Visualizing Incomplete and Partially Ranked Data

Paul Kidwell, Guy Lebanon, *Member, IEEE*, and William S. Cleveland

Abstract—Ranking data, which result from m raters ranking n items, are difficult to visualize due to their discrete algebraic structure, and the computational difficulties associated with them when n is large. This problem becomes worse when raters provide tied rankings or not all items are ranked. We develop an approach for the visualization of ranking data for large n which is intuitive, easy to use, and computationally efficient. The approach overcomes the structural and computational difficulties by utilizing a natural measure of dissimilarity for raters, and projecting the raters into a low dimensional vector space where they are viewed. The visualization techniques are demonstrated using voting data, jokes, and movie preferences.

Index Terms—Partial rankings, incomplete rankings, multidimensional scaling.

1 Introduction

Ranking data arise from m raters ordering, by some mechanism, n items to express their preferences for the items. The data can arise in many ways such as directly ranking items, by voting for a subset, or by rating items on a subjective measurement scale. A ranking can be complete, which means all n items are ranked, or incomplete, which means some items are not ranked. A ranking, whether it is complete or incomplete, can be with-ties or without-ties; it is with-ties if some of the ranked items are not clearly preferred to others. Raters are often people but can also be computer programs; one example is search engines that provide a partial ordering of web sites.

For example, suppose m members of a professional society vote for the top k of n candidates for the society council, where $k < n$, and supply ranks of 1 to k for their votes. Each rater establishes an ordering on all items in which there are $n-k$ ties for last place. As a result, the rankings are complete and with-ties. If $n = k$ then the rankings are complete and without-ties. If a rater can add names to the list, not all raters have the same write-in process, and the items are the original list plus write-ins, then the rankings are incomplete and with-ties.

A very common mechanism for rating is to have raters provide their rating of an item by using a subjective rating scale, say, of 1 to 10. Frequently, the numeric scores reported by different raters are not comparable since they interpret the subjective scale in very different ways. For example, a rating of 10 issued by one person for a highly desirable item may correspond to a score of 8 for a second individual who thinks of 10 as perfect and believes nothing is perfect. In other words, the scale has little metric content and provides just an ordering; in this case, the result is ranking data, typically with ties.

Effective visualization of ranking data can reveal important statistical properties of the population of raters, of the items, and of an item-rater interaction. For example, graphs may reveal that some products are generally more popular than others, that high preference of one product by a customer entails low preference of another product, or that there are multiple clusters of rankings representing several distinct types of customers.

The visualization of ranking data differs fundamentally from visualizing numeric data. Rankings — whether complete or incomplete, and whether with ties or not — are discrete objects, rather than numeric vectors.

- *Paul Kidwell is with the Department of Statistics, Purdue University, E-mail: kidwell@purdue.edu.*
- *Guy Lebanon is with the College of Computing, Georgia Institute of Technology, Atlanta GA. Email: lebanon@cc.gatech.edu.*
- *William S. Cleveland is with the Department of Statistics and Computer Science, Purdue University, E-mail: wsc@purdue.edu.*

Manuscript received 31 March 2008; accepted 1 August 2008; posted online 19 October 2008; mailed on 13 October 2008.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

In this paper we develop an intuitive, easy to use, and computationally efficient framework for the visualization of ranking data. The framework starts by considering incomplete or tied rankings as sets of permutations representing full ordering that are consistent with the ranking data. Based on the Kendall's tau distance on permutations, we define a dissimilarity score on incomplete and tied rankings which corresponds to the expected or average distance between the underlying permutations. Finally, the dissimilarity score is used in conjunction with multidimensional scaling to project the data into a low dimensional continuous vector space for easy visualization.

In the next section we provide a detailed description of total, tied, complete, and incomplete rankings as sets of permutations. We then proceed in Section 4 to describe the metric structure on permutations and the expected distance dissimilarity measure. Section 5 explores different visualization techniques for ranking data, and Section 6 demonstrates these concepts with an experimental study on voting data and ratings of jokes and movies. We conclude with Sections 7 and 8 which contain a description of related work and a discussion.

2 Complete or Incomplete and With-Ties or Without-Ties Rankings

Rankings can be classified as without-ties or with-ties and as complete or incomplete. We start by defining complete without-ties rankings which correspond to permutations and then proceed to discuss with-ties and incomplete rankings. Most of the definitions and notations are similar to the ones in the monographs [7, 9, 11, 16] where more information can be found. We discuss the metric structure of rankings in Section 4 followed by various visualization techniques.

A complete, without-ties ranking of the items $S = \{1,\ldots,n\}$ is a permutation of S which we denote as a bijection $\pi : S \to S$ mapping items to ranks. Identifying permutations with bijective functions, we have that the rank of item i is $\pi(i)$ and the item ranked j is $\pi^{-1}(j)$. The set of all permutations over n items is the symmetric group which we denote by $\mathfrak{S}_n$. We will represent a permutation by a sorted list of the items, most preferred to least, separated by vertical bars i.e. $\pi^{-1}(1)|\cdots|\pi^{-1}(n)$; for example, for $n = 5$ one permutation ranking item 3 as first and 2 as last is $3|5|1|4|2$.

Complete, with-ties rankings are similar to complete without-ties rankings but they allow some of the items in S to be of tied rank. We continue to represent such rankings using the vertical bar notation but now tied items are separated by commas, rather than vertical bars. For example, $3|1,2|4$ implies item 3 is the most preferred, item 4 is the least preferred, and items 1 and 2 are tied for the middle ranks.

Conceptually we take a tie to be a lack of information with the notion that more information could in principle break the tie. This means that a permutation with ties can be though of as a set of permutations where each member of the set is the potential true permutation if we had the full information. This idea is incorporated in the term full ranking which is defined as a ranking without-ties. For example the with-ties ranking $3|1|2,4$ corresponds to the following set of permuta-

1077-2626/08/$25.00 © 2008 IEEE Published by the IEEE Computer Society

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

tions

$$3|1|2,4 = \{3|1|2|4;3|1|4|2\} \subset \mathfrak{S}_4.$$

In our notation, each integer or sequence of integers separated by commas is a compartment. In the last example the compartments are 3, followed by 1, and finally 2,4. Denoting by r the number of compartments and by n_j the size of compartment j the size of the set of permutations corresponding to a complete with-ties ranking is $\prod_{j=1}^{r} n_j!$.

Popular types of complete with-ties rankings include top-k rankings, which are often encountered in polls or elections and in the ranked list of web-pages returned by search engines as a response to a query. Another popular complete with-ties ranking type is an unordered list of the top k items representing more preferred and less preferred items, for example, $3,5,7|2,4,6$.

When the number of items, n, is very large, rankings are often incomplete indicating that some items are missing. Our notional conventions continue in a similar fashion for incomplete rankings. For example, let $n = 20$ then $5|8|2$ is an incomplete without-ties ranking and $5|2,8$ is an incomplete with-ties ranking.

