Using Entropy-Related Measures in Categorical Data Visualization

Jamal Alsakran* The University of Jordan
Xiaoke Huang† Kent State University
Ye Zhao‡ Kent State University
Jing Yang§ UNC Charlotte
Karl Fast ¶ Kent State University

ABSTRACT

A wide variety of real-world applications generate massive high dimensional categorical datasets. These datasets contain categorical variables whose values comprise a set of discrete categories. Visually exploring these datasets for insights is of great interest and importance. However, their discrete nature often confounds the direct application of existing multidimensional visualization techniques. We use measures of entropy, mutual information, and joint entropy as a means of harnessing this discreteness to generate more effective visualizations. We conduct user studies to assess the benefits in visual knowledge discovery.

Keywords: Categorical data visualization, Dimension Management, Entropy, Mutual Information, Parallel Sets

1 INTRODUCTION

Categorical datasets include survey results in health/social studies, bank transactions, online shopping records, and taxonomy classifications. Such data usually contain a series of categorical variables (i.e., dimensions) whose values comprise a set of discrete categories, such as transaction types, county/town names, product codes, species characters, etc. High dimensional categorical data impose significant challenges for information visualization due to their unique data discreteness. Although major advances have been made on high dimensional data visualization, many successful visualization methods are often undermined when directly applied to categorical datasets.

Categorical multivariate visualization faces two major challenges: (1) the limited number of categories creates overlapping elements and visual clutter. (2) the lack of an inherent order (in contrast to numeric variables) of discrete categories confounds the visualization design. Therefore, good metrics are needed in managing high dimensional categorical data that can reduce clutter, optimize information disclosure, and provide navigation guide as users explore categorical data features. However, many statistic measures used in previous works cannot be easily applied to categorical data. For example, data correlation needs to compute deviation and covariance from the average, which is not defined over categorical variables.

As suggested in [6], entropy can be employed in quantifying data features for better visualization. However, that work does not explicitly study the use of entropy in categorical visualization. Moreover the use of entropy-derived measures, such as mutual information, has not been well investigated. In this paper, we extensively study the capacity of using the entropy-related measures in visualizing multidimensional categorical data.

The concept of entropy [9] is developed over the probabilistic explanation of data distribution. Essentially, entropy quantifies the variation, or diversity, of a *discrete* variable. Derivative measures address relationships over multiple variables such as joint entropy that describes associated uncertainty and mutual information that illustrates mutual dependence. These measures are naturally defined over categorical data. They can be computed straightforwardly and efficiently, leading to good tools in manipulating categorical visualization.

Our contribution in the paper is multi-fold: First, we show how entropy-related measures can help users understand and navigate categorical data. In particular, we show probability distribution of categories over word clouds to reflect data features. We color 2D projects according to their joint entropy and mutual information to indicate significant pairwise dimension relationship over scatter plot matrices [7]. Second, we employ these measures in managing and ordering dimensions within the parallel set visualization of categorical data. In particular, we perform a global optimization of the sum of mutual information over all pairs of dimensions. We also present a sorting method that uses joint entropy to find an optimal sequence of categories over neighboring axes. This reduces visual clutter and helps users discover new insights. Third, we conducted user studies on real-world data to evaluate multiple parallel sets layouts using different ordering methods. The results indicate that there are benefits of using entropy-related measures in parallel sets optimization. Statistical tests showed that the user study results were significant.

2 RELATED WORK

Categorical data visualization: Categorical data visualization has been addressed in many previous approaches. Sieve Diagram [28] visualizes two dimensional categorical data using a two-way contingency table of frequency values between all category pairs. Mosaic Display [15] extends Sieve Diagram to three-way and four-way contingency tables by representing categories as tiles whose area is proportional to frequency. Tablelens [27] combines graphical and symbolic visualizations inside a tabular view of multidimensional data which may include categorical variables. Contingency Wheel [2] presents an interactive visualization technique for analyzing associations in large 2-way contingency tables. However, these methods visualize datasets with only a few categorical dimensions. To overcome this limitation, CateRank [13] ranks the relationship between pairwise dimensions and uses the score to guide the selection of two variables. The ranking is achieved by variance, crossover, and uniformity of distribution. A rank-by-feature framework [31] helps find important 1D or 2D relationships. Entropy is used for 1D uniformity test and for 2D grid cell ranking. Even so this approach is not designed for categorical data. In comparison to these approaches, our work investigates broader use of entropy measures to manipulate visual layouts and guide navigation in multiple visualizations of categorical data including scatter plots, parallel sets, etc. Our work is mostly focused on joint entropy and mutual information in multidimensional categorical visualizations, which has not been well studied.

Parallel sets plot [20] improves classic parallel coordinates [18] for categorical data by substituting polylines for frequency based ribbons across dimensions. It greatly reduces clutter and helps in understanding categorical data. We adopt this visualization and enhance it with entropy-related information. Another approach [30] uses correspondence analysis to define the distance between categorical variables, which is then used to assign order and space in classic parallel coordinates visualization. Instead, we adopt the information theory measures related to variable diversity, joint distribution, and mutual relation in visual representations and visual

[*] e-mail: j.alsakran@ju.edu.jo
[†] e-mail: hxiaoke@cs.kent.edu
[‡] corresponding author, e-mail: zhao@cs.kent.edu
[§] e-mail: Jing.Yang@uncc.edu
[¶] e-mail: kfast@kent.edu

analysis.

Dimension management: Our approach supports visual discovery with ordering, spacing, and filtering by integrating entropy related measures into visualizations. Dimension management is important for order-sensitive multidimensional visualization techniques [3]. Clutter reduction has been studied through dimension reordering in multidimensional visualization techniques, including parallel coordinates, star glyphs, and scatter plot matrix [26], where pertinent metrics, such as minimizing the outliers between neighboring parallel axes, are developed. Interactive ordering, spacing, and filtering has been proposed for high dimensional data based on dimension hierarchies derived from similarities among dimensions [33]. Reordering and filtering methods are also developed based on finding relevant subspaces for clustering [12]. These approaches are not designed in particular for categorical data and do not involve the information theory measures used in our approach.

