} F_{\text{Haus}}(\tau_1, \tau_2) &= (k-z)(3k-z+1) + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i) \\ &= F^{(\frac{3k-z+1}{2})}(\tau_1, \tau_2). \end{aligned}$$

The last equality is obtained by formally substituting $\ell = (3k - z + 1)/2$ into the formula for $F^{(\ell)}$ given by Lemma 3.5. Thus, intuitively, $F_{\text{Haus}}(\tau_1, \tau_2)$ is a "dynamic" version of $F^{(\ell)}$ where $\ell = (3k - z + 1)/2$ actually depends on $\tau_1$ and $\tau_2$. Since $F_{\text{min}} = F_{\text{avg}} = F_{\text{Haus}}$ (Proposition 3.7), this gives us a formula for $F_{\text{min}}$ and $F_{\text{avg}}$ as well. Note that $\ell = (3k - z + 1)/2$ is the average of $k + 1$ and $2k - z$, where the latter number is the size of $D = D_{\tau_1} \cup D_{\tau_2}$. Since taking $\ell = (3k - z + 1)/2$ corresponds intuitively to "placing the missing elements at an average location," it is not surprising that the resulting formula gives $F_{\text{avg}}$. Unlike the situation with $K_{\text{min}}$ and $K_{\text{avg}}$, the next proposition tells us that $F_{\text{min}}$ and $F_{\text{avg}}$ are the same. Furthermore, the Hausdorff distance $F_{\text{Haus}}$ shares this common value.

**Proposition 3.7.** $F_{\text{min}} = F_{\text{avg}} = F_{\text{Haus}}$.

*Proof.* We first show that $F_{\text{min}} = F_{\text{avg}}$. Let $\tau_1$ and $\tau_2$ be top $k$ lists. Let $\sigma_1, \sigma_1', \sigma_2, \sigma_2'$ be permutations of $D = D_{\tau_1} \cup D_{\tau_2}$, where $\sigma_1$ and $\sigma_1'$ extend $\tau_1$, and where $\sigma_2$ and $\sigma_2'$ extend $\tau_2$. We need only show that $F(\sigma_1, \sigma_2) = F(\sigma_1', \sigma_2')$. Therefore, we need only show that $F(\sigma_1, \sigma_2) = F(\sigma_1, \sigma_2')$, where $\sigma_1$ is held fixed, since by symmetry (where $\sigma_2'$ is held fixed) we would then have $F(\sigma_1, \sigma_2') = F(\sigma_1', \sigma_2')$, and hence $F(\sigma_1, \sigma_2) = F(\sigma_1, \sigma_2') = F(\sigma_1', \sigma_2')$, as desired.

Now $F(\sigma_1, \sigma_2) = \sum_{i \in D} |\sigma_1(i) - \sigma_2(i)|$. So we need only show that

$$\sum_{i \in D} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in D} |\sigma_1(i) - \sigma_2'(i)|. \tag{10}$$

Now

$$\sum_{i \in D} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in S} |\sigma_1(i) - \sigma_2(i)|, \tag{11}$$

and similarly

$$\sum_{i \in D} |\sigma_1(i) - \sigma_2'(i)| = \sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2'(i)| + \sum_{i \in S} |\sigma_1(i) - \sigma_2'(i)|. \tag{12}$$

Now $\sigma_2(i) = \sigma_2'(i)$ for $i \in D_{\tau_2}$. Hence,

$$\sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2'(i)|. \tag{13}$$

From (11), (12), and (13), it follows that to prove (10), and hence complete the proof, it is sufficient to prove

$$\sum_{i \in S} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in S} |\sigma_1(i) - \sigma_2'(i)|. \tag{14}$$

If $i \in S$, then $\sigma_1(i) \leq k < \sigma_2(i)$. Thus, if $i \in S$, then $\sigma_1(i) < \sigma_2(i)$, and similarly $\sigma_1(i) < \sigma_2'(i)$. So it is sufficient to prove

$$\sum_{i \in S} \sigma_1(i) - \sigma_2(i) = \sum_{i \in S} \sigma_1(i) - \sigma_2'(i),$$

and hence to prove

$$\sum_{i \in S} \sigma_2(i) = \sum_{i \in S} \sigma_2'(i). \tag{15}$$

But both the left-hand side and the right-hand side of (15) equal $\sum_{\ell=k+1}^{|D|} \ell$, and hence are equal. This completes the proof that $F_{\min} = F_{\text{avg}}$.

We now consider $F_{\text{Haus}}$. We have shown that the minimal value $F_{\min}$ of $F(\sigma_1, \sigma_2)$, where $\sigma_1$ and $\sigma_2$ are each permutations of $D$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$, equals the average value $F_{\text{avg}}$. Since the minimal value equals the average value, it is clear that all of these values of $F(\sigma_1, \sigma_2)$ are necessarily the same. It follows easily that $F_{\text{Haus}}$ equals this common value, and so $F_{\min} = F_{\text{avg}} = F_{\text{Haus}}$. □

### 3.3 Metric properties

We have now introduced three distinct measures of closeness between top $k$ lists: (1) $K^{(p)}$, which has $K_{\min}$ and $K_{\text{avg}} = K_{\text{Haus}}$ as special cases for certain choices of $p$; (2) $F_{\min}$, which equals $F_{\text{avg}}$ and $F_{\text{Haus}}$; and (3) $F^{(\ell)}$. Perhaps the most natural question, and the main subject of our investigation, is to ask is whether or not they are metrics.

As a preview to our main results, we begin by observing that while $F^{(\ell)}$ is a metric, none of the other distance measures that we have defined (namely, $K^{(p)}$ and $F_{\min}$, hence also $K_{\min}, K_{\text{avg}}, F_{\text{avg}}, F_{\text{Haus}}$) is a metric.

**Proposition 3.8.** *The distance measure $F^{(\ell)}$ is a metric for every choice of the location parameter $\ell$.*

*Proof.* We need only show that the triangle inequality holds. Let $\tau_1, \tau_2, \tau_3$ be top $k$ lists. Let $n = |D_{\tau_1} \cup D_{\tau_2} \cup D_{\tau_3}|$. Define an $n$-dimensional vector $v_1$ corresponding to $\tau_1$ by letting $v_1(i) = \tau_1(i)$ for $i \in D_{\tau_1}$, and $\ell$ otherwise. Similarly, define an $n$-dimensional vector $v_2$ corresponding to $\tau_2$ and an $n$-dimensional vector $v_3$ corresponding to $\tau_3$. It is easy to see that $F^{(\ell)}(\tau_1, \tau_2)$ is the $L_1$ distance between $v_1$ and $v_2$, and similarly for $F^{(\ell)}(\tau_1, \tau_3)$ and $F^{(\ell)}(\tau_2, \tau_3)$. The triangle inequality for $F^{(\ell)}$ then follows immediately from the triangle inequality for the $L_1$ norm between two vectors in $n$-dimensional Euclidean space. □

The other two distinct distance measures, namely $K^{(p)}$ and $F_{\min}$, are not metrics, as we now show. Let $\tau_1$ be the top 2 list where the top 2 items in order are 1,2; let $\tau_2$ be the top 2 list where the top 2 items in order are 1,3; and let $\tau_3$ be the top 2 list where the top 2 items in order are 3,4. It is straightforward to verify that $K^{(p)}(\tau_1, \tau_2) = 1$; $K^{(p)}(\tau_1, \tau_3) = 4 + 2p$; and $K^{(p)}(\tau_2, \tau_3) = 2$. So the triangle inequality fails, because $K^{(p)}(\tau_1, \tau_3) > K^{(p)}(\tau_1, \tau_2) + K^{(p)}(\tau_2, \tau_3)$ for every $p \geq 0$. Therefore, $K^{(p)}$ is not a metric, no matter what the choice of the penalty parameter $p$ is; in particular, by Propositions 3.3 and 3.4, neither $K_{\min}$ nor $K_{\text{avg}}$ is a metric.

