Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data

Robert Kosara, *Member, IEEE,* Fabian Bendix, and Helwig Hauser, *Member, IEEE*

***Abstract*— Categorical data dimensions appear in many real-world data sets, but few visualization methods exist that properly deal with them.**

***Parallel Sets* are a new method for the visualization and interactive exploration of categorical data that shows data frequencies instead of the individual data points. The method is based on the axis layout of parallel coordinates, with boxes representing the categories and parallelograms between the axes showing the relations between categories.**

In addition to the visual representation, we designed a rich set of interactions. Parallel Sets allow the user to interactively remap the data to new categorizations, and thus to consider more data dimensions during exploration and analysis than usually possible. At the same time, a meta-level, semantic representation of the data is built. Common procedures, like building the cross product of two or more dimensions, can be performed automatically, thus complementing the interactive visualization.

We demonstrate Parallel Sets by analyzing a large CRM data set, as well as investigating housing data of two US states.

***Index Terms*— Information Visualization, Interaction, Nominal Data, Categorical Data, Multivariate Data.**

I. Introduction

CATEGORICAL dimensions play a very important role in the analysis of many real-world data sets. Numerical attributes often can only be understood in the context of categorizations, and users working with data often examine different classes before even looking at numbers. While numerical dimensions are well understood in both statistics and visualization, the categorization of products, customers, etc. provides a special challenge for visualization.

Categorical dimensions are generally data dimensions that only contain a small number of different values, which often have special meanings. Categories usually do not have an inherent order (e.g., bank account types, ethnic groups), which means that the mapping to numerical values is arbitrary, and also the differences between these values are not meaningful.

Dimensions with many categories also are often organized hierarchically: customer surveys contain sections with related questions, split one piece of information between several questions (e.g., education), or ask the same question several times for cross-checking; bank

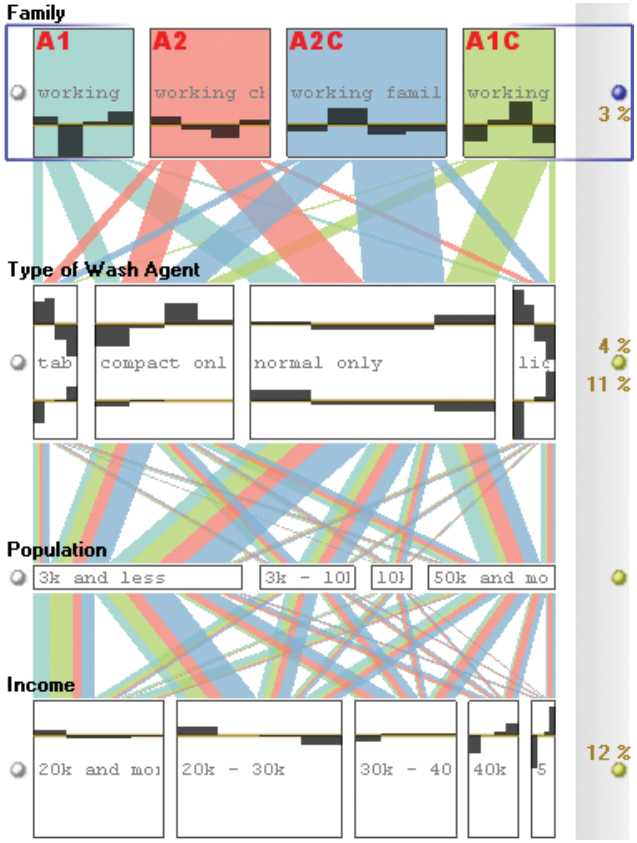

Fig. 1. Snapshot of an interactive visual analysis session using Parallel Sets. A large and complex CRM data set (Section IV) is analyzed and multi-dimensional relations within the data are revealed. The two meta-level categorizations shown in the top half of the visualization (family type and washing agent) are the result of an interactive process integrating multiple (simpler) data dimensions into new ones. The colored parallelograms and the histograms shown support the visual analysis of complex relations within this data, leading to insights such as the correlation between having no children, preferring liquid washing agents, and living in larger cities.