Incomplete rankings, with or without ties, may also be identified as sets of permutations that are consistent with it. For example, the incomplete ranking $4|2$ with $n = 4$ corresponds to $4|2|\bullet|\bullet$, $4|\bullet|2|\bullet$, $4|\bullet|\bullet|2$, $\bullet|4|2|\bullet$, $\bullet|4|\bullet|2$, and $\bullet|\bullet|4|2$, where each $\bullet$ symbol denotes a possible placement of the missing items $1,3$:

$$4|2 = \{4|2|1|3;4|2|3|1;4|1|2|3;4|3|2|1;4|1|3|2;4|3|1|2;1|4|2|3;$$
$$3|4|2|1;1|4|3|2;3|4|1|2;1|3|4|2;3|1|4|2\} \subset \mathfrak{S}_4.$$

3 Ranking Datasets and their Visualization

A ranking dataset $\mathscr{D}$ consists of a list of rankings over a common set of items $\{1,\ldots,n\}$. The rankings can be either without-ties or with-ties and either complete or incomplete. Some ranked datasets contain rankings that are all of the same type e.g.,

$$\mathscr{D} = \{3|1|2,4;1|3|2,4;1|2|3,4;1|4|2,3\}$$

while others are more heterogeneous, for example

$$\mathscr{D} = \{3|4;1|3|2,4;1|2|3|4;\ 1|3,4\}.$$

Since rankings are identified as sets of permutations consistent with it (see Section 2), a ranking dataset corresponds to a set of subsets of the symmetric group $\mathscr{D} = \{A_1,\ldots,A_m\}$, $A_i \subset \mathfrak{S}_n$.

The visualization of $\mathscr{D}$ is complicated for the following reasons. Its elements are not vectors which prevents the use of standard vector space visualization techniques. It is not clear how to relate one ranking to another, in particular if they are not of the same type. For example $2|1|3,4$ and $3|4$ correspond to two sets of permutations $A_i, A_j \subset \mathfrak{S}_n$ which are of different sizes and are neither disjoint nor contained in each other. Finally, if the number of items n is large (as is often the case), the number of permutations becomes overwhelming which raises substantial visualization and computational difficulties.

We address these difficulties by developing a natural dissimilarity measure between rankings which is then used by multidimensional scaling to embed the rankings in a low dimensional vector space. When viewed as points in a low dimensional vector space, the rankings may be effectively visualized and interpreted.

In Section 4 we develop the dissimilarity measure for both complete and incomplete and with-ties and without-tied. In Section 5 we describe various visualization techniques based on the developed dissimilarity measure. These techniques are then demonstrated in Section 6 on three ranking datasets.

4 Distances and Dissimilarities on Rankings

Kendall's tau $T(\pi,\sigma)$ [15] is the most popular choice of distance between permutations and the one we focus on in this paper. It can be interpreted as the number of pairs of items for which π and σ have opposing orderings (called disconcordant pairs) or the minimum number of transpositions of adjacent items needed to bring π to σ. For example, $T(\pi,\sigma) = 1$ for $\pi = 1|2|3$, $\sigma = 2|1|3$ since transposing the adjacent items 1 and 2 in π results in σ.

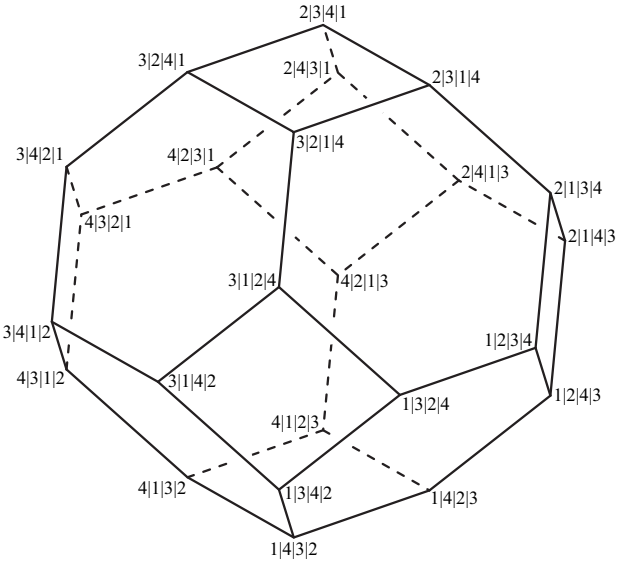

Fig. 1. Permutation polytope for 4 objects represented in 3D space.

A useful visualization tool for the metric structure $(\mathfrak{S}_4, T)$ is the permutation polytope [18] whose vertices correspond to permutations and whose edges correspond to adjacent transposition of items. In other words, two vertices that are connected by an edge correspond to Kendall's tau distance 1 and more generally the distance between two permutations is the length of the shortest path on the polytope between the two corresponding vertices. The permutation polytope for 4 items is displayed in Figure 1 where it is embedded in $\mathbb{R}^3$. When the number of items n is larger than 4 the polytope cannot be embedded in $\mathbb{R}^3$ and using it for visualization purposes [18, 2] is rather limited.

A natural way to extend T to rankings that are incomplete or with-ties is to consider the expected distance between the sets, A,B, of full and complete rankings which are consistent with the two observed rankings. The expectation is calculated with respect to a uniform distribution over the sets of consistent rankings i.e.,

$$T^*(A;B) = \frac{1}{|A|\cdot|B|}\sum_{\pi\in A}\sum_{\sigma\in B} T(\pi,\sigma). \tag{1}$$

We use a semicolon in $T^*(\cdot;\cdot)$ instead of a comma to avoid confusion with the commas in the with-ties rankings.

For any but the smallest A,B a direct calculation of (1) would involve an intractable summation. However, in some cases it is possible to efficiently compute (1), even for large n and large sets A,B. The efficient calculation of (1), which we present below, is based on the combinatorial properties of the Kendall's tau distance. These properties were first noted by Alvo and Cabilio, a more complete description is available in [16] and [1]. An alternative extension of T to incomplete and tied rankings may be obtained based on the Hausdorff distance construction. See [7] for more details.

In the case where A,B represent two incomplete full rankings σ_1,σ_2 each expressing preference over k_1,k_2 items respectively,

$$T^*(A;B) = \frac{n(n-1)}{4} + \sum_{i}^{n-1}\sum_{l>i} a_1(i,l)a_2(i,l) \tag{2}$$

where

$$a(i,l) = \begin{cases} 2I(\sigma(i)-\sigma(l))-1 & \text{if } i \text{ and } l \text{ are ranked} \\ 2\frac{\sigma(i)}{k+1}-1 & \text{if only } i \text{ is ranked} \\ 1-2\frac{\sigma(l)}{k+1} & \text{if only } l \text{ is ranked} \\ 0 & \text{otherwise} \end{cases}.$$

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

Intuitively, the terms $a_1(i,l)$ and $a_2(i,l)$ represent the contribution of each pair of items with respect to the two observed rankings. The subscript differentiates between the two observed rankings. The values correspond to the 4 scenarios of incompleteness: (1) both items were ranked, (2,3) only 1 item was ranked, and (4) neither item was ranked.