The same counterexample shows that $F_{\min}$ is not a metric. In this case, it is easy to verify that $F_{\min}(\tau_1, \tau_2) = 2$; $F_{\min}(\tau_1, \tau_3) = 8$; and $F_{\min}(\tau_2, \tau_3) = 4$. So the triangle inequality fails, because $F_{\min}(\tau_1, \tau_3) > F_{\min}(\tau_1, \tau_2) + F_{\min}(\tau_2, \tau_3)$.

The fact that $F_{\min}$ (and hence $F_{\text{avg}}$ and $F_{\text{Haus}}$) are not metrics shows that they are not special cases of $F^{(\ell)}$, since $F^{(\ell)}$ is a metric. This is in contrast to the situation with Kendall distances, where $K_{\min}$, $K_{\text{avg}}$, and $K_{\text{Haus}}$ are special cases of $K^{(p)}$. (As we noted earlier, the versions of $F_{\text{Haus}}$ and $K_{\text{Haus}}$ defined by Critchlow [Cri80] are indeed metrics, since the domain is fixed in his case.)

## 4 Metrics, near metrics, and equivalence classes

Motivated by the fact that most of our distance measures are not metrics (except for the somewhat strange measure $F^{(\ell)}$), we next consider a precise sense in which each is a "near metric." Actually, we shall consider

two quite different-appearing notions of being a near metric, which $K^{(p)}$ and $F_{\min}$ satisfy, and obtain the surprising result that these notions are actually equivalent.

Our first notion of near metric is based on "relaxing" the triangle inequality (or more generally, the polygonal inequality) that a metric is supposed to satisfy.

**Definition 4.1 (Relaxed inequalities).** A binary function $d$ satisfies the *$c$-triangle inequality* if $d(x,z) \leq c(d(x,y)+d(y,z))$ for all $x, y, z$ in the domain. A binary function $d$ satisfies the *$c$-polygonal inequality* if $d(x,z) \leq c(d(x,x_1)+d(x_1,x_2)+\cdots+d(x_{n-1},z))$ for all $n > 0$ and $x, z, x_1, \ldots, x_{n-1}$ in the domain.

The notion of $c$-triangle inequality, to our knowledge, appears to be rarely studied. It has been used in a paper on pattern matching [FS98], and in the context of the traveling salesperson problem [AB95, BC00]. We do not know if the $c$-polygonal inequality has ever been studied.

**Definition 4.2 (Relaxed metrics).** A *$c$-relaxed$_{\rm t}$ metric* is a distance measure that satisfies the $c$-triangle inequality. A *$c$-relaxed$_{\rm p}$ metric* is a distance measure that satisfies the $c$-polygonal inequality.

Of course, every $c$-relaxed$_{\rm p}$ metric is a $c$-relaxed$_{\rm t}$ metric. Theorem 4.7 below says that there is a $c$-relaxed$_{\rm t}$ metric that is not a $c'$-relaxed$_{\rm p}$ metric for any constant $c'$. We shall focus here on the stronger notion of being a $c$-relaxed$_{\rm p}$ metric.

The other notion of near metric that we now discuss is based on bounding the distance measure above and below by positive constant multiples of a metric.

**Definition 4.3 (Metric boundedness).** A *$(c_1, c_2)$-metric-bounded distance measure* is a distance measure $d$ for which there is a metric $d'$ and positive constants $c_1$ and $c_2$ such that $c_1 d'(x,y) \leq d(x,y) \leq c_2 d'(x,y)$.

Note that without loss of generality, we can take $c_1 = 1$ (by replacing the metric $d'$ by the metric $c_1 d'$). In this case, we say that $d$ is *$c_2$-metric bounded*.

The next theorem gives the unexpected result that our two notions of near metric are equivalent (and even with the same value of $c$).

**Theorem 4.4 (MAIN RESULT 1).** *Let $d$ be a distance measure. Then $d$ is a $c$-relaxed$_{\rm p}$ metric iff $d$ is $c$-metric-bounded.*

*Proof.* $\Longleftarrow$: Assume that $d$ is a $c$-relaxed$_{\rm p}$ metric. Define $d'$ by taking

$$d'(x,z) = \min_{\ell} \min_{y_0,\ldots,y_\ell \,|\, y_0=x \text{ and } y_\ell=z} \sum_{i=0}^{\ell-1} d(y_i, y_{i+1}). \tag{16}$$

We now show that $d'$ is a metric.

First, we have $d'(x,x) = 0$ since $d(x,x) = 0$. From (16) and the polygonal inequality with constant $c$, we have $d'(x,z) \geq (1/c)d(x,z)$. Hence, $d'(x,z) \neq 0$ if $x \neq z$. Symmetry of $d'$ follows immediately from symmetry of $d$. Finally, $d'$ satisfies the triangle inequality, since

$$\begin{aligned} d'(x,z) &= \min_{\ell} \min_{y_0,\ldots,y_\ell \,|\, y_0=x \text{ and } y_\ell=z} \sum_{i=0}^{\ell-1} d(x_i, x_{i+1}) \\ &\leq \min_{\ell_1} \min_{y_0,\ldots,y_{\ell_1} \,|\, y_0=x \text{ and } y_{\ell_1}=y} \sum_{i=0}^{\ell_1-1} d(y_i, y_{i+1}) + \min_{\ell_2} \min_{z_0,\ldots,z_{\ell_1} \,|\, z_0=y \text{ and } z_{\ell_2}=z} \sum_{i=0}^{\ell_2-1} d(z_i, z_{i+1}) \\ &= d'(x,y) + d'(y,z). \end{aligned}$$

Therefore, $d'$ is a metric.

We now show that $d$ is $c$-metric-bounded. By Equation (16), it follows easily that $d'(x,z) \leq d(x,z)$. By Equation (16) and the polygonal inequality with constant $c$, we have $d(x,z) \leq cd'(x,z)$.

$\Longrightarrow$: Assume that $d$ is $c$-metric-bounded. Then $0 = d'(x,x) \leq d(x,x) \leq cd'(x,x) = 0$. Therefore, $d(x,x) = 0$. If $x \neq y$, then $d(x,y) \geq d'(x,y) > 0$. We now show that $d$ satisfies the $c$-polygonal inequality.

$$\begin{aligned} d(x,z) &\leq cd'(x,z) \\ &\leq c(d'(x,x_1) + d'(x_1,x_2) + \cdots + d'(x_{n-1},z)) \text{ since } d' \text{ is a metric} \\ &\leq c(d(x,x_1) + d(x_1,x_2) + \cdots + d(x_{n-1},z)) \text{ since } d'(x,y) \leq d(x,y). \end{aligned}$$

Since also $d$ is symmetric by assumption, it follows that $d$ is a $c$-relaxed$_{\mathrm{p}}$ metric. □

Inspired by Theorem 4.4, we now define what it means for a distance measure to be "almost" a metric, and a robust notion of "similar" or "equivalent" distance measures.