R. Kosara is with the University of North Carolina at Charlotte; He can be reached via email at rkosara@uncc.edu.

F. Bendix and H. Hauser are with the VRVis Research Center in Vienna, Austria; Email {Bendix, Hauser}@VRVis.at.

Manuscript received Nov 1, 2005; revised Dec 23, 2005; accepted Jan 30, 2006.

accounts are classified in several ways that will often involve hierarchical categorizations, etc. Using these hierarchies for visualization is extremely helpful for the user, because they provide a natural way of aggregating and abstracting data. The visualization application has to know about those hierarchies in order to make use of them, of course, requiring additional data about the data set, or *meta data*. Interaction is also required, because the user will want to switch back and forth between a detailed investigation and a more general overview by means of these hierarchies.

Most existing work has focused on the visualization of numerical data, treating categories as a special case with only a few values. The approach presented in this paper had to be radically different in order to accomodate the special properties of categorical data and large categorical data sets in practice.

An implicit assumption in many visualization systems is also that the user will perform a whole analysis in one, uninterrupted session, and will never return to the same kind of analysis or the same data set. Our experience has shown that this is not the case, however. Users often deal with similar data sets and similar tasks, which consequently require them to go through the same or similar sets of actions for each new data set. Also, the analysis of a typical real-world data set requires many sessions, potentially spread out over a long time period. The user needs to be able to save results to continue where he or she left off as seamlessly as possible.

We present a new approach to information visualization, called *Parallel Sets* [1] (Figure 1), which was developed specifically for categorical data. This paper presents additional features as well as a new case study to demonstrate the method. Parallel Sets support interactive visual exploration and analysis [2] by combining a new visual metaphor with an advanced interaction scheme and automated procedures. Parallel Sets adopt the advantages of two older and well-proven visualization techniques: the *flexible layout* of Parallel Coordinates [3] (Figure 2b), treating all dimensions as visually independent – in contrast to recursive space-subdivision approaches like Mosaic Displays –, and *displaying frequencies* as representatives for the categories (Figure 2c) – as opposed to the usual one-by-one items-based visualization of data.

The following sections present the related work and the idea of Parallel Sets in terms of the visual metaphor and interaction scheme. The workflow is explained, which is essential to performing interactive visual analysis of heterogeneous and high-dimensional data. It is important to stress the fact that the interaction scheme is

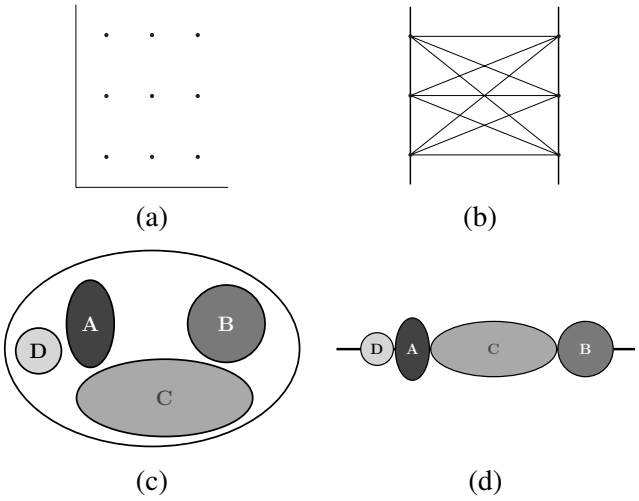

Fig. 2. Visualizing categorical data with traditional techniques such as a scatterplot (a) and parallel coordinates (b) do not yield useful visual representations, because usually only a small number of different values are given per categorical dimension. Traditional Venn diagrams (c, d) are based on the concept of showing data frequencies and therefore work well for categorical data. Lining categories up next to each other leads to the basic idea of Parallel Sets.

an integral part of Parallel Sets, and necessary to exploit the full potential of the approach. We demonstrate the use of Parallel Sets to reveal interesting information in a large customer relationship management (CRM) data set, as well as housing data of two US states.