For example, assuming $n = 4$ we have

$$T^*(4|2;3|4|1) = 3 + \sum_i^3 \sum_{l>i} a_1(i,l)a_2(i,l)$$
$$= 3 - \frac{1}{3}\cdot\frac{1}{2} + 0\cdot 1 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot\frac{1}{2} + 1\cdot 0 + \frac{1}{3}\cdot 0 = \frac{10}{3}.$$

Note that for two incomplete rankings that do not share any common objects the expected distance is $n(n-1)/4$ which agrees with the the average of Kendall's tau over the entire set of permutations $\mathfrak{S}_n$. Each common object adjusts this expectation by accounting for relationships among both the pairs of observed objects in the standard way and unobserved objects by considering their possible position relative to the observed objects. Computationally, using (2) requires $O(\min(k_1,k_2)n)$ where k_1 and k_2 are the number of observed rankings in A and B. This constitutes a dramatic improvement over the overwhelming complexity $\Omega((n-(k_1+k_2))!)$ required by a direct calculation of (1).

The computation of $T^*(A;B)$ for A,B corresponding to with-ties rankings can be efficiently computed using a common representation for all rankings in the class of compatible rankings. Objects considered to be tied are mapped to a shared position. A permutation is a mapping, τ, of objects to positions; however, unlike an ordinary permutation which maps each item to a distinct spot, ties are identified by mapping items to the same spot represented by a common number. For example, given preferences $3,4|1,2$ and $3|1,2|4$ the corresponding mappings are $\tau_1 = (2,2,1,1)$ and $\tau_2 = (2,2,1,3)$. The repetition of numbers indicate that multiple objects are tied, e.g. in τ_2 a 2-way tie exists for second place. The resulting formula for a with-ties ranking π,σ corresponding to the A,B is

$$T^*(A;B) = \sum_{i=1}^{n-1} \sum_{l>i} \phi((\tau_1(i)-\tau_1(l))(\tau_2(i)-\tau_2(l))) \qquad (3)$$

where $\phi(x) = 0$ for $x > 0$, $\phi(0) = \frac{1}{2}$, and $\phi(x) = 1$ for $x < 0$. For example using the rankings in the example above we obtain

$$T^*(3,4|1,2;3|1,2|4) = \phi(0\cdot 0) + \phi(1\cdot 1) + \cdots + \phi(0\cdot(-2))$$
$$= 0.5 + 0 + 1 + 0 + 1 + 0.5 = 3$$

Computationally, (3) requires $O(n^2)$ complexity which is substantially better than the $O(\prod_r N_r^A! + \prod_r N_r^B!)$ required by a direct calculation of (1). Here, N_r^A, N^B represent the number of tied items at different ranks in the with-ties rankings corresponding to A,B (see Section 2).

5 Visualization Techniques

The expected Kendall's tau distance T^* between two partial or incomplete rankings enables the embedding of ranking data in a Euclidean space $\mathbb{R}^2$ or $\mathbb{R}^3$ through multidimensional scaling. We start by describing briefly multidimensional scaling and then proceed to describe a variety of visualization techniques operating on the embedded data.

5.1 Multi-Dimensional Scaling

In our context of ranking data, multi-dimensional scaling (MDS) finds an embedding of ranking data $\{A_1,\ldots,A_m\}$ in Euclidean space i.e., $A_i \mapsto z_i$ with $\{z_1,\ldots,z_n\} \subset \mathbb{R}^2$, such that the distortion introduced by the embedding

$$R(z_1,\ldots,z_m) = \sum_{i,j} (T^*(A_i,A_j) - \|z_i - z_j\|)^2$$

is minimized. In other words, the coordinates $z_1,\ldots,z_m$ in $\mathbb{R}^2$ corresponding to the incomplete or with-ties ranking data are selected in a way that minimizes the total distortion of distances

$$(z_1,\ldots,z_m) = \arg\min_{z'_1,\ldots,z'_m} R(z'_1,\ldots,z'_m).$$

Note that multidimensional scaling is well-defined even if the function T^* is an expected distance rather than a formal distance function. In this case multi-dimensional scaling is a more appropriate choice than PCA since the distances are not formal. More details on multidimensional scaling may be found in [6].

To illustrate the application of multidimensional scaling to ranking data we demonstrate an MDS embedding of synthetic data in Figure 2. In the first case (left panel) we construct an embedding of the following fully ranking data

$$\{1|2|3|4|5|6,;\ldots;1|2|6|5|4|3\} \cup \{6|5|1|2|3|4;\ldots;6|5|4|3|2|1\}.$$

The embedded points correspond nicely to two clusters in the top and bottom parts of the two dimensional plane. The first cluster corresponds to the second set of rankings consisting of all permutations ranking items 6 and 5 in the top two positions. The second cluster corresponds to the first set of rankings consisting of permutations ranking items 1 and 2 in the top two positions.

In the second case (Figure 2, right) we construct an embedding of the incomplete ranking data

$$\{1|2|3;1|2|4;1|2|5;1|2|6\} \cup \{6|5|1;6|5|2;6|5|3;6|5|4\}.$$

As expected, we obtain two clusters in the top and bottom of the two dimensional plane corresponding to the two sets of ranking. The first cluster corresponds to rankings with items 1 and 2 occupying the top two ranks and the second cluster corresponds to rankings with items 6 and 5 occupying the top two ranks.

In both cases the points with largest and smallest distance with respect to $T^*(A_i,A_j)$ are also the most and least distant respectively in the the embedded space $\|z_i - z_j\|_2$. For example, in Figure 2 (right) the rankings $A_i = 6|5|1, A_j = 1|2|6$ are most distant in T^* and the the corresponding Euclidean distance $\|z_i - z_j\|_2$ is maximal. We thus conclude that while the precise spatial relationship between the ranking data is somewhat distorted (perfect embedding is impossible in this case), the major spatial qualities of the ranking data are mostly preserved by the MDS embedding.

5.2 Heat Maps

Plotting the embedded ranking data $\mathscr{D} = \{A_1,\ldots,A_m\}$ using a scatter plot as in Figure 2 is ineffective when the number of rankings m is large. Instead, assuming that the embedded rankings are drawn from a population $z_1,\ldots,z_m \sim p$, we use a non-parametric density estimation technique [20]

$$\hat{p}(z) = \frac{1}{m}\sum_{i=1}^{m} K_h(z,z_i), \quad K_h(x,y) = h^{-1}\exp(-h^{-1}\|x-y\|^2) \qquad (4)$$

to estimate the underlying density p on $\mathbb{R}^2$. We then proceed to plot the estimated density $\hat{p}$ by translating its numeric values to colors which are drawn as an image. For large datasets the resulting plot is less cluttered than a scatter plot and demonstrates nicely the distribution of points through the embedded two dimensional space.