Pargnostics [10] uses screen-space metrics for parallel coordinates. Together with line crossings, the pixel-based entropy is computed on the regions between a pair of coordinates for order optimization. They also use mutual information as a metric after applying bins to discretize numerical dimensions. Pargnostics is closely related to one important part of our study: parallel sets visualization of categorical data. However, the approach based on screen pixels is not directly applicable to categorical data. In contrast, our work uses entropy-related measures on original data records and applies optimizations over categorical parallel sets.

Some metrics have been proposed to steer data exploration during the knowledge discovery process. Johansson et al. [19] reduced dimensionality through user-defined quality measures. Albuquerque et al. [1] employed distance measures with an importance-aware sorting to reduce dimensions in scatter plot and parallel coordinates matrices. Recently, Lehmann et al. [22] presented an interactive framework to order plots in a scatter plot matrix. Their approach uses relevance visualization measures within a neighborhood and performs permutation to group relevant plots. Bertini et al. [4] conducted a systematic analysis of quality metrics to support the exploration of high-dimensional data. These approaches are not well suitable for discrete categorical data. Several experiments have also been conducted on assessing the tag cloud presentation techniques for various tasks [17, 29]. Our study of entropy-related measures presents a new metric, which facilitates the extension of these techniques to categorical data exploration.

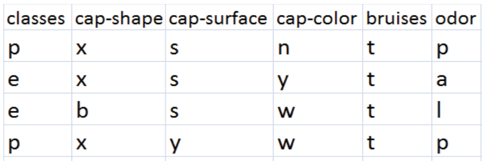
Chen et al. [6] conducted a study on how information theory can explain events and phenomena in information visualization, with the goal of establishing a theoretic framework. However, the discussion does not consider categorical visualization characteristics, which we deal with here.

3 Categorical Data and Measures

3.1 Data Features

Categorical datasets contain multiple variables (i.e. dimensions) whose values comprise a set of discrete categories. Each data record belongs to one of the categories in the corresponding dimension. We use two categorical datasets from the UC Irvine Machine Learning Repository [14] in our study. The first dataset contains 435 data items on congressional voting for congressmen in the U.S. House of Representatives. It includes votes for each of the congressmen on 16 key votes, and 1 classification variable of their party (Democrat or Republican). Each dimension refers to one voting record of a bill (e.g. education-spending) with three possible categories: y (yea), n (nay), or u (unknown).

The second dataset describes the physical characteristics of various mushroom species. The mushroom dataset includes 8,124 records and 23 categorical dimensions. The category values of each dimension are given in single indexing letters. Fig. 1 displays part of its categorical data table. The tabular view does not promote direct sense-making for analyzers. Users have to check their meanings in associated text files outside the table. To support sense-making activities, a visualization must orchestrate high dimensionality and tackle categorical discreteness. Entropy-related measures can provide important clues of data facts and effective instruments towards good designs in visualization and analytics.



classes	cap-shape	cap-surface	cap-color	bruises	odor
p	x	s	n	t	p
e	x	s	y	t	a
e	b	s	w	t	l
p	x	y	w	t	p

Figure 1: *Mushroom data table shown in part.*

3.2 Entropy-related Measures

Probability Distribution Considering a categorical variable as a discrete random variable $\mathbf{X}$, the probability mass function $p(x)$ defines the probability that $\mathbf{X}$ is exactly equal to a categorical value, x. The functions for all possible categorical values thus define the frequency distribution of all the data records within this dimension:

$$p(x) = \frac{count(x)}{count(\mathbf{X})}, \tag{1}$$

where $count(x)$ is the number of records has the value x and $count(\mathbf{X})$ is the number of all records of $\mathbf{X}$.

Entropy The entropy is computed as

$$H(\mathbf{X}) = \sum_{x \in \mathbf{X}} p(x) \log p(x), \tag{2}$$

which provides a measure of uncertainty (in the context of information theory) of $\mathbf{X}$. It provides information about the variation, or diversity, of the information contained in one data dimension.

Joint Entropy Joint entropy is defined over two variables $\mathbf{X}$ and $\mathbf{Y}$ as

$$H(\mathbf{X},\mathbf{Y}) = -\sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}} p(x,y) \log p(x,y), \tag{3}$$

where x and y are particular values of $\mathbf{X}$ and $\mathbf{Y}$, respectively. $p(x,y)$ is the probability of these values occurring together. Joint entropy measures data diversity associated with two variables.

Mutual Information Mutual information measures the reduction of uncertainty of one dimension due to the knowledge of another dimension. It defines a quantity that quantifies the mutual dependence of two random variables. The mutual information is defined as

$$I(\mathbf{X};\mathbf{Y}) = \sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

The highest mutual information indicates that if we know those data records having a category value in D_1, we will have more knowledge about their values in D_2. The lowest mutual information implies that the data records belonging to a category value in D_1 can be of any values in D_2.

3.3 Discussion

These measures are good metrics for representing categorical data features. They are naturally defined over discrete sets of categories, while many measures used in visualizations are not directly applicable to categorical variables. For example, the Pearson's correlation (e.g. used in scatter plots and parallel coordinates [23]) is computed from data deviation and covariance. Such computation relies on data average which however cannot be directly computed for categorical variables. We will study more advanced statistic measures to be used in categorical visualizations. For example, the covariance can be generalized for categorical data through Gini's

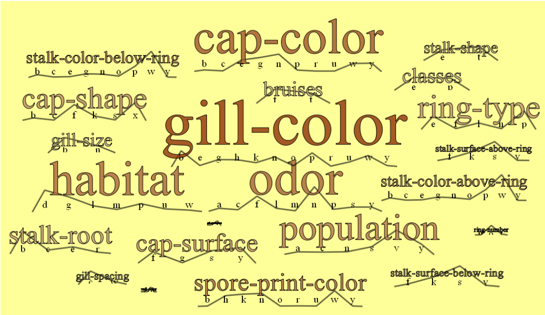


Figure 2: *Using word cloud (SparkClouds) to depict categorical data feature of the mushroom data. For each dimension, the font size reflects the number of categories and the color represents the entropy of the dimension. The graph line of each word shows the data frequency of each category.*

definition [25]. Moreover, Lambda is a measure of association suitable to show the strength of correlation (usually referred as association when dealing with nominal data) [16]. Similar to mutual information, it quantifies the proportional reduction of error when predicting the values of one variable from the known values of the other.

