SIAM J. DISCRETE MATH.
Vol. 17, No. 1, pp. 134–160

COMPARING TOP k LISTS*

RONALD FAGIN†, RAVI KUMAR†, AND D. SIVAKUMAR†

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 the same. 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* with respect to a large class of distance measures.

Key words. triangle inequality, polygonal inequality, metric, near metric, distance measures, top k list, rank aggregation

AMS subject classifications. 68R05, 68W25, 54E99

DOI. S0895480102412856

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 IR 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. Methods based on precision and recall yield a way to compare two top k lists by comparing them both to "ground truth." However, there are two limitations of such approaches: First, these methods typically give absolute (unary) ratings of top k lists, rather than give a relative, binary measure of distance. Second, for IR 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

*Received by the editors July 23, 2002; accepted for publication (in revised form) September 8, 2002; published electronically October 14, 2003. A preliminary version of this paper appeared in *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, New York, SIAM, Philadelphia, 2003, pp. 28–36.

http://www.siam.org/journals/sidma/17-1/41285.html

†IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120 (fagin@almaden.ibm.com, ravi@almaden.ibm.com, siva@almaden.ibm.com).

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 (i.e., how similar is the top k composite output to the top k of each of its components), 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 "metasearch" 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 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.[1] 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. Kamvar et al. [KHMG03] have used one of our distance measures in evaluating the quality of an approximate version of the PageRank ranking algorithm. 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 [CCF02] 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 the 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

[1] In fact, our first author is a coauthor of [CCF+01] and the need for comparing top k lists that arose in that paper is what led us to the research in this paper.

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, we 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. The first says that it satisfies a relaxed version of the "polygonal inequality" (the natural extension of the standard triangle inequality). The second says that there exists 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 Search and MSN Search 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 [Dia88, KG90]). 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 fairly straightforward attempt at defining a distance measure 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 IR [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 a metric 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 $n > 1$ 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. In section 7, we discuss two approaches based on Spearman's rho and symmetric difference. In section 8, we discuss the interpolation criterion—a natural and desirable property of a distance measure. In section 10, we conclude the paper.

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* σ 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 σ, we interpret $\sigma(i)$ as the position (or rank) of element i. We say that i *is ahead of* j *in* σ 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 σ_1, σ_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 σ_1 and σ_2, then let $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 0$; if i and j are in the opposite order (such as i being ahead of j in σ_1 and j being ahead of i in σ_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 σ_1 is the reverse of σ_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 $n^2/2$ when n is even, and $(n + 1)(n - 1)/2$ when n is odd. As with Kendall's tau, the maximum occurs when σ_1 is the reverse of σ_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 have only the top k members of the ordering. Formally, a *top* k *list* τ is a bijection from a domain D_τ (intuitively, the members of the top k list) to $[k]$. We say that i *appears in* the top k list τ if $i \in D_\tau$. Similar to our convention for permutations, we interpret $\tau(i)$ (for i in D_τ) as the rank of i in τ. As before, we say that i *is ahead of* j *in* τ if $\tau(i) < \tau(j)$. If τ is a top k list and σ is a permutation on $D \supseteq D_\tau$, then we say that σ is an *extension* of τ, which we denote $\sigma \succeq \tau$, if $\sigma(i) = \tau(i)$ for all $i \in D_\tau$.

Assume that τ_1 and τ_2 are top k lists. In this section, we give several measures

for the distance between τ_1 and τ_2. We begin by recalling the definition of a metric and formally defining 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 considered in this paper are distance measures.

Global notation. Here we set up some global notation that we use throughout the paper. When two top k lists τ_1 and τ_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 $|D| = 2k - z$.

Remark. An important feature of our work is that when we compare τ_1 and τ_2, we do not assume that these are top k lists of elements from a fixed domain D. This is a fairly natural requirement in many applications of our work. For example, if we wish to compare the top 10 lists produced by two search engines, it is unreasonable to expect any knowledge of the (possibly very large) universe to which elements of these lists belong; in fact, we cannot even expect to know the size of this universe. The drawback of our requirement is that it is one of the reasons why several very natural distance measures that we define between top k lists fail to be metrics (cf. section 3.3).