II. Related Work

Parallel Coordinates [3] (Figure 3a) are a popular visualization technique, in which the graphical axes are not arranged orthogonally, but they are placed side by side. An n-dimensional data point is represented by a polyline, which intersects the parallel axes at points which represent the values of the individual data dimensions along the respective axes. This view is capable of displaying high-dimensional data (up to about 10-15 axes in practice), because the axes are visually independent of each other.

Initially, parallel coordinates were designed to display continuous variables [3], but recent approaches have tried to integrate categorical variables into this visualization as well. Rosario et al. [4] suggest transforming categories to numbers by techniques similar to Multiple Correspondence Analysis (MCA). By this, the space on each axis is used more efficiently, because the spacing becomes meaningful (similar categories are positioned close to each other). A simpler approach is proposed by Teoh and Ma [5]: for each category, an interval is constructed on the continuous axes to make more polylines visible. By this, the space is used to give the user an impression of how many data items are visualized. Using alpha

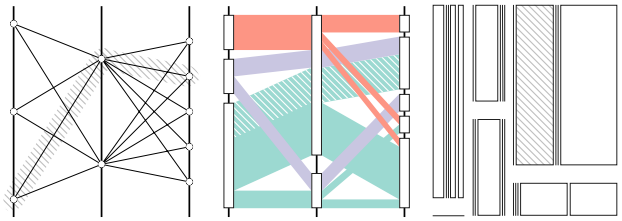

Fig. 3. Three different visualization techniques displaying the same data: (a) the categories are represented by points on continuous axes in parallel coordinates, (b) Parallel Sets show the frequencies of categories and relations, and (c) a Mosaic Display provides a compressed view of the data (the hatched parts represent the same subset).

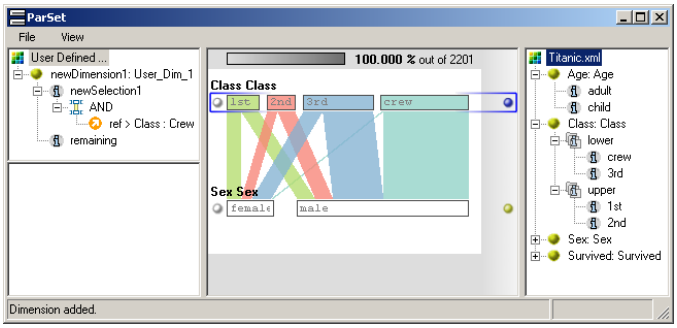

Fig. 4. The Parallel Sets prototype (showing the *Titanic* data set [15]) consists of four panels (clockwise from top left): the user panel (showing user-defined dimensions), the main visualization panel, the data panel (showing the source data), and the exclusion panel (for filtering).

blending, the Information Mural [6] also gives an impression of the number of values per category, but is still hard to read and imprecise for truly categorical data. One problem remains for all parallel coordinates techniques: the visualization implements a continuous design model, which does not match the discrete user model of the data. This discrepancy between the user's mental model and the presented image is eliminated by the use of frequency-based techniques: categories are represented by visual entities that are scaled according to their corresponding frequency.

There are several techniques that follow this approach: the Mosaic Display [7], [8] (Figure 3c) is a recursive space-subdivision technique (similar to Dimensional Stacking [9]), in which the frequency values of categories are represented by particular areas ("tiles") on the screen – interactive mosaic plots [10], [11] provide an even better approach for visual exploration, because they make use of the user's domain knowledge; Bargrams [12] and InfoZoom [13] are techniques that display the dimensions row by row and the categories are mapped to boxes whose widths are scaled according to their frequencies. The drawbacks of these frequency-based techniques are: (1) space-subdivision methods introduce a ranking of the displayed dimensions and are also limited in the number of dimensions that can be displayed, and (2) for the latter kind of visualizations, the relationships between dimensions are not shown explicitly, but the vertical alignment encodes the relation of different dimensions' categories, which can make the view difficult to understand when investigating multi-dimensional relations within the data.