4 Categorical Data Visualizations

In this section, we delineate the use of these measures to develop effective visualizations for large categorical datasets.

4.1 Navigation Guide for Categorical Data

Visualizing the frequency function, $p(x)$, of data records distributed over categories, and the entropy, $H(\mathbf{X})$, of dimensional uncertainty could give users an intuitive understanding of categorical datasets. High entropies reveal which dimensions have a highly diversified distribution of data records. Inversely, small entropies indicate which dimensions have more concentrated distributions. This information helps users rapidly identify critical characteristics of the categorical data they are exploring. Next we show two examples.

Fig. 2 shows a word cloud of mushroom dimension names. We use the SparkClouds representation to show information over words [21]. The font size shows the number of categories in the dimension. The font color reveals the entropy, where high entropy values are mapped to highly saturated colors. Here we follow Brewer's approach [5] of color design where saturation is used to show ordered difference within the same hue (also applied to scatter plot matrices below). Other color schemes can also be adopted. It is easy to recognize that "gill-color" has the largest entropy value, i.e., the color of the platelike structures on the underside of the mushroom cap has the most diverse distribution. Moreover, the categorical values are shown under the word, such as the color letters "b, e, g, h, $\cdots$" under "gill-color". The frequencies of data records distributed over those categories are drawn as a graph line and overlaid on the lower part of the word. We can see that mushroom gills have more "b" buff and "p" pink colors.

Displaying the relationship between a pair of dimensions is a primary visualization task. Scatter plots are a common example. For a high-dimensional dataset, a matrix of scatter plots over all pairs of variables reveals pairwise relations. Finding the important pairs among all pairs (theoretically $C(N,2) = \frac{N!}{2(N-2)!}$ pairs for N dimensions) usually requires careful investigation from users. Furthermore, categorical dimensions with discrete values tend to increase clutter due to excessive point overlaps. Joint entropy and mutual information can be used to find potentially salient pairs of dimensions for categorical data.

Joint entropy matrix: The pairwise joint entropies are visualized as salient cues by colors from low saturation to high saturation over the matrix as shown in Fig. 3(a). High joint entropy indicates diversely distributed data records in a scatter plot, and low joint entropy reveals that the records have lots of overlaps. In Fig. 3(a-1), we show the scatter plot linked to the highest joint entropy among all dimensional pairs. The particular pair refers to the two dimensions "cap-color" and "gill-color" and the plot shows diverse point distribution. Fig. 3(a-2) displays the plot between the dimensions "gill-attachment" and "veil-type" which has the lowest joint entropy. It shows that all mushrooms in the dataset have partial veil and most are gill-attached.

Mutual information matrix: Fig. 3(b) further utilizes the mutual information for visualization enhancement. Mutual information defines the reduction of uncertainty of one dimension due to the knowledge of another dimension. The highest mutual information is given a saturated color and the lowest is assigned with less saturation. Users can analyze one variable by comparing the row on its right and the column above it, so as to find dependency information between the dimension and others.

In Fig. 3(b), the first row is the mutual information between mushroom "class" (edible or poisonous) and any other dimension. Fig. 3(b-1) shows the scatter plot with the highest mutual information. It is between "class" and "odor". By clicking to view the scatter plots of dimension pairs with saliency, analyzers can easily find that poisonous mushrooms mostly have odors like creosote, foul, pungent, spicy and fishy. In contrast, most edible mushrooms have no odor, though some smell like almond or anise. Fig. 3(b-2) is the scatter plot with the lowest mutual information. It shows that whether a mushroom class is edible or poisonous it has no obvious relation with veil-type.

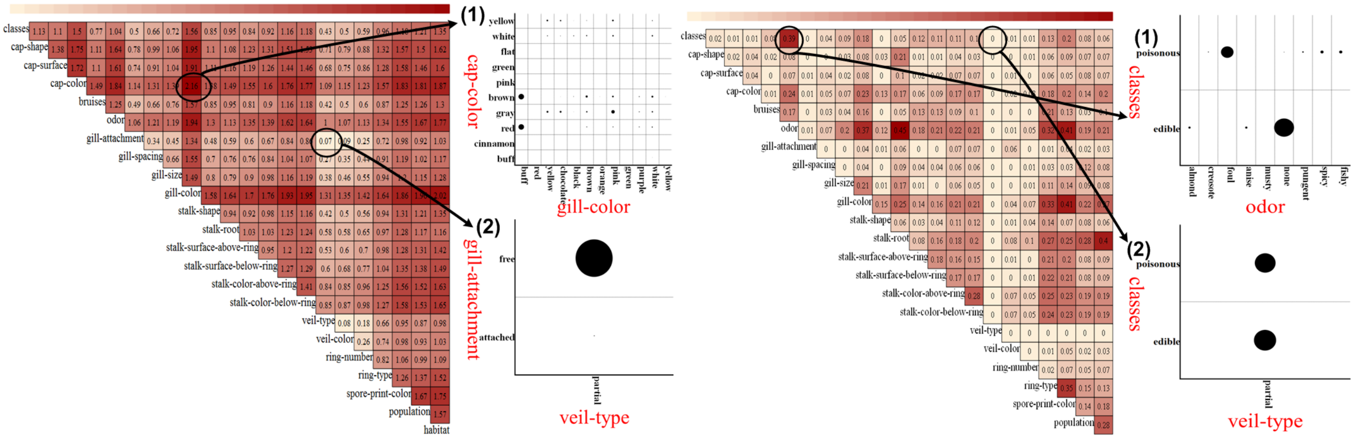
4.2 Dimension Management on Parallel Sets

Parallel coordinates plots have been widely used in visualizing multivariate data and perceiving their patterns. However, visual clutter in these plots increases along with data size. More data means it is more difficult for users to perceive patterns in the visualization. The situation deteriorates even further when many categorical dimensions are involved since a large number of polylines converge on individual categories. Parallel sets [20] is a parallel coordinates technique designed for categorical data. In this section, we use the entropy related measures to show how they can help users manage dimension spacing, ordering and filtering. Such application also leads to more tractable interaction.