We examine several ways of translating the numeric values of $\hat{p}(z)$ to colors based on the power transform, which is a continuous family of monotonic transformations parameterized by $\lambda > 0$

$$y \mapsto y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}. \qquad (5)$$

Thus, instead of mapping the values of $\hat{p}(z)$ to a colormap (say grayscale or red-blue) in a linear way we use the numeric values $\hat{p}^{(\lambda)}(z)$ which amounts to a power transformation of the estimated density $\hat{p}$. Specifically, varying $0 < \lambda < 1$ emphasizes differences in regions of low density over regions in high density (see Figure 3). In

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

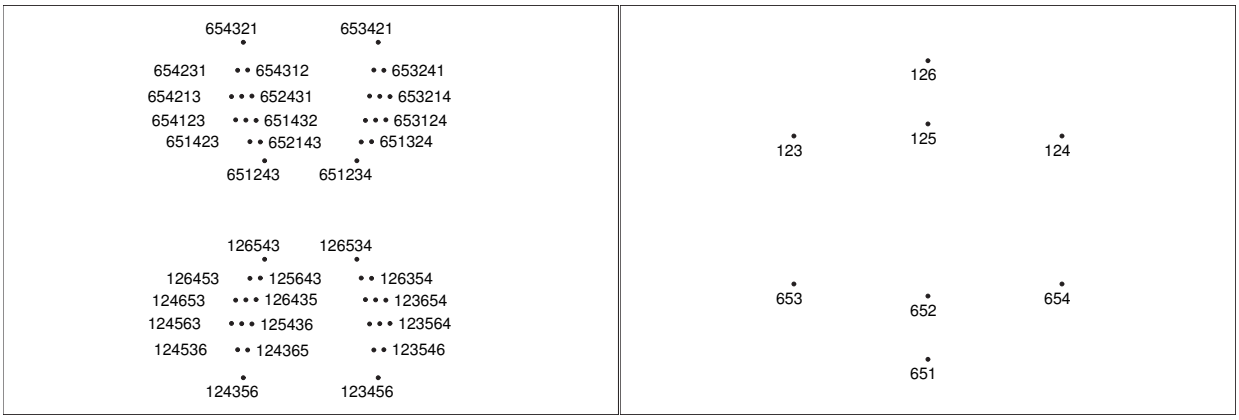

Fig. 2. Synthetic full rankings (left) and incomplete rankings (right) embedded in 2D using multidimensional scaling.

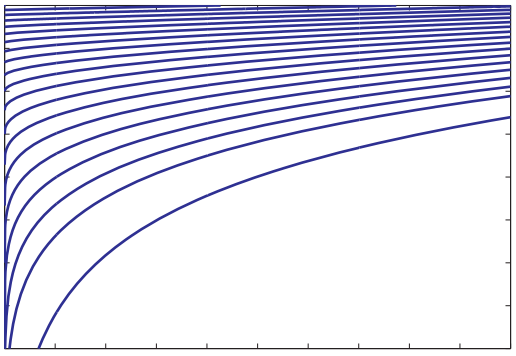

Fig. 3. Power transform for λ ranging from 0 (bottom) to 1 (top).

practice both λ and h can be regarded as tuning parameters for the visualization, a user iteratively interacting can select the parameter values yielding the most informative displays. As we see in the next section, this is an important component of the visualization system since in some cases visualizing $\hat{p}$ (or $\hat{p}^{(\lambda)}$ for $\lambda = 1$) focuses exclusively on a small region of very high density.

5.3 Subset Selection

In some situations the visualized density $\hat{p}$ should be computed based on only a subset of the data $\mathscr{D}' \subset \mathscr{D}$. For example, it may be desirable to focus only on raters satisfying a certain demographic criterion such as an age group or geographic location. In other cases it may be desirable to consider only rankings satisfying some constraints such as having a certain item in the top 10. These restrictions enable the visualization system to focus on a subset of the rankings that are of particular interest and that would otherwise be overwhelmed by the remaining data.

For large datasets, it is sometimes easier to visualize semantically meaningful parts of it. For example, the dataset $\mathscr{D}$ can be divided to l groups corresponding to the demographics of the raters $\mathscr{D}_1, \ldots, \mathscr{D}_l$. Due to the smaller size of each set and its semantic coherence the sets can be visualized separately at first producing a collection of visual cues. Then, the entire dataset can be visualized in order to relate the rankings found in the different subsets to each other.

5.4 Retention and Censoring

Two alternatives to subset selection that enable a different form of directed visualization are retention and censoring. These techniques visualize the data by conducting MDS and density estimation based on the original data $\mathscr{D}$, equipped with a modified geometry that emphasizes certain desired aspects.

It is easy to explain retention and censoring by transforming the original ranking data $\mathscr{D} = \{A_1, \ldots, A_m\}$ to a different but related set $\mathscr{D}' = \{A'_1, \ldots, A'_m\}$ of rankings that are then input to the MDS and density estimation procedures. Denoting the transformation by $g(A_i) = A'_i$ we have that visualizing the transformed data using MDS is equivalent to MDS on the original data, but under a different distance measure or geometry

$$T^*_g(A_i, A_j) = T^*(g(A_i), g(A_j)). \tag{6}$$

Different transformations g correspond to different geometric structures T^*_g emphasizing some aspects of the data over others. The two dimensional embedding obtained by MDS using T^*_g would therefore depend highly on the nature of g thus reflecting the desired emphasis.

In retention, a certain set of items S is selected and all the rankings A_i are mapped to $A'_i = g(A_i)$ corresponding to the preference relation in A_i restricted to S. For example, for $S = \{1,3\}$ we have

$$\mathscr{D} = \{3|1|2,4\,;1|3|2,4\,;1|2|3,4\,;1|4|2,3\} \mapsto \mathscr{D}' = \{3|1\,;1|3\,;1|3\,;1|3\}.$$

Visualizing the embedded coordinates of $\mathscr{D}$ through MDS using T^*_g emphasizes the importance of the ranking of items in S. This is often useful if there is a large number of items and it is desirable to concentrate on different subsets of items in different stages. For example, in the case of movie rankings it may be useful to first visualize rankings involving only a certain genre such as comedy followed by ranking of another genre such as drama.