A technique that is related in terms of interaction and application area are parallel coordinate trees [14]. By adding a tree-based navigation system, the data (customer surveys, similar to the data presented in the first part of the case study in this paper) can be analyzed in meaningful terms.

The Parallel Sets technique combines the advantages of frequency-based techniques (implementing a discrete design model and displaying the frequencies of categories) and parallel coordinates (treating dimensions independently).

III. Parallel Sets

Parallel Sets are not only a new visualization technique, but also an interaction framework. The visual metaphor serves as a natural way of mapping categorical variables to visual entities, which makes effective interactive exploration and analysis possible.

A. Basic Idea

Classical approaches (Section II) do not optimally deal with categorical data: either the frequency information is not visible or a ranking is imposed on the visual mapping transformation [16], influencing perception of the data (Figure 3).

Parallel Sets share the layout with parallel coordinates, but the point intersections are replaced with sets of boxes that represent the categories (Figure 3b). These boxes are scaled according to the frequencies of the corresponding categories (Section III-C) and are initially ordered according to meta information (Sections III-D and III-E). Using the frequency information means utilizing an aggregation [17] of a large categorical data set, reducing the amount of data to be displayed, and providing an image of the data that more closely resembles the way users think of large, categorical data sets.

This reduction also means that update rates of the visual representation only depend on the number of categories in the data, but not on the overall number of data points. Not only is the number of categories in

Class	Sex				
	female		male		
first	145	44.6%	180	55.4%	325
	30.8%	6.6%	10.4%	8.2%	14.8%
second	106	37.2%	179	62.8%	285
	22.6%	4.8%	10.4%	8.1%	12.9%
third	196	27.8%	510	72.2%	706
	41.7%	8.9%	29.5%	23.2%	32.1%
crew	23	2.6%	862	97.4%	885
	4.9%	1.1%	49.8%	39.1%	40.2%
	470		1731		2201
		21.4%		78.6%	100%

f_{ij} is the number of all data points which are categorized according to the ith row *and* the jth column (145 females travelled first class on the Titanic).

$r_{ij} = f_{ij}/f_{i+}$ is the individual row frequency of the same data subset, with $f_{i+} = \sum_{j=1}^{m} f_{ij}$ being the marginal row count for the ith row (almost 45% of all first class passengers on the Titanic were female).

$c_{ij} = f_{ij}/f_{+j}$ is the individual column frequency, whereas $f_{+j} = \sum_{i=1}^{m} f_{ij}$ is the marginal column count for the j-th column (e.g., 30.8% of the female passengers travelled first class).

$p_{ij} = f_{ij}/f_{++}$ is the a priori probability of this data subset, with $f_{++} = \sum\sum f_{ij}$ being the number of all data points (less than 7% of all passengers on the Titanic were females in the first class).

Fig. 5. The crosstabulation of the *Titanic* data set shows the absolute, relative, and marginal frequencies for dimensions *Class* and *Sex*.

a data set significantly smaller than the number of data points; categories are generally defined by the user, and therefore grow slowly – if at all – with data set size. As a consequence, Parallel Sets scale very well with data set size.

Because the sets of categories are placed independently side by side, the connections between categories (representing the relative number of attribute combinations) are also scaled according to their frequency values.

Parallel Sets are not restricted to categorical data, however. By means of binning, a continuous variable can be easily transformed into this kind of visualization (Section III-F).

A categorical dimension is a meaningful classification of the data, but rarely the only one. Hence, it is useful to give the user the possibility to create new classifications by combining existing dimensions. This process is user-driven: the user utilizes his or her domain knowledge to enrich the meta information about the data, and can consequently use this new information for further exploration and analysis.