Fig. 4(b) displays a parallel sets visualization of the mushroom dataset over 23 dimensions. The entropy values are also shown in Fig. 4(a). The order of coordinates is drawn as the reading order from the downloaded dataset. Between adjacent coordinates, those data records having two categories, x and y, are visualized as a ribbon. The ribbon connects two bars representing x and y over the coordinates respectively. The sequence of categories over each dimensional axis is plotted in alphabetical order. The category indexing letters are shown in the bars. Users can find their meanings from word clouds, such as those shown in Fig. 2. The ribbon width (size) is set proportional to the joint probability distribution $p(x,y)$ of the two categories. Here the ribbon colors are defined according to the leftmost dimension "class": green for edible and blue for poisonous. This allows users to quickly identify whether a mushroom is edible or poisonous given one characteristic (i.e., categories over data dimensions). Next, we show how the visualization can be improved by entropy related measures.

4.2.1 Sorting categories over coordinates

For numerical dimensions, the line endpoints are drawn on the corresponding axes from bottom to top with an incremental/decremental order of their numeric values. For categorical dimensions, the reading order, or alphabetical order, are usually used. Multiple ribbons starting from one category on the left axis are



(a) Joint entropy matrix and scatter plots

(b) Mutual information matrix and scatter plots

Figure 3: *Using joint entropy and mutual information in navigating scatter plot matrix.*

drawn with respect to the corresponding categories on the right from top to bottom, so as to reduce possible self-intersections. The empirical reading/alphabetical method can be improved by taking into account the relationship between neighbor axes. An optimal sequence of categories over axes can lead to clearer results. A very early work by Ma et al. [24] presents a method by minimizing order conflict between pre-clustered category values, which is achieved by a greedy algorithm. This solution is designed only for optimizing one pair of dimensions simultaneously and requires clustering. It is not easily extended to our problem, which aims to sort categories on all coordinates. .

Here we seek an optimization solution of sorting categories by utilizing joint probability distribution, $p(x,y)$, which is used in computing joint entropy in Eq. 3. It indeed measures the co-occurrence possibility of data records having both categories x and y. In practice, between a pair of adjacent coordinates d_1 and d_2, we first assume that the left coordinate d_1 has drawn its categories in a fixed sequence. This sequence might have been determined by the sorting executed between d_1 and its left dimension. A category Y of d_2 is assigned to an optimal position with regard to the values of all pairs $p(x,Y)$ for any x of d_1. The optimal position is computed by a linearly interpolated location using $\sum_{x \in d_1} p(x,Y)L(x)$ where $L(x)$ is the vertical location of one category x over d_1 axis. When two categories in d_2 are located at the same position, we put the one with the larger $p(TOP_{d_1},Y)$ value above the another one. Here TOP_{d_1} is the category on the highest position of d_1 axis. Fig. 5 illustrates the change of category sequence between two dimensions. After applying the algorithm, Fig. 5(right) improves the result of Fig. 5(left) with less intersections.

Fig. 4(c) shows the reorganization result for the mushroom dataset. When compared with the original result in Fig. 4(b), it has less clutter between many coordinate pairs. For example, compare the red circled areas in Fig. 4(c) with the corresponding areas in Fig. 4(b). Here the order of parallel coordinates is not changed for comparison.

4.2.2 Assessing crossings of categorical visualization

Reducing line crossings has been used to improve parallel coordinates results [11, 10, 32]. One line crossing is counted as one interaction of lines. For a quantitative measure, we extend the crossing computation to ribbons in parallel sets of categorical data. Here we need to distinguish ribbons with different sizes since the intersection of two big ribbons is more prominent than two small ones. So the crossing count is no longer an integer but computed as $p(x,y) \cdot p(z,w)$, where $p(x,y)$ and $p(z,w)$ are the joint probability distributions defining the sizes of the two interacted ribbons.

As a result, Fig. 4(c) has the ribbon crossing as 0.78, which is less than the crossing of 1.22 from Fig. 4(b). This provides a quantitative justification of the optimal sequence of categories.

4.2.3 Arranging space between neighboring coordinates

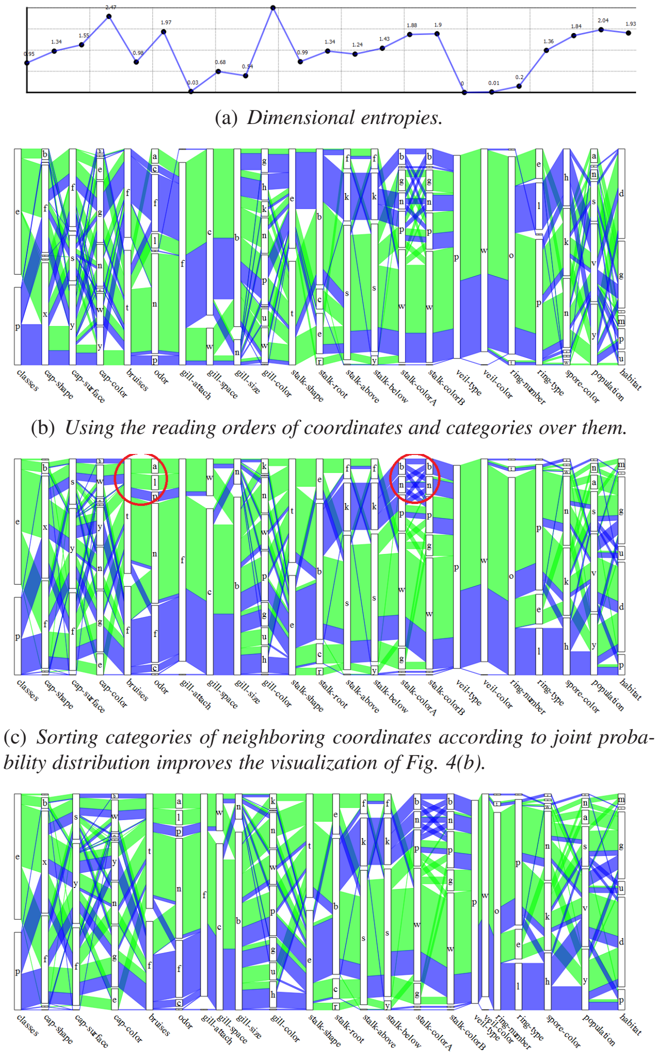
Giving more space to two adjacent coordinates with more interests can help users better perceive data relationships [33]. Fig. 4(c) can be further improved through an uneven distribution of space between coordinates. High joint entropy means the diversity of data records between the coordinate pair is high, which will usually produce a complex layout of lines. Thus without changing the coordinate order, we can compute the joint entropies of all neighboring pairs and use them to arrange the horizontal spacing. Fig. 4(d) shows the result after adjusting the spacing based on the result of Fig. 4(c). Many neighborhoods, such as the space between the two rightmost coordinates, are enlarged to depict their pairwise relations more clearly. The examples in this paper compute horizontal spaces linearly proportional to the joint entropies. This computation can also be implemented by other algorithms.