Censoring transforms the ranking data $A_i \mapsto g(A_i)$ to a consistent but more incomplete or with-ties version of it. For example, removing from the data all information not pertaining to the top two items we obtain

$$\mathscr{D} = \{3|1|2,4\,;1|3|2,4\,;1|2|3,4\,;1|4|2,3\} \mapsto \mathscr{D}' = \{3|1\,;1|3\,;1|2\,;1|4\}.$$

In contrast to retention which focuses on specific items, censoring focuses on specific ranks. This mechanism can be used to visualize the distribution of top k or bottom k rankings. If $k = 1$ we obtain a visualization of the population of the most preferred or least preferred item. More interestingly, selecting a small $k > 1$ we obtain a visualization of the ranking data restricted to the few most or least preferred items. Returning to the movie example, this approach could be used to visualize the viewers perspectives regarding the top three films of all time. Another example is web-search where censoring for the top k items emphasizes the distribution of websites relevant to a query.

The visualization system allows retention and censoring to be used in tandem thereby keeping only those orderings over a subset of objects, and then looking exclusively at the favorites among this group, e.g. Figure 4. In the case of the movie example, this would produce for example a visualization of the three best comedy films of all time.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

5.5 Clustering

The heat map density plot of the original dataset or a transformed one (via subset selection, retention, or censoring) provides a nice visual summary of the spatial features of the population p. Statistical analysis can aid further the visualization by automatically identifying clusters corresponding to regions of high density.

Standard clustering techniques such as k-means produce a partition of the data to k groups of spatially coherent clusters. Applying it to the ranking data provides an automatic partition of the data where distinct clusters correspond to a part of the population having similar preference relations. In other words, clustering stratifies the data to k different types of raters. For example in the case of voting data the rater types could correspond to different demographic such as geographic location, age, gender, and race.

In addition to providing an automatic division of the rankings to spatially coherent types, clustering enables the visualization system to add meaningful labels to the heat map density plot without the risk of adding unnecessary clutter. The labeling produced by the system correspond to certain statistics of interest such as the average rank of an item or the probability of an item appearing within the top l ranks. The labels are computed per cluster and are displayed in a point that visually identifies the cluster such as the cluster centroid.

5.6 Point Labeling and Zoom

An important feature in most visualization systems is the user's ability to interact with it. We already mentioned a few interactive features such as the translation of numeric value to color, subset selection, and censoring and retention. We describe next two additional highly interactive features - point labeling and zoom.

Point labeling refers to the situation in which a user is interested in locating the point on the two dimensional heat map plane corresponding to a certain ranking of interest. For example, the user may be interested in the embedded two dimensional coordinates of a certain top 10 ranking. Unfortunately, MDS applied to the ranking dataset $\mathscr{D}$ does not construct two dimensional coordinates for rankings not appearing in the dataset $\mathscr{D}$.

To resolve this difficulty, we augment the ranking dataset $\mathscr{D}$ with additional representative rankings called anchor points. Since the anchor points are added to $\mathscr{D}$ before the MDS is applied, they receive two dimensional coordinates as well. The list of anchor points is then made available to the user and may be marked on the heat map in order to aid its spatial interpretation.

Zooming into a certain region of the heat map can be done in two different ways. The first is by subset selection (Section 5.3 which focuses on a subset of the original data corresponding to rankings that fulfill a certain criterion. The second is by specifying a certain region of interest in the two dimensional map and redrawing the heat-map in that region over the entire figure.

The first zoom technique is more appropriate when the user knows in advance what type of rankings are relevant. The second zoom technique is more appropriate when the user unexpectedly observes an interesting spatial feature in the heat map such as a cluster and wishes to examine it in more detail.

6 Experiments

In this section we employ the framework described in the preceding section to analyze three ranked datasets: Jester, APA, and MovieLens.

6.1 Jester Dataset

The Jester dataset contains incomplete rankings of 100 jokes by 73,421 users during April 1999 and May 2003 [13]. Each user provided a numeric scores in the range from -10.00 to 10.00 for at least 20 out of the 100 jokes. Since the rating scale is nearly continuous there are very few ties and we can essentially consider this dataset as a set of incomplete full rankings. In the experiments below we visualize a set of incomplete ratings $\mathscr{D}$ obtained from 5,000 randomly selected users.

The heat map representing the estimate $\hat{p}$ obtained from $\mathscr{D}$ is displayed in Figure 4 using the power transform with $\lambda = 1, 1/3, 0$ (left, center, and right, respectively). The left panel corresponding to $\lambda = 1$ shows a massive cluster at the center of the two dimensional embedding space. Reducing λ to $1/3$ (middle) and 0 (right) reveals additional smaller clusters that are invisible without the use of a nonlinear transform. Note that $\lambda = 1/3$ reveals 6 additional smaller clusters while $\lambda = 0$ reveals additional sub-clusters within these six clusters.

	50	29	62	27	54	35	36
Joke	Old	Scott	Engi	Clinton	Celeb	Nat	Poles
Top 1	.037	.038	.026	.032	.031	.027	.025
Top 5	.038	.034	.029	.031	.029	.029	.028

Fig. 8. The probabilities of a joke ranking in the top k. First censoring is used to select the top k jokes, then probabilities are calculated given the censoring. The jokes listed correspond to those with the highest probability of appearing in the top 5. Joke 29 has the highest probability of being ranked first, but is not frequently ranked between 2 and 5.

We proceed next to illustrate the use of the censoring technique described in Section 5. Figure 5 displays the heat map obtained after censoring the incomplete rankings to top 5 (left) and top 3 (middle). Figure 5 (right) contains a zoomed version of bottom right cluster of the middle panel. The resulting visualization emphasizes differences between rankings in the top 5 or 3 items and ignores differences in lower ranked items. Such emphasis of differences among top ranked items is important in several applications. For example, in web search a mistakenly placed entry in the topmost rank could prove much more severe than in the bottom rank. Similarly, assuming that the jokes ranked at the bottom are of poor quality, it may make sense to consider visualizing the top k jokes in order to visualize the high quality jokes.

The embedded points were clustered using k-means and for each cluster we computed basic statistics that are displayed near the cluster's centroid. The displayed statistics in Figure 5 correspond to a list of prominent items accompanied by the average rank of items (within the cluster) in parenthesis. The average ranks represent the overall preference of the item for rankings belonging to different clusters. Below the items and their average ranks we display the size of the cluster in terms of the number of rankings it contains.

For example, in clusters in the top of the left panel of Figure 5, we can observe that joke number 50 (usually ranked 1 or 2) is usually preferred to joke number 29. On the other hand, in clusters at the bottom left of the panel joke 29 (usually ranked 1 or 2) is preferred to joke 50. The highly popular jokes which are used to label the clusters correspond with those identified by examining tabulations of preferences frequencies in Figure 8. We see that while joke 50 seems to be most frequently ranked in the top 5, joke 29 is more frequently listed as the top joke. The global statistics that indicates a close similarity between the preferences of jokes 29 and 50 is replaced by substantially different and finer local preferences. For example, some clusters in in the middle of the left panel have a relatively low average rank for joke 50 which is globally ranked high (Figure 8). The middle panel displays the heatmap corresponding to top 3 censoring and reveals a much coarser clusterings with 7 clusters, one of which is dominating in size the remaining clusters. The right panel contains a zoomed view of the bottom right cluster in the middle panel. The cluster which seemed homogenous is split into sub-clusters which all have joke 50 within the top 3 but they differ with respect to the remaining items.