**Definition 4.5 (Near metric).** A distance measure between top $k$ lists is a *near metric* if there is a constant $c$, independent of $k$, such that the distance measure is a $c$-relaxed$_{\mathrm{p}}$ metric (or, equivalently, is $c$-metric-bounded).[1]

**Definition 4.6 (Equivalent distance measures).** Two distance measures $d$ and $d'$ between top $k$ lists are *equivalent* if there are positive constants $c_1$ and $c_2$ such that $c_1 d'(\tau_1,\tau_2) \leq d(\tau_1,\tau_2) \leq c_2 d'(\tau_1,\tau_2)$, for every pair $\tau_1, \tau_2$ of top $k$ lists.[2]

It is easy to see that this definition of equivalence actually gives us an equivalence relation (reflexive, symmetric, and transitive). It follows from Theorem 4.4 that a distance measure is equivalent to a metric if and only if it is a near metric.

Our notion of equivalence is inspired by a classical result of Diaconis and Graham [DG77], which states that for every two permutations $\sigma_1, \sigma_2$, we have

$$K(\sigma_1,\sigma_2) \leq F(\sigma_1,\sigma_2) \leq 2K(\sigma_1,\sigma_2). \tag{17}$$

(Of course, we are dealing with distances between top $k$ lists, whereas Diaconis and Graham dealt with distances between permutations.)

Having showed that the notions of $c$-relaxed$_{\mathrm{p}}$ metric and $c$-metric-boundedness are identical, we compare these to the notions of $c$-relaxed$_{\mathrm{t}}$ metric and the classical topological notion of being a topoligical metric, that is, of generating a metrizable topology.

**Theorem 4.7.** *Every $c$-relaxed$_{\mathrm{p}}$ metric is a $c$-relaxed$_{\mathrm{t}}$ metric, but not conversely. In fact, there is a $c$-relaxed$_{\mathrm{t}}$ metric that is not a $c'$-relaxed$_{\mathrm{p}}$ metric for any constant $c'$.*

*Proof.* It is clear that every $c$-relaxed$_{\mathrm{p}}$ metric is a $c$-relaxed$_{\mathrm{t}}$ metric. We now show that the converse fails. Define $d$ on the space $[0,1]$ by taking $d(x,y) = (x-y)^2$. It is clear that $d$ is a symmetric function with $d(x,y) = 0$ iff $x = y$. To show the 2-triangle inequality, let $\alpha = d(x,z)$, $\beta = d(x,y)$, and $\gamma = d(y,z)$. Now $\sqrt{\alpha} \leq \sqrt{\beta} + \sqrt{\gamma}$, since the function $d'$ with $d'(x,y) = |x-y|$ is a metric. By squaring both sides, we get $\alpha \leq \beta + \gamma + 2\sqrt{\beta\gamma}$. But $\sqrt{\beta\gamma} \leq (\beta+\gamma)/2$ by the well-known fact that the geometric mean is bounded

[1] It makes sense to say that the constant $c$ is independent of $k$, since each of our distance measures is actually a family, parameterized by $k$. We need to make an assumption that $c$ is independent of $k$, since otherwise we are simply considering distance measures over finite domains, where there is always such a constant $c$.

[2] As before, the constants $c_1$ and $c_2$ are assumed to be independent of $k$.

above by the arithmetic mean. We therefore obtain $\alpha \leq 2(\beta + \gamma)$, that is, $d(x, z) \leq 2(d(x, y) + d(y, z))$. So $d$ is a 2-relaxed$_{\mathrm{t}}$ metric.

Let $n$ be an arbitrary positive integer, and define $x_i$ to be $i/n$ for $1 \leq i \leq n-1$. Then $d(0, x_1) + d(x_1, x_2) + \cdots + d(x_{n-1}, 1) = n(1/n^2) = 1/n$. Since this converges to 0 as $n$ goes to infinity, and since $d(0, 1) = 1$, there is no constant $c'$ for which $d$ satisfies the polygonal inequality. Therefore, $d$ is a $c$-relaxed$_{\mathrm{t}}$ metric that is not a $c'$-relaxed$_{\mathrm{p}}$ metric for any constant $c'$. □

**Theorem 4.8.** *Every $c$-relaxed$_{\mathrm{t}}$ metric is a topological metric, but not conversely. The converse fails even if we restrict attention to distance measures.*

*Proof.* By the *topological space induced by a binary function $d$*, we mean the topological space whose open sets are precisely the union of sets ("$\epsilon$-balls") of the form $\{y \mid d(x, y) < \epsilon\}$. A topological space is *metrizable* if there is a metric $d$ that induces the topology. A *topological metric* is a binary function $d$ such that the topology induced by $d$ is metrizable.

There is a theorem of Nagata and Smirnov [Dug66, pp. 193–195] that a topological space is metrizable if and only if it is regular and has a basis that can be decomposed into an at most countable collection of nbd-finite families. The proof of the "only if" direction can be modified in an obvious manner to show that every topological space induced by a relaxed$_{\mathrm{t}}$ metric is regular and has a basis that can be decomposed into an at most countable collection of nbd-finite families. It follows that a topological space is metrizable if and only if it is induced by a $c$-relaxed$_{\mathrm{t}}$ metric. That is, every $c$-relaxed$_{\mathrm{t}}$ metric is a topological metric.

We now show that the converse fails even if we restrict attention to distance measures (binary nonnegative functions $d$ that are symmetric and satisfy $d(x, y) = 0$ iff $x = y$). Define $d$ on the space $[1, \infty)$ by taking $d(x, y) = |y - x|^{\max\{x,y\}}$. It is not hard to verify that $d$ induces the same topology as the usual metric $d'$ with $d'(x, y) = |x - y|$. The intuition is that (1) the $\epsilon$-ball $\{y \mid d(x, y) < \epsilon\}$ is just a minor distortion of an $\epsilon$-ball $\{y \mid d_m(x, y) < \epsilon\}$ where $d_m(x, y) = |x - y|^m$ for some $m$ that depends on $x$ (in fact, with $m = x$), and (2) the function $d_m$ locally induces the same topology as the usual metric $d'$ with $d'(x, y) = |x - y|$. Condition (2) holds since the ball $\{y \mid |x - y|^m < \epsilon\}$ is the same as the ball $\left\{y \mid |x - y| < \epsilon^{1/m}\right\}$. So $d$ is a topological metric. We now show that $d$ is not a $c$-relaxed$_{\mathrm{t}}$ metric.

Let $x = 1$, $y = n + 1$, and $z = 2n + 1$. We shall show that for each constant $c$, there is $n$ such that

$$d(x, z) > c(d(x, y) + d(y, z)). \tag{18}$$

This implies that $d$ is not a relaxed$_{\mathrm{t}}$ metric. When we substitute for $x, y, z$ in (18), we obtain

$$(2n + 1)^{2n+1} > c((n + 1)^{n+1} + (n + 1)^{2n+1}). \tag{19}$$

But it is easy to see that (19) holds for every sufficiently large $n$. □

Thus, we have METRIC $\Rightarrow$ $c$-RELAXED$_{\mathrm{p}}$ METRIC $\Rightarrow$ $c$-RELAXED$_{\mathrm{t}}$ METRIC $\Rightarrow$ TOPOLOGICAL METRIC, and none of the reverse implications hold.

# 5 Relationships between measures

We now come to one of the main results of the paper, where we show that all of our distance measures we have discussed are in the same equivalence class, that is, are bounded by constant multiples of each other both above and below. The connections are proved via two proof methods. We use direct counting arguments to relate $F^*$ with $F_{\min}$, to relate the $K^{(p)}$ measures with each other, and to relate the $F^{(\ell)}$ measures with each other. The more subtle connection between $K_{\min}$ and $F_{\min}$—which provides the link between the measures based on Kendall's tau and the measures based on Spearman's footrule—is proved by applying Diaconis and Graham's inequalities (17) for permutations $\sigma_1, \sigma_2$.