Our prototype (Figure 4) shows the user not only the data itself, but also all the dimensions and categories, including hierarchies on both levels. A separate panel is used for creating new dimensions from existing ones, by simply clicking on categories or brushing value ranges in the visualization. The user can remove data values from the display (e.g., unknown values) by dragging the corresponding category to the data exclusion panel.

B. Statistical Basics

The information that is provided by the visualization is obtained by a crosstabulation [18]. Statistical examinations deal with categorical data quite frequently and usually analysts look at frequency (contingency) tables to get a first overview. Figure 5 shows an example of a two-way table: what is displayed by Parallel Sets is the information obtained by multi-way tables.

In each cell of the crosstabulation, the top left values show the occurrences f_{ij} (absolute numbers), the bottom right numbers show the absolute frequencies (probabilities) $p_{ij} = f_{ij}/f_{++}$ (where $f_{++} = \sum\sum f_{ij}$), and the remaining two values show the individual row frequencies $r_{ij} = f_{ij}/f_{i+}$ and column frequencies $c_{ij} = f_{ij}/f_{+j}$ (where f_{i+} and f_{+j} are the marginal row and column frequencies, respectively). The crosstabulation, which is calculated for each attribute combination of the displayed dimensions, builds the basis for the visual metaphor: each category is scaled according to the corresponding marginal frequency f_{i+} and f_{+j} respectively, and the connection between each pair of categories is scaled according to the absolute frequency f_{ij}. The visualization of actual data records is replaced by that of frequency information, which gives the user insight into the distribution of the data records.

C. Visual Metaphor

The basic building block of Parallel Sets is a box that represents the size of a category on one axis relative to all the data samples. Parallelograms connect categories

to show how many data points are in any of the combinations between two or more categories. Color is used to differentiate the categories and to make the connections between them easier to see.

At any point in time, there is one selected dimension – the *active* dimension. This dimension defines the color-coding of the connections: each category of the active dimension gets one color from a predefined set of equally-spaced, iso-luminant colors which differentiate the connections well [19], and all connections obtain the color of the respective active category. Then a visual ordering of the displayed dimension is introduced: starting at the active dimension, neighboring dimensions split the connections into sub-connections according to their number of categories. This is analogous to imagining a subset with a particular attribute (e.g., first-class passengers) and subdividing it according to a second feature (e.g., gender), then a third feature, and so on.

In this flexible display only the absolute frequencies are visualized, but there is room to offer more information: the user can vertically resize the boxes (representing the categories) and inside this additional space histograms can provide a more detailed view of the data. Aside from the absolute frequencies, the individual row and column frequencies of the contingency table (Figure 5) can be integrated into the visualization by the use of histograms [20] for the selected dimensions. In statistical terms, these relative frequencies are conditional probabilities. Because comparing conditional probabilities can be misleading (similar to Simpson's paradox [21]), the relative frequencies have to be standardized. One way is displaying the deviations of conditional probabilities from the a-posteriori probabilities ($\Delta P_i = P(A_i \mid B_j) - P(A_i)$). If the deviation is zero, then the particular category (with associated probability B_j) is independent of all categories of the neighboring dimension.

Figure 6 shows an example of dependent relations: one can see the absolute distribution of the upper dimension and additionally, how the particular frequencies change if only data records of the lower left categorical attribute are considered. For instance, the positive difference (9%) means that data records of the associated category (crew members) are more frequent in the considered category (male passengers, 49% of all males were crew members) than in the absolute distribution (40% of all passengers were crew).

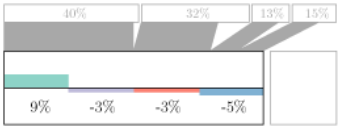

Fig. 6. The level of dependence between the class and the sex of passengers on the Titanic, shown for male passengers. The crew was overrepresented in the male population on the ship, with all the passenger classes being underrepresented.