6.2 APA voting

The APA dataset contains 15,449 votes for the presidency of the American Psychological Association in 1980. Each ballot contained a ranking of the top 5 candidates, except in some cases where ties were observed. The dataset has been extensively analyzed in [9] and other sources which found that the voting population was divided into 3 distinct blocs.

We visualize the rankings by displaying in Figure 6 (left) the heatmap corresponding to 4,000 randomly selected ballots from the data. The middle panel zooms in on the bottom cluster which appears to be two clusters close to each other. The right panel zooms in the top cluster which appear to be somewhat bean-shaped.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

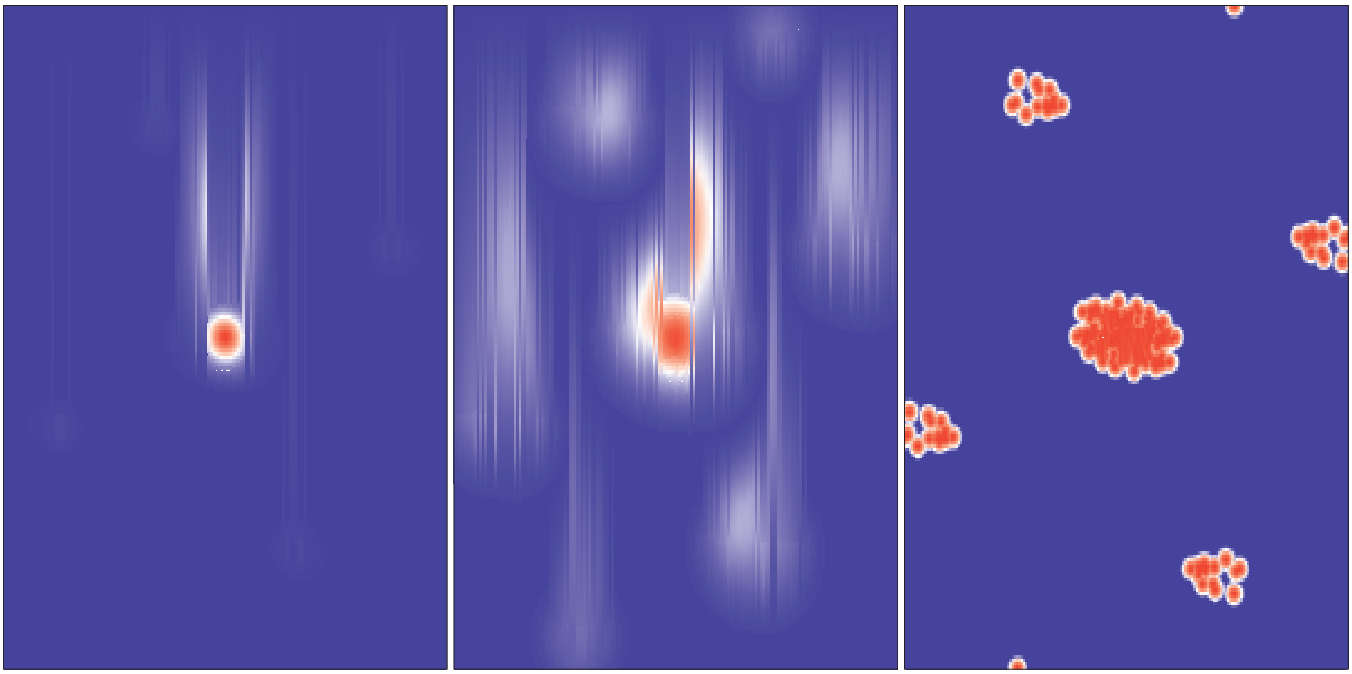

Fig. 4. The heat map corresponding to $\hat{p}$ obtained from the Jester data. The three panels represent the use of the power transform with values $\lambda = 1, 1/3, 0$ (left, center, and right, respectively). Decreasing the value of λ emphasizes differences in regions of low density as opposed to the central region of high density.

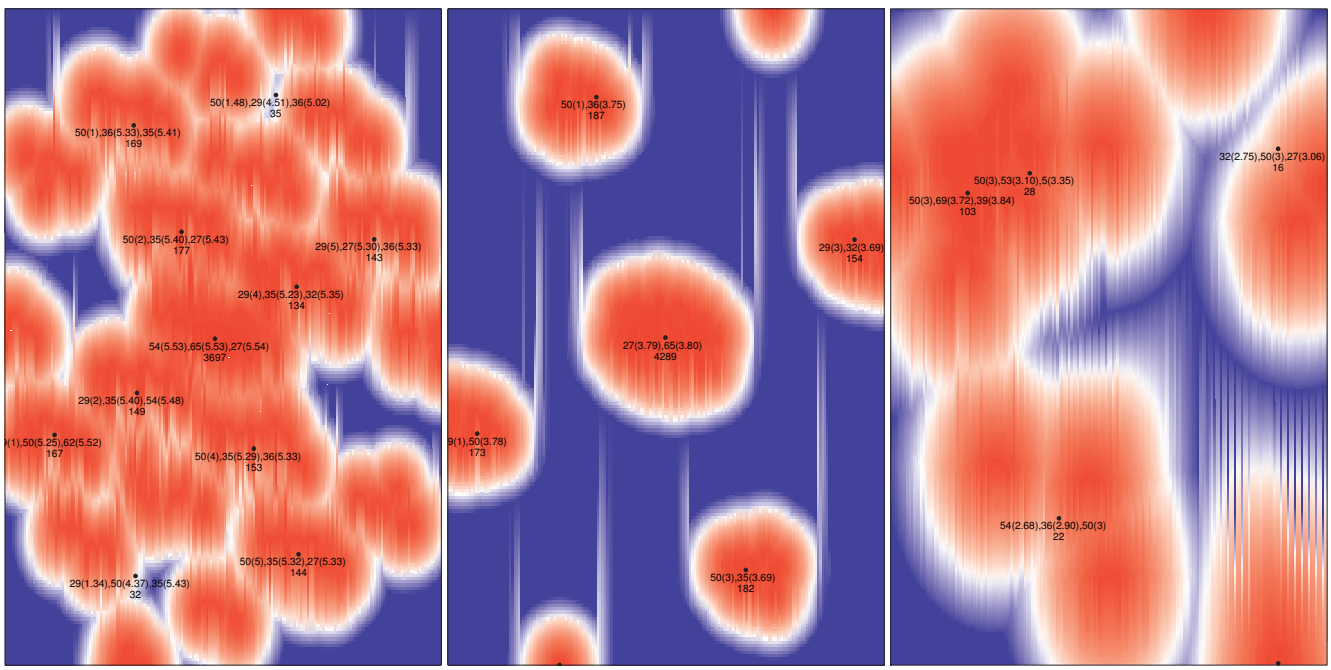

Fig. 5. Heatmap corresponding to the Jester data with censoring. Left panel corresponds to top 5, middle panel corresponds to top 3, and right panel corresponds to a zoomed version of the middle panel. Cluster labels indicate the highest ranked jokes followed by their mean ranking below which the cluster size is displayed.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