4.2.4 Dimension filtering with user interaction

Entropy-related measures can be very helpful in the visual sense making process. Fig. 6 shows a filtered visualization after users set an entropy threshold of 1.0 to remove low entropy dimensions. This provides more space to analyze dimensions with complex relations. Moreover, users can use their domain knowledge, along with the information depicted by the aforementioned entropy curve, word cloud, and treemap visualizations, to select, remove, and exchange dimensions.

4.2.5 Optimal ordering of multiple coordinates

A long-standing problem of multivariate visualization techniques is how to arrange the order of dimensions in visual layouts. A good example is the order of axes in parallel coordinates plots, which determines what information is directly manifested between neighboring coordinates. While users might be given the capability to adjust the order interactively, the initial order inarguably plays a critical role in understanding data and guiding user exploration. Moreover, users may only interactively adjust the order of a few dimensions for a large amount of dimensions. For example, it is unlikely that users would manually adjust the order of all 23 dimensions in the mushroom dataset. Consequently, there remains a need for automatic order arrangement for the dimensions that are not adjusted. In many cases, the success of knowledge discovery will rely, in large part, on this ordering. Entropy-related measures provide an important metric, contributing to the orchestration of dimensions for optimal categorical visualization.



(a) *Dimensional entropies.*

(b) *Using the reading orders of coordinates and categories over them.*

(c) *Sorting categories of neighboring coordinates according to joint probability distribution improves the visualization of Fig. 4(b).*

(d) *Arranging space between neighboring coordinates with joint entropy improves the visualization of Fig. 4(c).*

Figure 4: *Visualizations of the mushroom dataset with parallel sets.*

A variety of approaches have proposed empirical or optimal ordering algorithms to provide clear visualization layouts. For example, reducing line crossings between neighboring coordinates to reduce visual clutter. However, most methods are not designed for the unique characteristics of categorical data. For instance, a graph-theory based method [32] achieves an optimal order of parallel coordinates by extending the ordering problem to a complete graph in metric space, whose solution is associated to the metric Hamiltonian path of the Traveling Salesman Problem. The cost values between every pair of dimensions are used in the optimization algorithm for global cost minimization. In this approach [32], the "cost" is defined as the number of line crossings. Reducing line crossings creates less tangled layouts. However, reducing crossings does not necessarily lead to more effective insight discovery.

In our study, mutual information is employed as the "cost" function in the global optimization algorithm. For two axes, given a categorical dimension (X) on the left, selecting a right dimension (Y_l) with larger mutual information can accommodate more definite relationships of data records than using a dimension (Y_s) with smaller mutual information. Thus, with $I(X;Y_l) > I(X;Y_s)$, the two-coordinates visualization leads to more insights of data dependency over categories for user's investigation. Globally maximizing the sum of mutual information of a series of dimensions tends to create visualization results that promote better knowledge discovery. We modify the minimization optimization [32] to a maximization solution of the sum of mutual information with respect to all different orders of categorical dimensions. This implementation uses Hamiltonian path algorithms and creates optimal ordering.

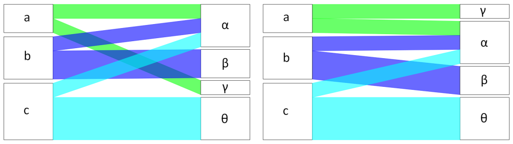


Figure 5: *Sorting categories between two categorical dimensions*

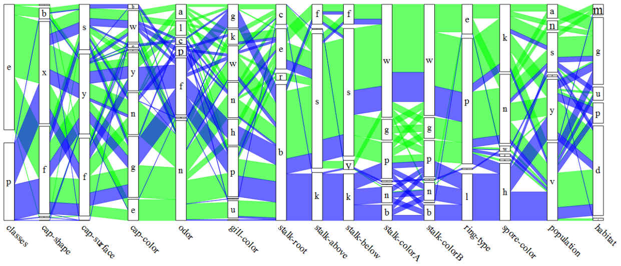


Figure 6: *Visualization of the mushroom data filtered by setting an entropy threshold of 1.0 to remove low diversity dimensions.*

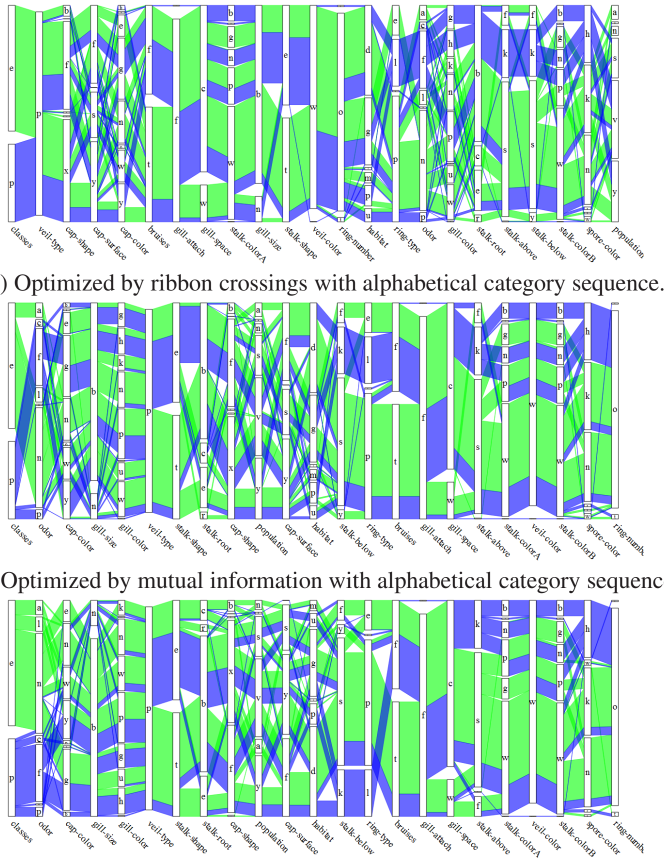
Mutual information based ordering has an explicit benefit: the cost is not related to the sequences of categories over axes. In contrast, the computation of ribbon crossings relies on such sequences, which does not have a predisposed solution for categorical data. In other words, the crossing based optimization is not as definite as the mutual information based one. This makes the latter method more appropriate for categorical data visualization.