D. Interaction and Workflow

Parallel Sets implement several common interaction schemes: selection and highlighting, interactive querying, filtering, and reordering of dimensions and categories – thus also heeding Shneiderman's visual information seeking mantra: *overview first, zoom and filter, then details-on-demand* [22] (Figure 7).

The interactive data exploration starts with an undirected investigation of the available data variables. The user chooses interesting data variables, adds them to the visualization panel, and explores their relationships. The visualization can easily become very complex if the number of displayed categories increases to more than just a few, so the user is able to create new views by defining new dimensions. These dimensions are typically more meaningful to the user or more suitable for the analysis task at hand, and compress the relevant information into one or just a few dimensions, while leaving out unwanted detail. Later, the user may need more detailed information about the relation between two or more dimensions, and can then go back to the original ones.

This high-level view of interactive visual analysis is implemented by Parallel Sets. The investigation starts with choosing interesting data variables: the data panel offers the data dimensions and the user panel shows the user-defined dimensions. The user can drag dimensions from both panels, drop them in the visualization panel, and create his or her own view of the data. The dynamic layout permits the reordering of dimensions with immediate visual feedback which is useful to look at the relationship of different dimensions more closely. Also, the categories can be reordered along their respective axis, as there may not be a natural ordering among categorical values (Figure 7a,b). A function that we found very useful is to let the program arrange the categories by absolute or relative frequencies, since often the first question is which category is the largest, and how does it relate to the categories on another dimension.

Having added interesting dimensions to the visualization (overview), the user can group selected categories (zoom and filter: Figure 7c,d), by which he or she can organize categories hierarchically. The user can also drag uninteresting categories into the exclusion panel to filter the data (Figure 7e,f), thus using the available screen