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 τ_1 and τ_2, we define $\mathcal{P}(\tau_1, \tau_2) = \mathcal{P}_{D_{\tau_1} \cup D_{\tau_2}}$ to be the set of all unordered pairs of distinct elements in $D_{\tau_1} \cup D_{\tau_2}$.

For top k lists τ_1 and τ_2, the *minimizing Kendall distance* $K_{\min}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the minimum value of $K(\sigma_1, \sigma_2)$, where σ_1 and σ_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 τ_1 and τ_2, the *averaging Kendall distance* $K_{\text{avg}}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the expected value $\mathrm{E}(K(\sigma_1, \sigma_2))$, where σ_1 and σ_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 $0 \leq p \leq 1$. Similar to our definition of $\bar{K}_{i,j}(\sigma_1, \sigma_2)$ for permutations σ_1, σ_2, we define a penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ for top k lists τ_1, τ_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 $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 0$; 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 τ_1 and j being ahead of i in τ_2), then let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$.

Case 2 (i and j both appear in one top k list (say τ_1), and exactly one of i or j, say i, appears in the other top k list (τ_2)). If i is ahead of j in τ_1, then let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 0$, and otherwise let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$. Intuitively, we know that i is ahead of j as far as τ_2 is concerned, since i appears in τ_2 but j does not.

Case 3 (i, but not j, appears in one top k list (say τ_1), and j, but not i, appears in the other top k list (τ_2)). Then let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$. Intuitively, we

know that i is ahead of j as far as τ_1 is concerned and j is ahead of i as far as τ_2 is concerned.

Case 4 (i and j both appear in one top k list (say τ_1), but neither i nor j appears in the other top k list (τ_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 penalty score of $1/2$. Later, we show that the optimistic approach gives precisely $K_{\min}$ and the neutral approach gives precisely K_{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

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

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_ℓ and its rank $\tau_1(j_\ell)$ in the first top k list τ_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_ℓ is ahead of i in τ_1; denote this net contribution of ℓ 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_ℓ is ahead of in τ_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 ℓ, 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

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

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

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

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 σ_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 τ_1 and τ_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 σ_1 and σ_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, as remarked earlier, in this paper we do not assume a fixed domain. Therefore, we define K_{Haus}, the Hausdorff version of the Kendall distance between top k lists, to be given by (5) with $d(\sigma_1,\sigma_2)$ as the Kendall distance $K(\sigma_1,\sigma_2)$, but where, unlike Critchlow, we take σ_1 and σ_2 to be permutations of $D_{\tau_1}\cup D_{\tau_2}$.

Critchlow obtains a closed form for his version of (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 (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_{Haus}.

LEMMA 3.2.

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

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_{avg}, respectively. Furthermore, we show the somewhat surprising result that the Hausdorff distance K_{Haus} also equals $K^{(1/2)}$.

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

Proof. Let τ_1 and τ_2 be top k lists. We must show that $K_{\min}(\tau_1,\tau_2) = K^{(0)}(\tau_1,\tau_2)$. Define σ_1 to be the extension of τ_1 over D where the elements are, in order, the elements of D_{τ_1} in the same order as they are in τ_1, followed by the elements of T in the same order as they are in τ_2. For example, if $k = 4$, if the top 4 elements of τ_1 are, in order, 1, 2, 3, 4, and if the top 4 elements of τ_2 are, in order, 5, 4, 2, 6, then the ordering of the elements for σ_1 is 1, 2, 3, 4, 5, 6. We similarly define the extension σ_2 of τ_2 by reversing the roles of τ_1 and τ_2. First, we show that $K_{\min}(\tau_1,\tau_2) = K(\sigma_1,\sigma_2)$, and then we show that $K(\sigma_1,\sigma_2) = K^{(0)}(\tau_1,\tau_2)$.

To show that $K_{\min}(\tau_1,\tau_2) = K(\sigma_1,\sigma_2)$, it is clearly sufficient to show that if σ_1' is an arbitrary extension of τ_1 (over D) and σ_2' is an arbitrary extension of τ_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 τ_1,τ_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. $K_{\text{avg}} = K^{(1/2)} = K_{\text{Haus}}$.