Visualization results with different ordering methods of the mushroom dataset are shown in Fig. 7. They improve the original visualization of Fig. 4(b) which uses the reading order of dimensions and the alphabetical sequence of categories on axes. Fig. 7(a) is optimized by ribbon crossings while using the alphabetical category sequence. Fig. 7(b) is optimized by mutual information while using the alphabetical category sequence. Fig. 7(c) is optimized by mutual information and also sorts the sequence of categories on axes as described in Sec. 4.2.1.

We also perform the order optimization over the congressional voting dataset. Fig. 8(a) shows the visualization of the congressional voting dataset using the reading order of dimensions and the alphabetical sequence of categories. Then the optimized ordering is applied by using mutual information for dimension and sorting categories on axes, as shown in Fig. 8(b). Classification of the congressmen is the leftmost dimension. Green represents Democrats and red represents Republicans. Our system supports users to interactively select the observed dimension. Fig. 9 shows the results of selecting one voting dimension, education-spending, as the leftmost dimension. Fig. 9(a) and (b) are results of using reading order and after optimization, respectively. Here, blue represents voting for nay, red represents voting for yea, and gray is for unknown. Users can visually identify the votes of congressmen for other bills, when they vote for education spending with yea, nay and unknown. It is shown in both Fig. 8 and Fig. 9 that the optimization separates green and red ribbons automatically, which reduces the clutter and helps users to quickly identify insights. In next section we present user studies to support the benefit of using the optimized visualization.

5 User Studies

We conducted user studies of using entropy-related measures for categorical data visualization. More specifically, the studies were on the ordering of parallel coordinates, which has long been a key problem in multidimensional visualizations, especially when cate-



(a) Optimized by ribbon crossings with alphabetical category sequence.

(b) Optimized by mutual information with alphabetical category sequence.

(c) Optimized by mutual information with optimized category sequence.

Figure 7: *Visualizations with different ordering methods of multiple coordinates, improving the original visualization of Fig. 4(b) which uses the reading order of dimensions and the alphabetical sequence of categories on axes.*

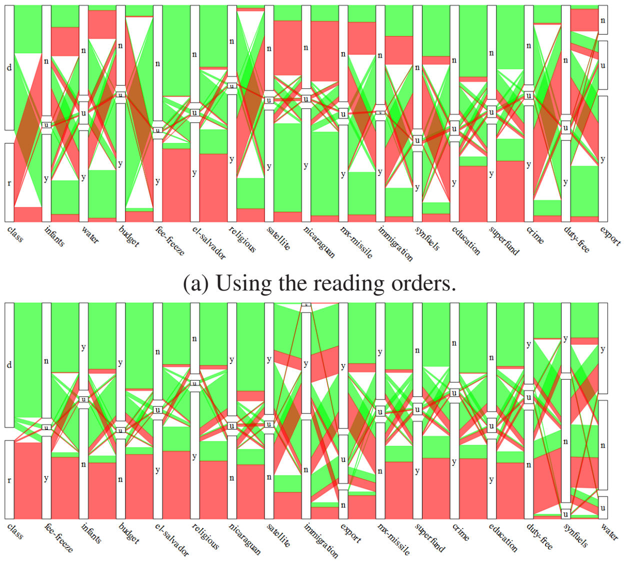
gorical data are involved.

5.1 User Study of the Mushroom Dataset

The user study assessed user performance on insight discovery with four parallel sets visualizations of the mushroom data, each using a distinct ordering approach:

- **C1**: original reading order, with alphabetical sequence of categories over axes, as shown in Fig. 4(b);
- **C2**: ribbon-crossings minimized order, with alphabetical sequence of categories over axes, as shown in Fig. 7(a);
- **C3**: mutual information maximized order, with alphabetical sequence of categories over axes, as shown in Fig. 7(b);
- **C4**: mutual information maximized order, with optimized sequence of categories on coordinates, as shown in Fig. 7(c).

Tasks: We designed four tasks based on a question people are likely to have about mushrooms: how do different mushroom characteristics, such as odor or gill color, relate to mushroom edibility? In other words, do certain characteristics indicate that a mushroom is likely to be edible or poisonous? We studied how the four visualization cases can help users discover such insights. Users were asked to use the visualizations to find characteristics which are: (**T1**) All-edible; (**T2**) All-poisonous; (**T3**) Mostly-edible; (**T4**) Mostly-poisonous. For example, if the data records of category "red" in the dimension "cap-color" are all "poisonous" in dimension "class", this means that a red cap-color is an all-poisonous characteristic. Here mostly-edible means that for a given characteristic, the number of edible mushrooms is greater than the number of poisonous ones. Mostly-poisonous is defined in the same manner.



(a) Using the reading orders.

(b) After optimized ordering.

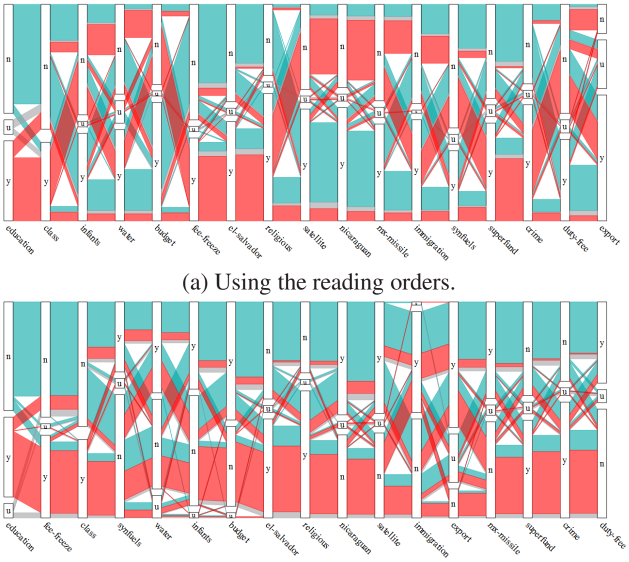
Figure 8: *Visualizations of the congressional voting dataset. The optimization is applied by using mutual information for dimension ordering and sorting categories over axes. Classification of the congressmen is the leftmost dimension where green represents Democracy and reed represents Republican.*

Each participant was given 90 seconds to find characteristics belonging to one of the tasks T1 to T4, based on a given visualization among case C1 to C4. We gave a score for each task of each participant. Once a participant found one mushroom characteristic belonging to the current task, he/she wrote down the index letter of this character. If the finding was correct compared to the ground truth, the participant's score of this task was increased by one. Note that each participant worked on visualization cases C1 to C4 in a random order, so as to remove possible bias created by task sequence.