**Theorem 5.1 (MAIN RESULT 2).** *The distance measures $K_{\min}$, $K_{\text{avg}}$, $K_{\text{Haus}}$, $K^{(p)}$ (for every choice of $p$), $F_{\min}$, $F_{\text{avg}}$, $F_{\text{Haus}}$, and $F^{(\ell)}$ (for every choice of $\ell$) are all in the same equivalence class.*

The fact that $F^{(\ell)}$ is a metric now implies that all our distance measures are near metrics.

**Corollary 5.2.** *Each of $K^{(p)}$ and $F_{\min}$ (thus also $K_{\min}, K_{\text{avg}}, K_{\text{Haus}}, F_{\text{avg}}, F_{\text{Haus}}$) is a near metric.*

We discuss the proof of this theorem shortly. We refer to the equivalence class that contains all of these distance measures as the *big equivalence class*. The big equivalence class seems to be quite robust. As we have seen, it consists of distance measures, some of which are metrics.

In later sections, we shall find it convenient to deal with normalized versions of our distance measures, by dividing each distance measure by its maximum value. The normalized version is then a distance measure that lies in the interval $[0, 1]$.[3] The normalized version is a metric if the original version is a metric, and is a near metric if the original version is a near metric. It is easy to see that if two distance measures are in the same equivalence class, then so are their normalized versions.

Theorem 5.1 is proven by making use of the following theorem (Theorem 5.3), along with Propositions 3.3, 3.4, and 3.7. The bounds in Theorem 5.3 are not tight (we have improved some of them, with more complicated proofs). Our goal was simply to prove enough to obtain Theorem 5.1. If we really wished to obtain tight results, we would have to compare every pair of the distance measures we have introduced, such as $K^{(p)}$ versus $F^{(\ell)}$ for arbitrary $p, \ell$.

**Theorem 5.3.** *Let $\tau_1, \tau_2$ be top $k$ lists.*
*(1)* $K_{\min}(\tau_1, \tau_2) \le F_{\min}(\tau_1, \tau_2) \le 2K_{\min}(\tau_1, \tau_2)$;
*(2)* $F^*(\tau_1, \tau_2) \le F_{\min}(\tau_1, \tau_2) \le 2F^*(\tau_1, \tau_2)$;
*(3)* $K^{(p)}(\tau_1, \tau_2) \le K^{(p')}(\tau_1, \tau_2) \le (\frac{1+p'}{1+p})K^{(p)}(\tau_1, \tau_2)$, *for* $0 \le p \le p' \le 1$;
*(4)* $F^{(\ell)}(\tau_1, \tau_2) \le F^{(\ell')}(\tau_1, \tau_2) \le (\frac{\ell'-k}{\ell-k})F^{(\ell)}(\tau_1, \tau_2)$, *for* $k < \ell \le \ell'$;

*Proof.* (Part (1))

For the first inequality of Part (1), let $\sigma_1, \sigma_2$ be permutations so that $\sigma_1 \succeq \tau_1$, $\sigma_2 \succeq \tau_2$, and $F_{\min}(\tau_1, \tau_2) = F(\sigma_1, \sigma_2)$. Then $F_{\min}(\tau_1, \tau_2) = F(\sigma_1, \sigma_2) \ge K(\sigma_1, \sigma_2) \ge K_{\min}(\tau_1, \tau_2)$, using the first inequality in (17) and the fact that $K_{\min}$ is the minimum over all extensions $\sigma_1$ of $\tau_1$ and $\sigma_2$ of $\tau_2$.

For the second inequality of Part (1), let $\sigma_1, \sigma_2$ be permutations so that $\sigma_1 \succeq \tau_1$, $\sigma_2 \succeq \tau_2$, and $K_{\min}(\tau_1, \tau_2) = K(\sigma_1, \sigma_2)$. Then $K_{\min}(\tau_1, \tau_2) = K(\sigma_1, \sigma_2) \ge (1/2)F(\sigma_1, \sigma_2) \ge (1/2)F_{\min}(\tau_1, \tau_2)$ using the second inequality in (17) and the fact that $F_{\min}$ is minimum over all extensions $\sigma_1$ of $\tau_1$ and $\sigma_2$ of $\tau_2$.

(Part (2))

Let $\sigma_1, \sigma_2$ be permutations so that $\sigma_1 \succeq \tau_1$, $\sigma_2 \succeq \tau_2$, and $F_{\min}(\tau_1, \tau_2) = F(\sigma_1, \sigma_2)$. For $s \in \{1, 2\}$, let $v_s$ be a vector such that $v_s(i) = \tau_s(i)$ if $i \in D_{\tau_s}$ and $v_s(i) = k + 1$ otherwise. Given $\tau_1, \tau_2$, recall that $F^*(\tau_1, \tau_2)$ is exactly the $L_1$ distance between the corresponding vectors $v_1, v_2$. If $i \in Z = D_{\tau_1} \cap D_{\tau_2}$, then $|v_1(i) - v_2(i)| = |\sigma_1(i) - \sigma_2(i)|$. If $i \in S = D_{\tau_1} \setminus D_{\tau_2}$, then $|v_1(i) - v_2(i)| = |\tau_1(i) - (k+1)| = |\sigma_1(i) - (k+1)| \le |\sigma_1(i) - \sigma_2(i)|$, since $\sigma_2(i) \ge k + 1 > \tau_1(i) = \sigma_1(i)$. The case of $i \in T = D_{\tau_2} \setminus D_{\tau_1}$ is similar. Thus, for every $i$, we have $|v_1(i) - v_2(i)| \le |\sigma_1(i) - \sigma_2(i)|$. It follows by definition that $F^*(\tau_1, \tau_2) \le F(\sigma_1, \sigma_2) = F_{\min}(\tau_1, \tau_2)$. This proves the first inequality.

We now prove the second inequality. First, we have

$$F_{\min}(\tau_1, \tau_2) = \sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in S} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in T} |\sigma_1(i) - \sigma_2(i)|. \quad (20)$$

[3] For metrics on permutations, such as Kendall's tau and Spearman's footrule, it is standard to normalize them to lie in the interval $[-1, 1]$, with $-1$ corresponding to the situation where the permutations are the reverse of each other, and with 1 corresponding to the situation where the permutations are equal.

On the other hand, we have

$$F^*(\tau_1, \tau_2) = \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| + \sum_{i \in S} |\tau_1(i) - (k+1)| + \sum_{i \in T} |(k+1) - \tau_2(i)|. \tag{21}$$

Furthermore, if $z = |Z|$, note that

$$\begin{aligned} \sum_{i \in S} |\tau_1(i) - (k+1)| &\geq \sum_{r=z+1}^{k} |r - (k+1)| \\ &= (k-z) + \cdots + 1 \\ &= \frac{(k-z)(k-z+1)}{2}. \end{aligned} \tag{22}$$

By symmetry, we also have $\sum_{i \in T} |(k+1) - \tau_2(i)| \geq (k-z)(k-z+1)/2$.