Proof. Let τ_1,τ_2 be top k lists. Then

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

We shall show that

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

This proves that $K_{\text{avg}} = K^{(1/2)}$, since the result of substituting $\bar{K}^{(1/2)}_{i,j}(\tau_1,\tau_2)$ for $\mathrm{E}(\bar{K}_{i,j}(\sigma_1,\sigma_2))$ in (8) gives $K^{(1/2)}(\tau_1,\tau_2)$. Similar 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 $K_{\text{avg}} = K^{(1/2)}$ if we show that when $\{i, j\}$ is a special pair, then $\text{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2)) = 1/2$. Assume without loss of generality that i, j are both in D_{τ_1} but neither is in D_{τ_2}. The ordering of i, j in σ_1 is forced by τ_1. Further, there is a one-to-one correspondence between those permutations σ_2 that extend τ_2 with i ahead of j and those that extend τ_2 with j ahead of i (the correspondence is determined by simply switching i and j). Therefore, for each choice of σ_1, exactly half of the choices for σ_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 $\text{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2)) = 1/2$, as desired.

We now show that $K_{\text{Haus}} = K^{(1/2)}$. 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, $K_{\text{Haus}} = K^{(1/2)}$. 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 (6).

Let τ_1, τ_2 be top k lists. Then $K_{\text{Haus}}(\tau_1, \tau_2)$ 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 σ_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. Let σ_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 not hard to see that $K_{\text{Haus}}(\tau_1, \tau_2) = K(\sigma_1^*, \sigma_2^*)$. 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 τ_1 or τ_2, but neither appears in the other top k list. If both i and j appear in τ_1 but neither appears in τ_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 τ_2 but neither appears in τ_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 τ_1 but neither appears in τ_2) as there are of the second type (where both i and j appear in τ_2 but neither appears in τ_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 τ_1 and τ_2, the *minimizing footrule distance* $F_{\min}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the minimum value of $F(\sigma_1, \sigma_2)$, where σ_1 and σ_2 are each permutations of D and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$.

For top k lists τ_1 and τ_2, the *averaging footrule distance* $F_{\text{avg}}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the expected value $\text{E}(F(\sigma_1, \sigma_2))$, where σ_1 and σ_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 ℓ be a real number greater than k. The *footrule distance with location parameter* ℓ, denoted $F^{(\ell)}$, is obtained—intuitively—by placing all missing elements in each of the lists at position ℓ and computing the usual footrule distance between them. More formally, given top k lists τ_1 and τ_2, define functions τ_1' and τ_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 τ_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 ℓ 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.

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

Similar to our definition of K_{Haus}, we define F_{Haus}, the Hausdorff version of the footrule distance between top k lists, to be given by (5) with $d(\sigma_1, \sigma_2)$ as the footrule distance $F(\sigma_1, \sigma_2)$, where, as before, we take σ_1 and σ_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 (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. 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_{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 τ_1 and τ_2. Since $F_{\min} = F_{\text{avg}} = F_{\text{Haus}}$ (Proposition 3.7), this gives us a formula for $F_{\min}$ and F_{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_{avg}.

Unlike the situation with $K_{\min}$ and K_{avg}, the next proposition tells us that $F_{\min}$ and F_{avg} are the same. Furthermore, the Hausdorff distance F_{Haus} shares this common value.

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

Proof. Let τ_1 and τ_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 σ_1 and σ_1' extend τ_1 and where σ_2 and σ_2' extend τ_2. We need only show that $F(\sigma_1, \sigma_2) = F(\sigma_1', \sigma_2')$, that is, that the value of $F(\sigma_1, \sigma_2)$ is independent of the choice of σ_1, σ_2. Therefore, we need only show that $F(\sigma_1, \sigma_2) = F(\sigma_1, \sigma_2')$,

where σ_1 is held fixed, since by symmetry (where σ_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 $\sigma_1(i) \le k < \sigma_2(i)$. 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 $F_{\min} = F_{\text{avg}} = F_{\text{Haus}}$. □

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_{avg} and F_{Haus}; and (3) $F^{(\ell)}$. Perhaps the most natural question, and the main subject of our investigation, is to ask 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 $K_{\min}, K_{\text{avg}}, K_{\text{Haus}}, F_{\text{avg}}, F_{\text{Haus}}$) is a metric.