Participants and training: The participants were mostly graduate students in the computer science department with two professor participants. We had a total of 11 participants (2 female and 9 male) aged from twenties to fifties. All reported spending more than 20 hours per week using computers. Most of them had a basic understanding of statistics and information visualization, but they did not have previous experience with, or knowledge of, parallel sets visualizations and the proposed research work. We spent about 15 to 20 minutes to brief the participants about the user study. This involved explaining the tasks, introducing the mushroom data, describing the parallel set visualization results, and the meaning of T1 to T4. Prior to starting the study, each participant was given 10 minutes to practice with similar visualization results from different datasets. This was to familiarize them with the procedure. After this introductory briefing and practice session, participants reported confidence to conduct the tasks.

Results: The ground truth of the mushroom data is directly computed from the dataset (not from visualizations): (T1) All-edible, 9 characteristics; (T2) All-poisonous, 25 characteristics; (T3) Mostly-edible, 51 characteristics; (T4) Mostly-poisonous, 13 characteristics. These are the maximum number of characteristics people can find, which are directly computed from the dataset.

Fig. 10(a) illustrates the results of average scores achieved for T1 to T4, using visualizations C1 to C4, respectively. Each bar represents a percentage of the average score of each task from all participants over the corresponding ground truth value. The variance is shown on the bar as a line segment. Given the limited time, users achieved the highest scores in all tasks when using the visualization C4 (shown in purples bars), which employs our approach of coordi-



(a) Using the reading orders.

(b) After optimized ordering.

Figure 9: *Visualizations of the congressional voting dataset. The optimization is applied by using mutual information for dimension ordering and sorting categories over axes. The leftmost dimension is the votes of education-spending. Here, blue represents voting for nay, red represents voting for yea, and gray is for unknown.*

nate ordering and category sequence optimization. Without optimal category sorting on axes, mutual information based ordering (green bars) also achieved good results. These results indicate that by using entropy-based measures users can identify more insights in a limited time. In contrast, the crossing based optimization (red bars) does not provide better results in T2/T3 than original visualization (blue bars).

We further computed the average performance of C1 to C4 over the four tasks. In Fig. 10(b): using C4, users found approximately 47% of all answers on average, while the best performance reached more than 60%. C3 provided the second best average performance. On average, C2 provided no difference from the original visualization C1. Note that these measures were acquired when users were given the limited 90 seconds for each task. Participants made errors when they incorrectly identified the characteristics for the given tasks. These errors are an important indicator of the visualization's quality. In Fig. 10(c), it clearly shows C3 and C4 reduced users error rate to around 11% and 10.5% in all their reports, compared with 16% and 15% in C1 and C2.

Statistical test: Since the collected user data did not form a normal distribution for most cases, we applied the Friedman test of variance by ranks, a non-parametric statistical test [8], to determine statistical significance among the visualizations C1-C4. Statistical significant differences are discovered between C1 and C4 (p-value = 0.011), between C2 and C4 (p-value = 0.035), as well as between C3 and c4 (p-value = 0.007). Such results indicate that, from the user study, the optimized visualization C4 has a significant effect at the level of p-value < 0.05.

5.2 User Study of the Voting dataset

This user study assessed user performance on the congressional voting dataset with the four visualizations:

Fig. 8: Classification as the leftmost dimension:

- C1: original reading order, with alphabetical sequence of categories over axes, as shown in Fig. 8(a).
- C2: mutual information maximized order, with optimized sequence of categories on axes, as shown in Fig. 8(b).

Fig. 9: Eduction-spending vote as the leftmost dimension:

- C3: original reading order, with alphabetical sequence of categories over axes, as shown in Fig. 9(a).
- C4: mutual information maximized order, with optimized sequence of categories on axes, as shown in Fig. 9(b).

The two comparable groups of visualizations, (1) C1 and C2 and (2) C3 and C4, are studied to find the benefits of our optimization approach.

Task: With the visualizations of C1 and C2, we designed tasks to determine how Democrat and Republican congressmen voted for different bills, such as immigration and export. For each dimension (i.e. each bill), the participants were asked to identify: **(T1)** which party vote more for yea? **(T2)** which party vote more for nay? For both tasks, the participants chose from three options: Democrat, Republican, or hard to identify.

With the visualizations of C3 and C4, for each dimension, the participants were asked to identify: **(T3)** which congressmen group vote more for yea? **(T4)** which congressmen group vote more for nay? For both tasks, the participants chose from three options: those voting yea for education bill, those voting nay for education bill, or hard to identify. In our study, each participant was given 120 seconds to work on tasks **(T1)** to **(T4)**. As before, the task sequence was randomized.

Participants and training: We recruited 35 participants (5 female and 30 male) aged from 22 to 30. They are all graduate students in computer science having little experience with information visualization. We gave an introductory briefing and allowed them to practice for 5 minutes on similar visualizations other than the experiment cases.