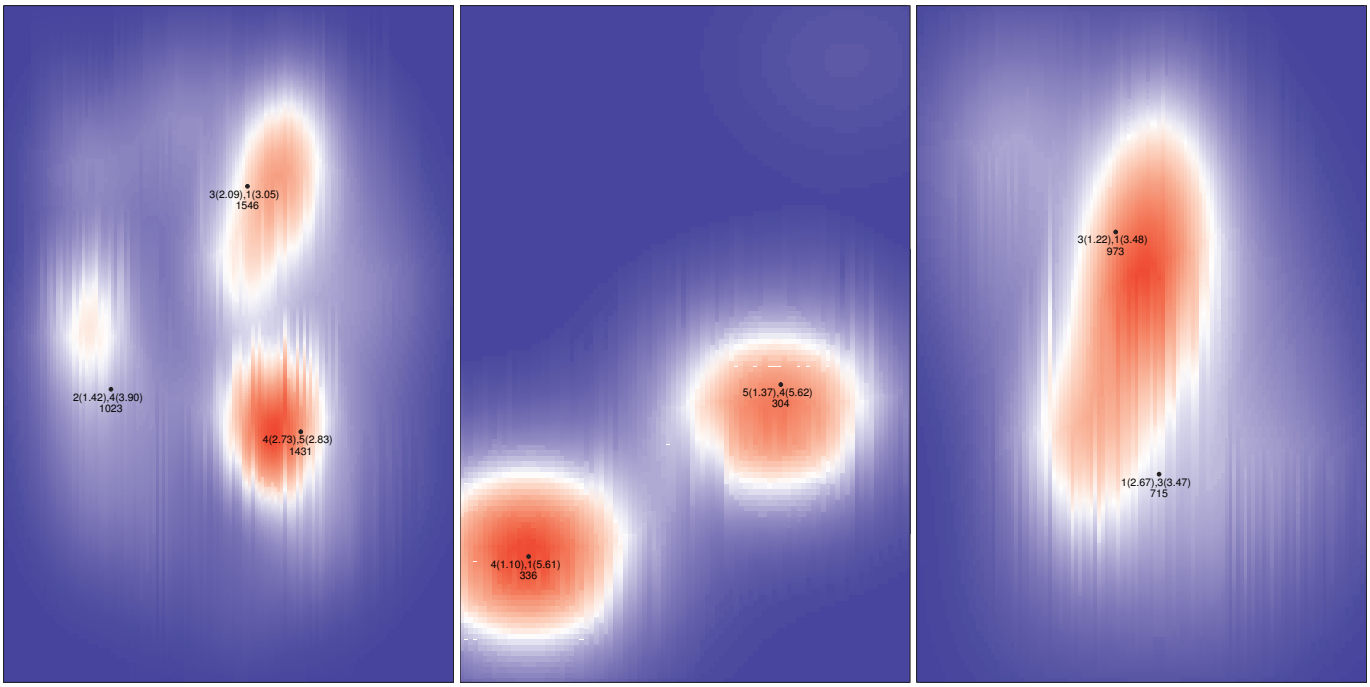

Fig. 6. The overall picture of the voting distribution shows 3 distinct voting blocs (left). Zooming in on the bottom cluster (middle) reveals two sub-clusters – one with candidate 4 ranked first and one with candidate 5 ranked first. Zooming in on voters in the top cluster (right) reveals a clear preference for candidate 3 over 1 in the top region and a preference of candidate 1 over 3 in the bottom region.

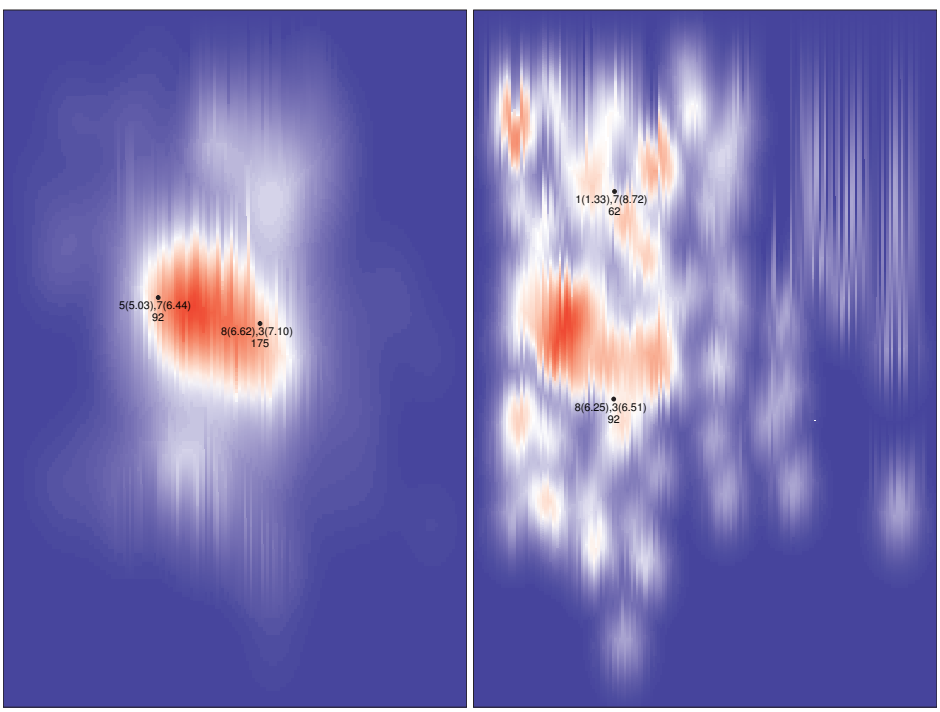

Fig. 7. The overall picture (left panel) shows a single large cluster, but inspecting the results of k-means clustering reveals preferences differ from the left to the right. Zooming in (right panel) on the right half of the cluster produces a separation among the fans of romance films. Fans of dramatic romance are in the bottom cluster, while fans of lighter romance fall in the top cluster.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.

The visualization seems to indicate the voters corresponding to the bottom cluster are indifferent to candidates 4 and 5 (average ranks are in the middle ranks of 2 and 3). However, zooming into this cluster (middle panel) reveals that there are actually two sub-clusters within it with the top one corresponding to candidate 5 typically selected as the topmost choice and the second corresponding to candidate 4 typically selected as the topmost choice. Similarly, the right panel reveals that although both items 3 and 1 receive middle ranks in the top cluster, the top of that cluster correspond to voters that favor candidate 3 while the bottom correspond to voters than prefer candidate 1.