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

Proof. We need only show that the triangle inequality holds. Let τ_1, τ_2, τ_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 τ_1 by letting $v_1(i) = \tau_1(i)$ for $i \in D_{\tau_1}$ and ℓ otherwise. Similarly, define an n-dimensional vector v_2 corresponding to τ_2 and an n-dimensional vector v_3 corresponding to τ_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 τ_1 be the top 2 list where the top 2 items in order are 1,2; let τ_2 be the top 2 list where the top 2 items in order are 1,3; let τ_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_{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_{avg} and F_{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_{avg}, and K_{Haus} are special cases of $K^{(p)}$. (As we noted earlier, the versions of F_{Haus} and K_{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 seemingly different notions of being a near metric, which our distance measures 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 $n > 1$ 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$_{\text{t}}$ metric *is a distance measure that satisfies the c-triangle inequality. A* c-relaxed$_{\text{p}}$ metric *is a distance measure that satisfies the c-polygonal inequality.*

Of course, every c-relaxed$_{\text{p}}$ metric is a c-relaxed$_{\text{t}}$ metric. Theorem 4.7 below says that there is a c-relaxed$_{\text{t}}$ metric that is not a c'-relaxed$_{\text{p}}$ metric for any constant c'. We shall focus here on the stronger notion of being a c-relaxed$_{\text{p}}$ 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 $c_1 d'(x, y) \le d(x, y) \le c_2 d'(x, y)$.*

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$_{\mathrm{p}}$ metric if and only if d is c-metric-bounded.*

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

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

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 $d'(x, z) \ge (1/c) d(x, z)$. Hence, $d'(x, z) \ne 0$ if $x \ne z$. Symmetry of d' follows immediately from symmetry of d. Finally, d' satisfies the triangle inequality, since

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

Therefore, d' is a metric.

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

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

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

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).*[2]

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 $c_1 d'(\tau_1, \tau_2) \le d(\tau_1, \tau_2) \le c_2 d'(\tau_1, \tau_2)$ for every pair τ_1, τ_2 of top k lists.*[3]

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 σ_1, σ_2, we have

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

(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 topological 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 $d(x,y) = 0$ if and only if $x = y$. To show the 2-triangle inequality, let $\alpha = d(x,z)$, $\beta = d(x,y)$, and $\gamma = d(y,z)$. Now $\sqrt{\alpha} \le \sqrt{\beta} + \sqrt{\gamma}$, since the function d' with $d'(x,y) = |x-y|$ is a metric. By squaring both sides, we get $\alpha \le \beta + \gamma + 2\sqrt{\beta\gamma}$. But $\sqrt{\beta\gamma} \le (\beta+\gamma)/2$ by the well-known fact that the geometric mean is bounded above by the arithmetic mean. We therefore obtain $\alpha \le 2(\beta+\gamma)$, that is, $d(x,z) \le 2(d(x,y)+d(y,z))$. 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 $1 \le i \le n-1$. 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 ("ϵ-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

[2] 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.

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

into an at most countable collection of neighborhood-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 neighborhood-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 $d(x,y)=0$ if and only if $x=y$). 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 ϵ-ball $\{y \mid d(x,y)<\epsilon\}$ is just a minor distortion of an ϵ-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 the second main result 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 σ_1,σ_2.

THEOREM 5.1 (main result 2). *The distance measures $K_{\min}$, K_{avg}, K_{Haus}, $K^{(p)}$ (for every choice of p), $F_{\min}$, F_{avg}, F_{Haus}, and $F^{(\ell)}$ (for every choice of ℓ) 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 $K_{\min}, K_{\mathrm{avg}}, K_{\mathrm{Haus}}, F_{\mathrm{avg}}, F_{\mathrm{Haus}}$) 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, some members of the big equivalence class 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 $[0,1]$.[4] 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; while we have improved some of them with more complicated proofs, our goal here is 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, ℓ.