For $i \in Z$, we have $|\sigma_1(i) - \sigma_2(i)| = |\tau_1(i) - \tau_2(i)|$ and so,

$$\sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in Z} |\tau_1(i) - \tau_2(i)|. \tag{23}$$

Since $\sigma_2(i) \geq k+1$ and $\tau_1(i) \leq k$ if and only if $i \in S$, we have, for $i \in S$, that $|\tau_1(i) - \sigma_2(i)| = |\tau_1(i) - (k+1)| + (\sigma_2(i) - (k+1))$. Furthermore, since $\sigma_2$ is a permutation, the list of values $\sigma_2(i), i \in S$ is precisely $k+1, \ldots, 2k-z$. Summing over all $i \in S$, this yields

$$\begin{aligned} \sum_{i \in S} |\sigma_1(i) - \sigma_2(i)| &= \sum_{i \in S} |\tau_1(i) - \sigma_2(i)| \\ &= 0 + 1 + \cdots + (k-z-1) + \sum_{i \in S} |\tau_1(i) - (k+1)| \\ &= \frac{(k-z-1)(k-z)}{2} + \sum_{i \in S} |\tau_1(i) - (k+1)| \\ &\leq 2 \sum_{i \in S} |\tau_1(i) - (k+1)| \qquad \text{by Equation (22).} \end{aligned} \tag{24}$$

Similarly, we also have

$$\sum_{i \in T} |\sigma_1(i) - \sigma_2(i)| \leq 2 \sum_{i \in T} |(k+1) - \tau_2(i)|. \tag{25}$$

Now, using Equations (20)–(25), we have $F_{\min}(\tau_1, \tau_2) \leq 2F^*(\tau_1, \tau_2)$.

(Part (3)) From the formula given in Lemma 3.1, we have

$$K^{(p')}(\tau_1, \tau_2) - K^{(p)}(\tau_1, \tau_2) = (k-z)(p'-p)(k-z-1). \tag{26}$$

The first inequality is immediate from Equation (26), since $k \geq z$.

We now prove the second inequality. If $K^{(p)}(\tau_1, \tau_2) = 0$, then $\tau_1 = \tau_2$, so also $K^{(p')}(\tau_1, \tau_2) = 0$, and the second inequality holds. Therefore, assume that $K^{(p)}(\tau_1, \tau_2) \neq 0$. Divide both sides of Equation (26) by $K^{(p)}(\tau_1, \tau_2)$, to obtain

$$\frac{K^{(p')}(\tau_1, \tau_2)}{K^{(p)}(\tau_1, \tau_2)} = 1 + \frac{(k-z)(p'-p)(k-z-1)}{K^{(p)}(\tau_1, \tau_2)}. \tag{27}$$

Since $\frac{1+p'}{1+p} = 1 + \frac{p'-p}{1+p}$, the second inequality would follow from Equation (27) if we show

$$K^{(p)}(\tau_1, \tau_2) \geq (k-z)(k-z-1)(1+p) \tag{28}$$

In the derivation of the formula for $K^{(p)}(\tau_1, \tau_2)$ in the proof of Lemma 3.1, we saw that the contribution from Case 3 is $(k-z)^2$ and the contribution from Case 4 is $p(k-z)(k-z-1)$. Hence, $K^{(p)}(\tau_1, \tau_2) \geq (k-z)^2 + p(k-z)(k-z-1) \geq (k-z)(k-z-1) + p(k-z)(k-z-1) = (k-z)(k-z-1)(1+p)$, as desired.

(Part (4)) From the formula given in Lemma 3.5, we have

$$F^{(\ell')}(\tau_1, \tau_2) - F^{(\ell)}(\tau_1, \tau_2) = 2(k-z)(\ell' - \ell). \tag{29}$$

The first inequality is immediate from Equation (29), since $k \geq z$.

We now prove the second inequality. If $F^{(\ell)}(\tau_1, \tau_2) = 0$, then $\tau_1 = \tau_2$, so also $F^{(\ell')}(\tau_1, \tau_2) = 0$, and the second inequality holds. Therefore, assume that $F^{(\ell)}(\tau_1, \tau_2) \neq 0$. Divide both sides of Equation (29) by $F^{(\ell)}(\tau_1, \tau_2)$, to obtain

$$\frac{F^{(\ell')}(\tau_1, \tau_2)}{F^{(\ell)}(\tau_1, \tau_2)} = 1 + \frac{2(k-z)(\ell' - \ell)}{F^{(\ell)}(\tau_1, \tau_2)}. \tag{30}$$

Since $\frac{\ell'-k}{\ell-k} = 1 + \frac{\ell'-\ell}{\ell-k}$, the second inequality would follow from Equation (30) if we show

$$F^{(\ell)}(\tau_1, \tau_2) \geq 2(k-z)(\ell - k). \tag{31}$$

To see Equation (31), observe that $|S| + |T| = 2(k-z)$ and each element in $S$ and $T$ contributes at least $\ell - k$ (which is positive since $k < \ell$) to $F^{(\ell)}(\tau_1, \tau_2)$. □

# 6 An algorithmic application

In the context of algorithm design, the notion of near metrics has the following useful application. Given $r$ ranked lists $\tau_1, \ldots, \tau_r$ (either full lists or top $k$ lists) of "candidates," the *rank aggregation* problem [DKNS01] with respect to a distance measure $d$ is to compute a list $\tau$ (again, either a full list on the union of the domains of the $\tau_j$'s or another top $k$ list) such that $\sum_{j=1}^{r} d(\tau_j, \tau)$ is minimized.

This problem arises in the context of information retrieval, where possible results to a search query may be ordered with respect to several criteria, and it is useful to obtain an ordering (often a top $k$ list) that is a good aggregation of the rank orders produced. It is argued in [DKNS01] that Kendall's tau and its variants are good measures to use, both in the context of full lists and top $k$ lists. Our experiments at IBM Almaden (see also Section 9.1) have confirmed that, in fact, producing an ordering with small Kendall's tau distance yields qualitatively excellent results. Unfortunately, computing an optimal aggregation of several full or top $k$ lists is NP-hard for each of the Kendall measures. In this context, our notion of an equivalence class of distance measures comes in handy.

**Proposition 6.1.** *Let $\mathcal{C}$ be an equivalence class of distance measures. If there is at least one distance measure $d$ in $\mathcal{C}$ so that the rank aggregation problem with respect to $d$ has a polynomial-time exact or constant-factor approximation algorithm, then for every $d'$ in $\mathcal{C}$, there is a polynomial-time constant-factor approximation algorithm for the rank aggregation problem with respect to $d'$.*

*Proof.* Given $\tau_1, \ldots, \tau_r$, let $\tau$ denote an aggregation with respect to $d$ that is within a factor $c \geq 1$ of a best possible aggregation $\pi$ with respect to $d$, that is, $\sum_j d(\tau_j, \tau) \leq c \sum_j d(\tau_j, \pi)$. Let $c_1, c_2$ denote positive constants such that for all $\sigma, \sigma'$ (top $k$ or full lists, as appropriate) $c_1 d(\sigma, \sigma') \leq d'(\sigma, \sigma') \leq c_2 d(\sigma, \sigma')$. Also, let $\pi'$ denote a best possible aggregation with respect to $d'$. Then we have

$$\sum_j d'(\tau_j, \tau) \leq \sum_j c_2 d(\tau_j, \tau) \leq c \sum_j c_2 d(\tau_j, \pi) \leq c c_2 \sum_j d(\tau_j, \pi') \leq \frac{c c_2}{c_1} \sum_j d'(\tau_j, \pi').$$

□

Via an application of minimum-cost perfect matching, the rank aggregation problem can be solved optimally in polynomial time for any of the $F^{(\ell)}$ metrics. Together with Theorem 5.1, this implies polynomial time constant-factor approximation algorithms for the rank aggregation problem with respect to the Kendall measures.