6.3 Movie ratings

The MovieLens web site collected data over the seven-month period beginning September 19th, 1997 and ending April 22nd, 1998. The data set contains a total of 100,000 ratings (1-5) from 943 users on 1682 movies. Each movie was categorized by genre although a single film could fall into multiple genre. We examined the user's movie preferences by choosing subset of 12 movies including action, drama, romance, and children's films. Our sample included 267 users rating at least two of the 12 movies.

The movie viewing population initially appears to be a single large cluster in the left panel of Figure 7. A closer look identifies a dichotomy within the single large cluster where preferences vary from the left to the right. The left half of the cluster appears to appreciate action films, Highlander (5) and Die Hard II (7), while the right half leans towards romance, The Piano (8) and Gone With the Wind (3). A closer look at the right side of the cluster (left panel) produces two clusters within the romance genre. The bottom cluster contains dramatic romances, 8 and 3, while the top cluster has films of the romance genre not categorized as dramatic, Sleepless in Seattle (1) and Dirty Dancing (4).

7 Related Work

Visualizing ranking data has been an ongoing challenge that has received attention from a number of communities ranging from pioneering efforts by Spearman motivated in psychology to more recent efforts by the computer science and statistics communities. The ubiquity of ranking data and the vast number of ranking forms have led to a variety of visualization techniques.

A popular visual representation of the ranking space is the permutation polytope as coined by Yemelichev, Kovalev, and Krasov [21]. It is defined to be the convex hull of $n!$ points in $\mathbb{R}^n$ where each vertex represents a complete ordering of the n items. A vehicle for the visualization of probability distributions over the polytope was developed by Cohen and Mallows [5] and extended by Thompson [18] and Baggerly [2]. This method replaces each vertex of the polytope by a sphere with diameter proportional to the probability of the corresponding permutation. The polytope approach is extremely effective for preserving relationships between objects when the number of items are small but is ineffective for large n.

Marginal statistics and pairs have been established as an effective approach to overcoming the problem of large n by [8], [10], and [14]. Basic marginal plots can be built by constructing a matrix of object by rank filled with spheres whose radii are proportional to ranking probability. Many relationships can be identified by expanding from pairs to triples and larger as in [17]. Parallel coordinate axis plots developed by Inselberg [14] place objects on the x-axis and rankings on the y, while simultaneously plotting each ranking as a series. This has been shown to be effective for identifying relationships among a small number of raters . More recently, Batty [3] changed the parallel coordinate method into a ranking clock to view rankings over time.

Projecting polytopes is a method that has been frequently used for overcoming large n. Gabriel's [12] bi-plot is similar to principal component analysis (PCA) or MDS; the idea is to locate the center of the polytope and calculate the Euclidean distance between this point and each ranking. This projection method maximizes the spread of points in 2D where the points of interest are outlying. Other efforts utilizing PCA and MDS include Ukkonen's [19] scatter plots which use an alternative distance to deal with incomplete rankings.

The structure of rank data and the computational advantages afforded by Kendall's tau has led to its popularity [9], [1]. Alvo and Cabilio [1] have shown it to be convenient for analyzing rankings in the face of missing data. Recently, Busse et al. [4] have used this metric to cluster heterogeneous rank data using top k partial rankings.

Our approach leverages the strengths of previous approaches in the analysis of ranking data together with visualization. This combination enables our approach to deal with a variety of real world data structures including both with-ties and incomplete rankings. The use of kernel smoothing facilitates the analysis of datasets with a large number of raters. Kendall's tau creates computational efficiencies necessary for distance calculations involving large numbers objects while preserving the notion of similarity described by the permutation polytope.

8 Discussion

Our visualization framework is based on three main components. First we express a natural measure of dissimilarity of rater's rankings using Kendall's tau T for complete and without-ties ranking data, and extent this to incomplete or with-ties rankings by taking expectations. Second, we use multidimensional scaling to embed the rankings in two dimensions in a way that provides a best fit of the 2-D distances to the actual distances as measured by the expected Kendall's tau. The third component is a nonparametric estimate of the density of the projected rankings, which allows us to visualize the behavior of the raters.

In addition to these three components we examine several visualization techniques: heat maps, power transformation, retention and censoring, clustering and labeling, and zoom and subset selection. Using these techniques, the visualization system is flexible enough to emphasize desirable aspects of the data while de-emphasizing less desirable aspects. Furthermore, the visualization system is computationally efficient due to Alvo and Cabilio's efficient computation of T^*.

References

[1] M. Alvo and P. Cabilio. Rank correlation methods for missing data. *The Canadian Journal of Statistics*, 23(4):345–358, 1995.

[2] K. A. Baggerly. *Visual Estimation of Structure in Ranked Data*. PhD thesis, Rice University, 1995.

[3] M. Batty. Rank clocks. *Nature*, 444:592–596, 2006.

[4] L. Busse, P. Orbanz, and J. Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[5] A. Cohen and C. Mallows. Analysis of ranking data. Technical report, Bell Laboratories, 1980.

[6] T. F. Cox and M. A. Cox. *Multidimensional Scaling*. CRC Press, 1994.

[7] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*. Lecture Notes in Statistics, volume 34, Springer, 1985.

[8] H. David. *The Method of Paired Comparisons*. Oxford Press, 1988.

[9] P. Diaconis. *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, 1988.

[10] P. Diaconis. A generalization of spectral analysis with application to ranked data. *Annals of Statistics*, 17(3):949–979, 1989.

[11] M. A. Fligner and J. S. Verducci, editors. *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, 1993.

[12] K. Gabriel. Biplot. *Encyclopedia of Statistical Science*, 1:263–271, 1982.

[13] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[14] A. Inselberg. Visualizing multi-dimensional structure using parallel coordinates. In *American Statistical Association Proceedings of the Section on Statistical Graphics*, 1989.

[15] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.

[16] J. I. Marden. *Analyzing and modeling rank data*. CRC Press, 1996.

[17] P. McCullagh. Permutations and regression models. In *Probability Models and Statistical Analysis for Ranking Data*, 1993.

[18] G. L. Thompson. Generalized permutation polytopes and exploratory graphical methods for ranked data. *The Annals of Statistics*, 21(3):1401–1430, 1993.

[19] A. Ukkonen. Visualizing sets of partial rankings. In *Advances in Intelligent Data Analysis VII*, 2007.

[20] M. P. Wand and M. C. Jones. *Kernel Smoothing*. CRC Press, 1995.

[21] V. A. Yemelichev, M. M. Kovalev, and M. K. Kravtsov. *Polytopes, Graphs, and Optimisation*. Cambridge University Press, 1984.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 4, 2009 at 10:42 from IEEE Xplore. Restrictions apply.