THEOREM 5.3. *Let τ_1, τ_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. (1) For the first inequality of part (1), let σ_1, σ_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 σ_1 of τ_1 and σ_2 of τ_2.

For the second inequality of part (1), let σ_1, σ_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 σ_1 of τ_1 and σ_2 of τ_2.

(2) Let σ_1, σ_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 τ_1, τ_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

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

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

[4] 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. However, this normalization immediately precludes one from studying metric-like properties.

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

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

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

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 σ_2 is a permutation, the list of values $\sigma_2(i), i \in S$, is precisely $k+1,\ldots,2k-z$. Summing over all $i \in S$ yields

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

Similarly, we also have

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

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

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

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

The first inequality is immediate from (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 (26) by $K^{(p)}(\tau_1,\tau_2)$ to obtain

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

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

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

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.

(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 (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 (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 (30) if we show

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

To see (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 τ (again, either a full list or another top k list) such that $\sum_{j=1}^{r} d(\tau_j,\tau)$ is minimized.

This problem arises in the context of IR, 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 the IBM Almaden Research Center (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 τ denote an aggregation with respect to d that is within a factor $c \geq 1$ of a best possible aggregation π 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 σ,σ' (top k or full lists, as appropriate) $c_1 d(\sigma,\sigma') \leq d'(\sigma,\sigma') \leq c_2 d(\sigma,\sigma')$. Also, let π' denote a best possible aggregation with respect to d'. Then we have

$$\begin{aligned}\sum_j d'(\tau_j,\tau) &\leq \sum_j c_2 d(\tau_j,\tau) \leq c\sum_j c_2 d(\tau_j,\pi) \\ &\leq cc_2 \sum_j d(\tau_j,\pi') \leq \frac{cc_2}{c_1}\sum_j d'(\tau_j,\pi').\end{aligned}$$

□

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.[5] The maximum value of $\rho(\sigma_1, \sigma_2)$ is $(n(n+1)(2n+1)/3)^{\frac{1}{2}}$, which occurs when σ_1 is the reverse of σ_2. Spearman's rho is a popular metric between permutations. Analogous to the footrule case, we can define the notions of $\rho_{\min}$, ρ_{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 ρ, since $\rho_{\min}$, ρ_{avg}, and $\rho^{(\ell)}$ all equal ρ. But the maximum value of F^* is $\Theta(n^2)$ and that of ρ is $\Theta(n^{\frac{3}{2}})$. Therefore, $\rho_{\min}$, ρ_{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 $\rho_{\min}$, ρ_{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 $\rho_{\min}$, ρ_{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 σ_1 so that the elements in order are $1, \ldots, n$, and define the permutation σ_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 ρ 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. □

7.2. The intersection metric. A natural approach to defining the distance between two top k lists τ_1 and τ_2 is to capture the extent of overlap between D_{τ_1} and D_{τ_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

[5] Spearman's rho is usually defined without the exponent of $\frac{1}{2}$, that is, without the square root. However, 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.

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

(Here, Δ 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_{τ_1} and D_{τ_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 with strictly positive coefficients 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 τ_1 be the top k list where the top k elements in order are $1, 2, \ldots, k$, and let τ_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. □

7.3. Goodman and Kruskal's gamma. Goodman and Kruskal [GK54] have defined a "correlation statistic" for rank orders (and partial orders), which can be used to define a distance measure for top k lists. Let τ_1 and τ_2 be top k lists. As before, let $\mathcal{P}(\tau_1, \tau_2) = \mathcal{P}_{D_{\tau_1} \cup D_{\tau_2}}$ be the set of all unordered pairs of distinct elements in $D_{\tau_1} \cup D_{\tau_2}$. Let C be the set of all pairs $\{i, j\} \in \mathcal{P}$ (τ_1, τ_2) where both τ_1 and τ_2 implicitly or explicitly place one of i or j above the other (τ_1 and τ_2 can differ on this placement). In other words, C consists of all pairs $\{i, j\} \in \mathcal{P}$ (τ_1, τ_2) such that (1) either i or j is in D_{τ_1} and (2) either i or j is in D_{τ_2}. Note that C consists exactly of all pairs $\{i, j\}$ that occur in the first three cases in our definition of $K^{(p)}$. Now define $\gamma(\tau_1, \tau_2)$ to be the fraction of pairs $\{i, j\} \in C$ where τ_1 and τ_2 disagree on whether i is ahead of j.