# 7 Other approaches

## 7.1 Spearman's rho

Spearman's rho is the $L_2$ distance between two permutations. Formally,

$$\rho(\sigma_1, \sigma_2) = \left( \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|^2 \right)^{1/2}$$

and it can be shown that $\rho(\cdot, \cdot)$ is a metric.[4] The maximum value of $\rho(\sigma_1, \sigma_2)$ is $(n(n+1)(2n+1)/3)^{\frac{1}{2}}$, which occurs when $\sigma_1$ is the reverse of $\sigma_2$. Spearman's rho is a popular metric between permutations. Analogous to the footrule case, we can define the notions of $\rho_{\min}$, $\rho_{\text{avg}}$, and $\rho^{(\ell)}$. They are not in the big equivalence class, for the following reason. Consider the case where $k = n$, that is, where we are considering full lists, which are permutations of all of the elements in a fixed universe. In this case, we need only consider $\rho$, since $\rho_{\min}$, $\rho_{\text{avg}}$, and $\rho^{(\ell)}$ all equal $\rho$. But the maximum value of $F^*$ is $\Theta(n^2)$ and that of $\rho$ is $\Theta(n^{\frac{3}{2}})$. Therefore, $\rho_{\min}$, $\rho_{\text{avg}}$, and $\rho^{(\ell)}$ cannot be in the same equivalence class as $F^*$. What if we consider normalized versions of our distance measures, as discussed after Theorem 5.1? We now show that the normalized versions of $\rho_{\min}$, $\rho_{\text{avg}}$, and $\rho^{(\ell)}$ are not in the normalized version of the big equivalence class. If $d$ is a distance measure, we will sometimes denote the normalized version of $d$ by $\dot{d}$.

**Proposition 7.1.** *The distance measures $\rho_{\min}$, $\rho_{\text{avg}}$ and $\rho^{(\ell)}$ do not belong to the big equivalence class, even if all distance measures are normalized.*

*Proof.* As before, we consider full lists. We will show that $\dot{F}^*$ and $\dot{\rho}$ do not bound each other by constant multiples. We will present a family of pairs of full lists, one for each $n$, such that $\dot{F}^*(\sigma_1, \sigma_2) = \Theta(1/n)$ and $\dot{\rho}(\sigma_1, \sigma_2) = \Theta(1/n^{\frac{3}{4}})$. For every $n$, let $r = \lceil \sqrt{n} \rceil$. Assume $n$ is large enough so that $n \geq 2r$. Define the permutation $\sigma_1$ so that the elements in order are $1, \ldots, n$, and define the permutation $\sigma_2$ so that the elements in order are $r+1, \ldots, 2r, 1, \ldots, r, 2r+1, \ldots, n$. The unnormalized versions of Spearman's footrule and Spearman's rho can be easily calculated to be $F^*(\sigma_1, \sigma_2) = 2r^2 = \Theta(n)$ and $\rho(\sigma_1, \sigma_2) = (2r)^{\frac{3}{2}} = \Theta(n^{\frac{3}{4}})$. As we noted, the maximum value of $F^*$ is $\Theta(n^2)$ and that of $\rho$ is $\Theta(n^{\frac{3}{2}})$. Therefore, $\dot{F}^*(\sigma_1, \sigma_2) = \Theta(1/n)$ and $\dot{\rho}(\sigma_1, \sigma_2) = \Theta(1/n^{\frac{3}{4}})$. Thus $\dot{F}^*$ and $\dot{\rho}$ cannot bound each other by constant multiples, so $\dot{\rho}_{\min}$, $\dot{\rho}_{\text{avg}}$ and $\dot{\rho}^{(\ell)}$ do not belong to the normalized version of the big equivalence class. □

[4] Spearman's rho is usually defined without the exponent of $\frac{1}{2}$, that is, without the square root. However, it turns out that if we drop the exponent of $\frac{1}{2}$, then the resulting distance measure is not a metric, and is not even a near metric.

### 7.2 The intersection metric

A natural approach to defining the distance between two top $k$ lists $\tau_1$ and $\tau_2$ is to capture the extent of overlap between $D_{\tau_1}$ and $D_{\tau_2}$. We now define a more robust version of this distance measure. For $1 \leq i \leq k$, let $\tau^{(i)}$ denote the restriction of a top $k$ list to the first $i$ items. Let

$$\delta_i^{(w)}(\tau_1, \tau_2) = |D_{\tau_1^{(i)}} \Delta D_{\tau_2^{(i)}}|/(2i).$$

Finally, let

$$\delta^{(w)}(\tau_1, \tau_2) = \frac{1}{k} \sum_{i=1}^{k} \delta_i^{(w)}(\tau_1, \tau_2).$$

(Here, $\Delta$ represents the symmetric difference. Thus, $X \Delta Y = (X \setminus Y) \cup (Y \setminus X)$.) It is straightforward to verify that $\delta^{(w)}$ lies between 0 and 1, with the maximal value of 1 occurring when $D_{\tau_1}$ and $D_{\tau_2}$ are disjoint. In fact, $\delta^{(w)}$, as defined above, is just one instantiation of a more general paradigm: any convex combination of the $\delta_i^{(w)}$'s yields a metric on top $k$ lists.

We now show that the distance measure $\delta^{(w)}$ is a metric.

**Proposition 7.2.** *$\delta^{(w)}(\cdot, \cdot)$ is a metric.*

*Proof.* It suffices to show that $\delta_i^{(w)}(\cdot, \cdot)$ is a metric for $1 \leq i \leq k$. To show this, we show that for any three sets $A, B, C$, we have $|A \Delta C| \leq |A \Delta B| + |B \Delta C|$. For $x \in A \Delta C$, assume without loss of generality that $x \in A$ and $x \notin C$. We have two cases: if $x \in B$, then $x \in B \Delta C$ and if $x \notin B$, then $x \in A \Delta B$. Either way, each $x \in A \Delta C$ contributes at least one to the right-hand side, thus establishing the inequality. □

Since $\delta^{(w)}$ is bounded (by 1), and $F^*$ is not bounded, it follows that $\delta^{(w)}$ is not in the big equivalence class. Of course, $\delta^{(w)}$ is normalized; we now show that $\delta^{(w)}$ is not in the normalized version of the big equivalence class.

**Proposition 7.3.** *$\delta^{(w)}$ does not belong to the equivalence class, even if all distance measures are normalized.*

*Proof.* Let $\tau_1$ be the top $k$ list where the top $k$ elements in order are $1, 2, \ldots, k$, and let $\tau_2$ be the top $k$ list where the top $k$ elements in order are $2, \ldots, k, 1$. The normalized footrule can be calculated to be $\dot{F}^*(\tau_1, \tau_2) = \Theta(1/k)$, whereas $\delta^{(w)}(\tau_1, \tau_2) = (1/k) \sum_{i=1}^{k} 1/i = \Theta((\ln k)/k)$. Therefore, $\delta^{(w)}$ and $\dot{F}^*$ cannot bound each other by constant multiples, and so $\delta^{(w)}$ does not belong to the normalized version of the big equivalence class. □

## 8 The interpolation criterion

In practical situations where one compares two top $k$ lists, it would be nice if the distance value has some natural real-life interpretation associated with it. There are three possible extreme relationships between two top $k$ lists: (a) they are identical; (b) they contain the same $k$ elements in the exact opposite order, or (c) they are disjoint. We feel that it is desirable that the value in case (b) be about halfway between the values in cases (a) and (c).

Let $d$ denote any one of our distance measures between top $k$ lists $\tau_1$ and $\tau_2$. Analogous to the normalization given in footnote 3 of Section 5, let us obtain a normalized version $\nu$ that maps the distance values into the interval $[-1, 1]$ so that

(a) $\nu(\tau_1, \tau_2) = 1$ iff $\tau_1 = \tau_2$;

(b) $\nu(\tau_1, \tau_2) = -1$ iff $D_{\tau_1}$ and $D_{\tau_2}$ are disjoint, that is, $Z = \emptyset$.