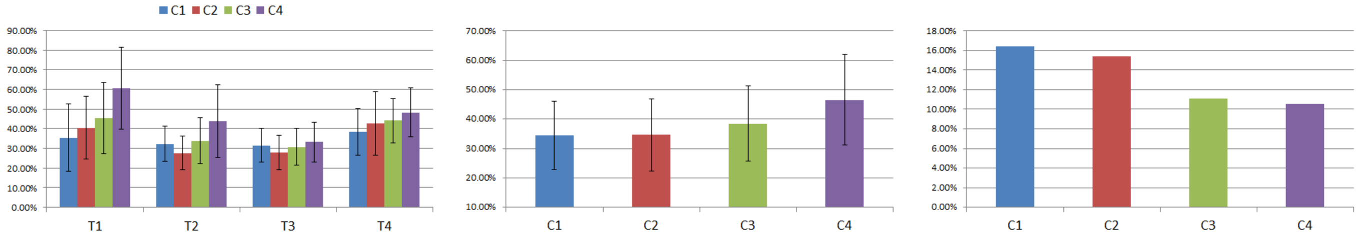
Results: We computed the ground truth of the tasks directly from the dataset. Then we graded each participant by giving 1 point if the answer was correct, -1 point if the answer was incorrect, and 0 points if they said it was hard to identify. The average score of using C1 is 11.5, while the average score of using C2 with optimization was 20.1. When education was on the left, the average score of using C3 was 13.2, and the average score of using C4 with optimization was 18.0. The results show that the participants found more correct information within 120 seconds by using our optimization approach.

Statistical test: One-way analysis of variance (ANOVA) [8] was conducted to compare the effect of using different visualizations for insight discovery. One test was performed with the user scores of C1 and C2. In the ANOVA F-test, we have F = 17.1 and p-value = 0.0001. Another test is conducted based on the user scores of C3 and C4. Here the ANOVA F-test has F = 5.64 and p-value = 0.02. Such results show that there was a significant effect of using the optimization method at the level of p-value < 0.05. This indicates the significance of our user study results.

6 Conclusion

Categorical data visualization is particularly challenging due to the discrete nature of the data. In this work, we utilize measures from information theory to enhance the visualization of high dimensional categorical data and to support exploratory visual reasoning. In particular, entropy, joint entropy and mutual information are employed for categorical visualization enhancement. In the future, we will study more measures from information theory in visualization enhancement. We will also extend the use of information measures to numeric and hybrid datasets.

The measures can support users to browse data facts among dimensions, to determine starting points of data analysis, and to test-and-tune parameters for visual reasoning. In the paper, we briefly introduce possible user interactions to find interesting dimensions through word clouds, to find 2D projects through scatter plots, and to change ordering, spacing, and filtering in parallel sets. Further investigation and user studies of such interactive approaches are needed and will be an important topic in our future work.



(a) (b) (c)

Figure 10: *User study results of the mushroom dataset. (a) Average percentage of user findings over ground truth on each task, attained with different visual representations. (b) Total performance of user findings using different visualizations. (c) Total error rate of user findings using different visualizations.*

Acknowledgements

This work is partially supported by NSF IIS-1352927, IIS-1352893 and Google Faculty Research Award.

References

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Quality-based visualization matrices. In *Proc. Vision, Modeling and Visualization (VMV) 2009*, pages 341–349, Braunschweig, Germany, Nov. 2009.

[2] B. Alsallakh, E. Gröller, S. Miksch, and M. Suntinger. Contingency wheel: Visual analysis of large contingency tables. In *Proceedings of the International Workshop on Visual Analytics*, 2011.

[3] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 52–60, 1998.

[4] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212.

[5] C. A. Brewer. In J. Bares, editor, *Color hard copy and graphic arts III, Proceedings of SPIE*, volume 2171, pages 54–63, Feb. 1994.

[6] M. Chen and H. Janicke. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, 2010.

[7] W. S. Cleveland and M. E. McGill. *Dynamic Graphics for Statistics*. Statistics/Probability Series. Wadsworth & Brooks/Cole, 1988.

[8] W. Conover. *Practical nonparametric statistics*. Wiley, 3th edition, 1999.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[10] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1017–1026, Nov. 2010.

[11] G. Ellis and A. Dix. Enabling automatic clutter reduction in parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):717–724, Sept. 2006.

[12] B. J. Ferdosi and J. B. T. M. Roerdink. Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Comput. Graph. Forum*, 30(3):1121–1130, 2011.

[13] D. Filippova and B. Shneiderman. Interactive exploration of multivariate categorical data: Exploiting ranking criteria to reveal patterns and outliers. *HCIL Technical Report*, 2009.

[14] A. Frank and A. Asuncion. UCI machine learning repository [http://archive.ics.uci.edu/ml]. *University of California, Irvine, School of Information and Computer Sciences*, 2010.

[15] M. Friendly. Visualizing categorical data: Data, stories and pictures. *SAS User Group International Conference Proceedings*, pages 190–200, 1992.

[16] G. D. Garson. *Measures of Association*. Statistical Associates Publishers, Asheboro, NC, 2012.

[17] M. J. Halvey and M. T. Keane. An assessment of tag presentation techniques. In *Proc. 16th International Conference on World Wide Web*, pages 1313–1314, 2007.

[18] A. Inselberg. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer-Verlag New York, 2009.

[19] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15:993–1000, 2009.

[20] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.

[21] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. SparkClouds: Visualizing trends in tag clouds. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1182–1189, 2010.

[22] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, and H. Theisel. Selecting coherent and relevant plots in large scatterplot matrices. *COMPUTER GRAPHICS forum*, 2012.

[23] J. Li, J.-B. Martens, and J. J. van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, Mar. 2010.

[24] S. Ma and J. L. Hellerstein. Ordering categorical data to improve visualization. *IEEE Information Visualization Symposium Late Breaking Hot Topics*, pages 15–18, 1999.

[25] T. Okada. A note on covariances for categorical data. In *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents*, pages 150–157. Springer-Verlag, 2000.

[26] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 89–96, 2004.

[27] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 318–322, 1994.

[28] H. Riedwyl and M. Schupbach. Graphische darstellung von kontingenztafeln. *Technical Report 12, Institute for Mathematical Statistics, University of Bern*, June 1983.

[29] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tag clouds. In *Proc. CHI'07*, pages 995–998, 2007.

[30] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, June 2004.

[31] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

[32] Y. Xiang, D. Fuhry, R. Jin, Y. Zhao, and K. Huang. Visualizing clusters in parallel coordinates for visual knowledge discovery. In P.-N. Tan, S. Chawla, C. Ho, and J. Bailey, editors, *PAKDD*, volume 7301 of *Lecture Notes in Computer Science*, pages 505–516. Springer Berlin / Heidelberg, 2012.

[33] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 105–112, 2003.