Goodman and Kruskal defined this quantity for rank orders τ_1 and τ_2 that are more general than top k lists, namely, "bucket orders," or total orders with ties.[6] However, this quantity is not well defined for all pairs of bucket orders, since the set C as defined above can be empty in general. In ongoing work, we are exploring the issue of bucket orders in more detail. Here we simply remark that if τ_1 and τ_2 are top k lists, then C is always nonempty, and so we do obtain a meaningful distance measure on top k lists via this approach.

[6] As with Kendall's tau and Spearman's footrule (see footnote 4), Goodman and Kruskal's gamma is traditionally normalized to lie in the interval $[-1, 1]$, although we shall not do so, so that we can discuss metric properties.

We now show that γ is not a metric. Let τ_1 be the top 4 list where the top 4 items in order are 1,2,3,4; let τ_2 be the top 4 list where the top 4 items in order are 1,2,5,6; and let τ_3 be the top 4 list where the top 4 items in order are 5,6,7,8. It is straightforward to verify that $\gamma(\tau_1, \tau_3) = 1$, $\gamma(\tau_1, \tau_2) = 4/13$, and $\gamma(\tau_2, \tau_3) = 8/13$. So the triangle inequality fails, because $\gamma(\tau_1, \tau_3) > \gamma(\tau_1, \tau_2) + \gamma(\tau_2, \tau_3)$.

We now show that γ belongs to the normalized version of our big equivalence class and is therefore a near metric. Let τ_1 and τ_2 be top k lists, and let C be as earlier. Let $c = |C|$, and let s be the number of pairs $\{i, j\} \in C$ where τ_1 and τ_2 disagree on whether i is ahead of j. Thus, $\gamma(\tau_1, \tau_2) = s/c$. Note that since $c \le k^2$, we have $s/c \ge s/k^2 = K_{\min}(\tau_1, \tau_2)/k^2$, which equals the normalized $K_{\min}$ distance between τ_1 and τ_2. Finally, note that since $c \ge \binom{k}{2}$, we have $s/c \le s/\binom{k}{2} \le 4s/k^2$ (for $k \ge 2$). Therefore, s/c is at most 4 times the normalized $K_{\min}$ distance between τ_1 and τ_2 if $k \ge 2$. (It is easy to see that γ and the normalized version of $K_{\min}$ are both 0 or both 1 when $k = 1$.)

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 τ_1 and τ_2. Analogous to the normalization given in footnote 4 of section 5, let us obtain a normalized version ν that maps the distance values into the interval $[-1, 1]$ so that

(a) $\nu(\tau_1, \tau_2) = 1$ if and only if $\tau_1 = \tau_2$;

(b) $\nu(\tau_1, \tau_2) = -1$ if and only if D_{τ_1} and D_{τ_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 τ_1 and τ_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 (www.hotbot.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

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

	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

the top 50 list.[7] 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.

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 twofold: 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.

(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 coverage, resistance to spam, and ability to mitigate the quirks of crawl. As we mentioned earlier, our methods can be used to evaluate the behavior of a metasearch engine. Such

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

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

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

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 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 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 τ with elements from the universe $\{1, 2, \ldots, 2k\}$, let $\hat{d}(\tau) = d(\tau, I)$. If τ 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 τ 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 τ 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 τ is picked at random, then there is up to "2 digits worth of uncertainty in the value of $\hat{d}(\tau)$."

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 $\widehat{d_1}(\tau)$.