Clearly, this can be achieved via a linear map of the form $\nu(\tau_1, \tau_2) = a \cdot d(\tau_1, \tau_2) + b$. The question now is: how close to zero is $\nu(\tau_1, \tau_2)$ when $\tau_1$ and $\tau_2$ contain the same $k$ elements in the exact opposite order?

It turns out that the answer is asymptotic (in $k$) to $p/(1+p)$ for $K^{(p)}$. Therefore, it is asymptotic to 0 for $K_{\min} = K^{(0)}$. In fact, for $K_{\min}$, it is $\Theta(1/k)$. For $F_{\min}$, it is $\frac{1}{2}$, and for $F^{(\ell)}$, with $\ell = k + \frac{1}{2} + \alpha$, it is $\Theta(\frac{\alpha}{k+\alpha})$. In fact, for $F^{(k+\frac{1}{2})}$, where $\alpha = 0$, it is $\Theta(1/k^2)$. Thus, from this viewpoint, the preferable distance measures are $K_{\min}$ and $F^{(k+\beta)}$ for $\beta = o(k)$ (which includes $F^*$).

# 9 Experiments

## 9.1 Comparing Web search engines

As we mentioned earlier, one of the important applications of comparing top $k$ lists is to provide an objective way to compare the output of different search engines. We illustrate the use of our methods by comparing the outputs of seven popular Web search engines: AltaVista (`www.altavista.com`), Lycos (`www.lycos.com`), AllTheWeb (`www.alltheweb.com`), HotBot (`hotbot.lycos.com`), NorthernLight (`www.northernlight.com`), AOL Search (`search.aol.com`), and MSN Search (`search.msn.com`). Comparing the output in this manner will shed light both on the similarities between the underlying indices and the ranking functions used by search engines. We selected $K_{\min}$ as the measure of comparison between the search engines. This choice is arbitrary, and as we argued earlier, we could just as well have chosen any other measure from the big equivalence class.

We made use of 750 queries, that were actually made by real users to a metasearch engine developed at the IBM Almaden Research Center [DKNS01]. For each of these queries, and for each of the seven Web search engines we are considering, we obtained the top 50 list.[5] We then computed the normalized $K_{\min}$ distance between every pair of search engine outputs. Finally, we averaged the distances over the 750 queries. The results are tabulated in Table 1. The values are normalized to lie between 0 and 1, with smaller values representing closer matches. Note, of course, that the table is symmetric about the main diagonal.

| | AltaVista | Lycos | AllTheWeb | HotBot | NorthernLight | AOL Search | MSN Search |
|---|---|---|---|---|---|---|---|
| AltaVista | 0.000 | 0.877 | 0.879 | 0.938 | 0.934 | 0.864 | 0.864 |
| Lycos | 0.877 | 0.000 | 0.309 | 0.888 | 0.863 | 0.796 | 0.790 |
| AllTheWeb | 0.879 | 0.309 | 0.000 | 0.873 | 0.866 | 0.782 | 0.783 |
| HotBot | 0.938 | 0.888 | 0.873 | 0.000 | 0.921 | 0.516 | 0.569 |
| NorthernLight | 0.934 | 0.863 | 0.866 | 0.921 | 0.000 | 0.882 | 0.882 |
| AOL Search | 0.864 | 0.796 | 0.782 | 0.516 | 0.882 | 0.000 | 0.279 |
| MSN Search | 0.864 | 0.790 | 0.783 | 0.569 | 0.882 | 0.279 | 0.000 |

Table 1: $K_{\min}$ distances between search engines for $k = 50$.

Several interesting conclusions can be derived from this table. Some of the conclusions are substantiated by the alliances between various search engines (for a detailed account of the alliances, see `www.searchenginewatch.com/reports/alliances.html`).

(1) AOL Search and MSN Search yield very similar results! The reason for this (surprising) behavior is two-fold: both AOL Search and MSN Search index similar sets of pages and probably use fairly similar ranking functions. These conclusions are substantiated by the fact that AOL Search uses search data from OpenDirectory and Inktomi, and MSN Search uses LookSmart and Inktomi. HotBot uses DirectHit and Inktomi, and can be seen to be moderately similar to AOL Search and MSN Search.

[5] For some queries, we had to work with a slightly smaller value of $k$ than 50, since a search engine returned some duplicates.

(2) Lycos and AllTheWeb yield similar results. Again, the reason for this is because Lycos gets its main results from DirectHit and AllTheWeb.

(3) AltaVista and NorthernLight, since they use their own crawling, indexing and ranking algorithms, are far away from every other search engine. This is plausible for two reasons: either they crawl and index very different portions of the Web or their ranking functions are completely unrelated to the ranking functions of the other search engines.

(4) The fact that $K_{\min}$ is a near metric allows us to draw additional interesting inferences from the tables (together with observations (1) and (2) above). For example, working through the alliances and partnerships mentioned above, and exploiting the transitivity of "closeness" for a near metric, we obtain the following inference. The data services LookSmart and OpenDirectory are closer to each other than they are to DirectHit. Given that DirectHit uses results from its own database and from OpenDirectory, this suggests that the in-house databases in DirectHit and OpenDirectory are quite different. A similar conclusion is again supported by the fact that Lycos and HotBot are far apart, and their main results are powered by OpenDirectory and DirectHit respectively.

### 9.2 Evaluating a metasearch engine

Recall that a metasearch engine combines the ranking of different search engines to produce an aggregated ranking. There are several metasearch engines available on the Web (for a list of popular ones, see the site `searchenginewatch.com`). Metasearch engines are quite popular for their ability to mitigate the quirks of crawl, coverage and their resistance to spam. As we mentioned earlier, our methods can be used to evaluate the behavior of a metasearch engine. Such an analysis will provide evidence to whether the metasearch is highly biased towards any particular search engine or is reasonably "close" to all the search engines.

For our purposes, we use a metasearch engine that we developed. Our metasearch engine uses a Markov Chain approach to aggregate various rankings. The underlying theory behind this method can be found in [DKNS01]. We used a version of our metasearch engine that combines the outputs of the seven search engines described above. We measured the average $K_{\min}$ distance of our metasearch engine's output to the output of each of the search engines for the same set of $750$ queries. The results are tabulated in Table 2. From this table and Table 1, we note the following. There is a strong bias towards the AOL Search/MSN

| AltaVista | Lycos | AllTheWeb | HotBot | NorthernLight | AOL Search | MSN Search |
|---|---|---|---|---|---|---|
| 0.730 | 0.587 | 0.565 | 0.582 | 0.823 | 0.332 | 0.357 |

Table 2: $K_{\min}$ distance of our metasearch engine to its sources for $k = 50$.

Search cluster, somewhat less bias towards Lycos, AllTheWeb, and HotBot, and very little bias towards AltaVista and NorthernLight. This kind of information is extremely valuable for metasearch design (and is beyond the scope of this paper). For example, the numbers show that the output of our metasearch engine is a reasonable aggregation of its sources—it does not simply copy of its components, nor does it exclude any component entirely. Finally, the degree to which our metasearch engine aligns itself with a search engine depends on the various reinforcements among the outputs of the search engines.

### 9.3 Correlations among the distance measures

The following experiment is aimed at studying the "correlations" between the distance measures. We seek to understand how much information the distance measures reveal about each other. One of the goals of this experiment is to find empirical support for the following belief motivated by our work in this paper: the

distance measures within an equivalence class all behave similarly, whereas different equivalence classes aim to capture different aspects of the distance between two lists.

Let $I$ denote the top $k$ list where the top $k$ elements in order are $1, 2, \ldots, k$. For a distance measure $d(\cdot, \cdot)$ and a top $k$ list $\tau$ with elements from the universe $\{1, 2, \ldots, 2k\}$, let $\hat{d}(\tau) = d(\tau, I)$. If $\tau$ is a randomly chosen top $k$ list, then $\hat{d}(\tau)$ is a random variable.

Let $d_1$ and $d_2$ denote two distance measures. Consider the experiment where a random top $k$ list $\tau$ is picked. Informally, the main question we ask here is the following: if we know $\widehat{d_1}(\tau)$ (namely, the distance, according to $d_1$, of $\tau$ to the list $I$), to what extent can we predict the value of $\widehat{d_2}(\tau)$? To address this question, we use two basic notions from information theory.

Recall that the entropy of a random variable $X$ is

$$H(X) = -\sum_x \Pr[X = x] \log \Pr[X = x].$$

If we truncate the precision to two digits and use logarithms to the base 10 in the entropy definition, then for each $d$, the quantity $H(\hat{d}(\tau))$ is a real number between 0 and 2. In words, when $\tau$ is picked at random, then there is up to "2 digits worth of uncertainty in the value of $\hat{d}(\tau)$."

The conditional entropy of a random variable $X$ with respect to another random variable $Y$ is

$$H(X \mid Y) = \sum_y \Pr[Y = y] H(X \mid Y = y).$$

Informally, the conditional entropy measures the uncertainty in $X$, assuming that we know the value of $Y$. In our case, we ask the question: for a random $\tau$, if we know the value of $\widehat{d_1}(\tau)$, how much uncertainty is left in the value of $\widehat{d_2}(\tau)$?[6]

For all pairs of our distance measures $d_1$ and $d_2$, we measure $H(\widehat{d_2}(\tau) \mid \widehat{d_1}(\tau))$, and present the results in Table 3. We consider a universe of 20 elements and let $k = 10$. (These choices enable us to exhaustively enumerate all possible top $k$ lists and perform our experiments on them.) The entry $(d_1, d_2)$ in this table denotes $H(\widehat{d_2}(\tau) \mid \widehat{d_1}(\tau))$. Therefore, the closer the value is to 2, the less information $\widehat{d_1}$ reveals about $\widehat{d_2}$. The value of 1 is an interesting case, since this roughly corresponds to saying that on the average, given $\widehat{d_1}(\tau)$, one can predict the leading digit of $\widehat{d_2}(\tau)$.

Some conclusions that can be drawn from the table are the following:

(1) Every distance measure reveals a lot of information about symmetric difference $\delta$. A reason for this is that $\delta$ uses only 10 distinct values between 0 and 1, and is not sharp enough to yield finer information. This suggests that the other measures are preferable to symmetric difference.

(2) The distance measure $\rho^{(k+1)}$ reveals much information about the other measures, as is evident from the row for $\rho^{(k+1)}$; on the other hand, as can be seen from the column for $\rho^{(k+1)}$, the other measures do not reveal much information about $\rho^{(k+1)}$. The weighted symmetric difference metric $\delta^{(w)}$ seems fairly unrelated to all the others.

(3) The measures in the big equivalence class all appear to have a stronger correlation between themselves than to the ones not in the class. In fact, each of the footrule measures $F_{\min}$, $F^*$ is strongly correlated with the other footrule measures as is evident from the entries corresponding to their submatrix. Similarly, the Kendall measures $K_{\min}, K_{\text{avg}}, K^{(1)}$ are all strongly correlated. This suggests that the footrule and Kendall measures form two 'mini'-equivalence classes that sit inside the big equivalence class.

**Acknowledgments.** We thank Moni Naor and Gagan Aggarwal for helpful suggestions.

[6] We chose conditional entropy instead of statistical notions like correlation for the following reason. Correlation (covariance divided by the product of standard deviations) measures linear relationships between random variables. For example, if $X = \alpha Y + \beta$ for some constants $\alpha$ and $\beta$, then the correlation between $X$ and $Y$ is zero. On the other hand, consider $X = \alpha Y^2 + \beta Y + \gamma$; even though given the value of $Y$, there is absolutely no uncertainty in the value of $X$, their correlation is not zero. Conditional entropy, however, can measure arbitrary functional relationships between random variables. If $X = f(Y)$ for any fixed function $f$, then $H(X \mid Y) = 0$.

| | $\delta$ | $\delta^{(w)}$ | $\rho^{(k+1)}$ | $F^*$ | $F_{\min}$ | $K_{\min}$ | $K_{\text{avg}}$ | $K^{(1)}$ |
|---|---|---|---|---|---|---|---|---|
| $\delta$ | 0.000 | 1.409 | 1.469 | 1.203 | 1.029 | 1.235 | 1.131 | 0.991 |
| $\delta^{(w)}$ | 0.580 | 0.000 | 1.193 | 0.863 | 0.945 | 1.087 | 1.091 | 1.043 |
| $\rho^{(k+1)}$ | 0.530 | 1.083 | 0.000 | 0.756 | 0.838 | 0.670 | 0.773 | 0.760 |
| $F^*$ | 0.497 | 0.985 | 0.989 | 0.000 | 0.434 | 0.848 | 0.845 | 0.819 |
| $F_{\min}$ | 0.388 | 1.132 | 1.131 | 0.499 | 0.000 | 0.885 | 0.748 | 0.650 |
| $K_{\min}$ | 0.490 | 1.170 | 0.863 | 0.808 | 0.780 | 0.000 | 0.454 | 0.500 |
| $K_{\text{avg}}$ | 0.421 | 1.210 | 1.002 | 0.841 | 0.680 | 0.490 | 0.000 | 0.354 |
| $K^{(1)}$ | 0.361 | 1.240 | 1.068 | 0.894 | 0.660 | 0.615 | 0.433 | 0.000 |

Table 3: Conditional entropy values for pairs of distance measures. The entry $(d_1, d_2)$ of the table may be interpreted as the average uncertainty in $\widehat{d}_2(\tau)$, assuming we know $\widehat{d}_1(\tau)$.

## References


[AB95] T. Andreae and H. S. Bandelt. Performance guarantees for approximation algorithms depending on parametrized triangle inequalities. *SIAM Journal of Discrete Mathematics*, 8(1):1–16, 1995.

[BC00] M. A. Bender and C. Chekuri. Performance guarantees for the TSP with a parameterized triangle inequality. *Information Processing Letters*, 73(1-2):17–21, 2000.

[CCF+01] D. Carmel, D. Cohen, R. Fagin, E. Farchi, M. Herscovici, Y. Maarek, and A. Soffer. Static index pruning for information retrieval systems. In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–50, 2001.

[CCFC02] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming*, 2002.

[Cri80] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*. Number 34 in Lecture Notes in Statistics. Springer-Verlag, Berlin, 1980.

[DG77] P. Diaconis and R. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B*, 39(2):262–268, 1977.

[Dia88] P. Diaconis. *Group Representation in Probability and Statistics*. Number 11 in IMS Lecture Series. Institute of Mathematical Statistics, 1988.

[DKNS01] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International World Wide Web Conference*, pages 613–622, 2001.

[Dug66] J. Dugundji. *Topology*. Allyn and Bacon, Inc., Boston, 1966.

[FS98] R. Fagin and L. Stockmeyer. Relaxing the triangle inequality in pattern matching. *International Journal of Computer Vision*, 30(3):219–231, 1998.

[KG90] M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Edward Arnold, London, 1990.

[Lee95] J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–188, 1995.

[Lee97] J. H. Lee. Combining multiple evidence from different relevant feedback methods. In *Database Systems for Advanced Applications*, pages 421–430, 1997.