	δ	$\delta^{(w)}$	$\rho^{(k+1)}$	γ	F^*	$F_{\min}$	$K_{\min}$	K_{avg}	$K^{(1)}$
δ	0.000	1.409	1.469	1.415	1.203	1.029	1.235	1.131	0.991
$\delta^{(w)}$	0.580	0.000	1.193	1.282	0.863	0.945	1.087	1.091	1.043
$\rho^{(k+1)}$	0.530	1.083	0.000	1.057	0.756	0.834	0.670	0.773	0.760
γ	0.503	1.197	1.082	0.000	1.039	1.025	0.533	0.525	0.507
F^*	0.497	0.985	0.989	1.246	0.000	0.434	0.848	0.845	0.819
$F_{\min}$	0.388	1.132	1.131	1.297	0.499	0.000	0.885	0.748	0.650
$K_{\min}$	0.490	1.170	0.863	0.700	0.808	0.780	0.000	0.454	0.500
K_{avg}	0.421	1.210	1.002	0.729	0.841	0.680	0.490	0.000	0.354
$K^{(1)}$	0.361	1.240	1.068	0.789	0.894	0.660	0.615	0.433	0.000

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 τ, if we know the value of $\widehat{d_1}(\tau)$, how much uncertainty is left in the value of $\widehat{d_2}(\tau)$?[8]

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 δ. A reason for this is that δ 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 to 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

[8] 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 α and β, 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 $H(X \mid Y) = 0$.

Kendall measures form two "mini"-equivalence classes that sit inside the big equivalence class.

(4) The distance measure γ reveals much information about the Kendall measures, and vice versa. This is to be expected, since γ is very similar to $K_{\min}$, except for the normalization factor.

10. Conclusions. We have introduced various distance measures between top k lists and have shown that these distance measures are equivalent in a very natural sense. We have also introduced a robust notion of "near metric," which we think is interesting in its own right. We have shown that each of our distance measures that is not a metric is a near metric. Our results have implications for IR (since we can quantify the differences between search engines, by measuring the difference between their outputs). Our results also have implications for algorithm design (since we can use our machinery to obtain polynomial-time constant-factor approximation algorithms for the rank aggregation problem).

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

REFERENCES

[AB95] T. Andreae and H.-J. Bandelt, *Performance guarantees for approximation algorithms depending on parametrized triangle inequalities*, SIAM J. Discrete Math., 8 (1995), pp. 1–16.

[BC00] M. A. Bender and C. Chekuri, *Performance guarantees for the TSP with a parameterized triangle inequality*, Inform. Process. Lett., 73 (2000), pp. 17–21.

[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, 2001, pp. 43–50.

[CCF02] 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, Lecture Notes in Comput. Sci. 2380, Springer-Verlag, Berlin, 2002, pp. 693–703.

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

[DG77] P. Diaconis and R. Graham, *Spearman's footrule as a measure of disarray*, J. Roy. Statist. Soc., Ser. B, 39 (1977), pp. 262–268.

[Dia88] P. Diaconis, *Group Representation in Probability and Statistics*, IMS Lecture Notes Monogr. Ser. 11, Institute of Mathematical Statistics, Hayward, CA, 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, ACM, New York, 2001, pp. 613–622.

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

[FKS03] R. Fagin, R. Kumar, and D. Sivakumar, *Comparing top k lists*, in Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2003, pp. 28–36.

[FS98] R. Fagin and L. Stockmeyer, *Relaxing the triangle inequality in pattern matching*, Int. J. Comput. Vision, 30 (1998), pp. 219–231.

[GK54] L. A. Goodman and W. H. Kruskal, *Measures of association for cross classifications*, J. Amer. Statist. Assoc., 49 (1954), pp. 732–764.

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

[KHMG03] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub, *Extrapolation methods for accelerating PageRank computations*, in Proceedings of the 12th International World Wide Web Conference, ACM, New York, 2003, pp. 261–270.

[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, 1995, pp. 180–188.

[Lee97] J. H. Lee, *Combining multiple evidence from different relevant feedback methods*, in Database Systems for Advanced Applications '97, World Scientific, Singapore, 1997, pp. 421–430.