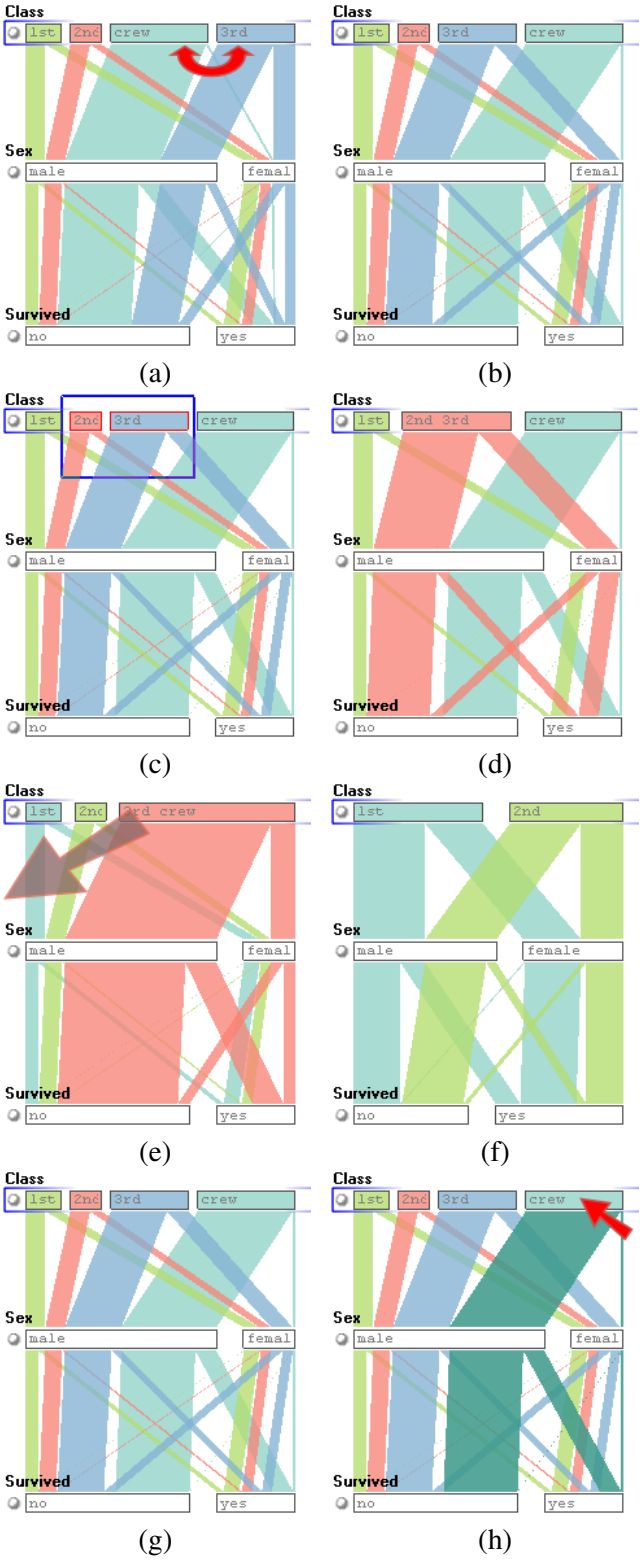

Fig. 7. Basic interaction elements in Parallel Sets: reordering categories (a, b) helps to generate a more meaningful layout; grouping categories (c, d) enables a hierarchical analysis/exploration; excluding categories from the visualization (e, f) allows for interactive filtering; and category highlighting (g, h) enables the selective investigation of high-dimensional relations.

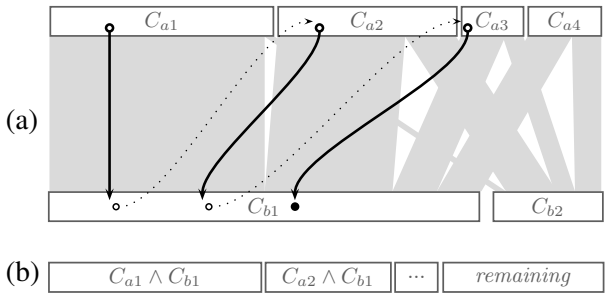

Fig. 8. An example of dimension composition: the user is interested in a classification into the following four categories: $C_{a1} \wedge C_{b1}$, $C_{a2} \wedge C_{b1}$, $C_{a3} \wedge C_{b1}$, and *remaining*. The displayed interaction path in (a) illustrates the sequence of selections (the dotted lines indicate that the user finishes the current brush and starts a new category); the resulting user-defined classification is displayed in (b).

space more effectively. Moving the mouse over a category highlights it and its connections to other dimensions (Figure 7g,h), supporting the user in understanding high-dimensional relationships.

One fundamental interaction technique in the design of Parallel Sets is dimension composition. The use of this feature is to reduce the dimensionality of the visualization – both screen space and human perception limit the maximal dimensionality of the visual mapping – and to build more practical and meaningful categorizations (Figure 8a). In contrast to *data-driven* approaches (like PCA [23] or VHDR [24]), interactive dimension composition enables the integration of the user's domain knowledge. A categorical dimension is a classification of all data records according to a particular data attribute (e.g., regarding the attribute age, a binning could classify the data into ten years intervals). In general, the data can be classified according to multiple aspects of the data. Hence, during the exploration process, it is useful to allow the user to build his or her own classifications of the data and to also *reuse* this information during further exploration and analysis. Figure 8 gives an example of the process: a new classification is created by selection activities. The path illustrates the sequence of selections; firstly, the category C_{a1} is selected, then the category C_{b1}, and so on. These selections are recorded by the user panel: for the first selection, a new dimension, an *active category* (equal to the selected category), and a *default category* (which contains all the remaining data items) are created. All successive selections are added to the current active category (by default, all selected categories are combined by a conjunction). In the example, after every two successive selections, the user indicates to start a new category (not visible). The result of the process is a new categorical dimension with four categories that represents a new classification of

the data. This dimension (representing part of the user's domain knowledge) can be dragged into the visualization again and the user can continue working with just this one dimension, because it contains all the information the user considers to be relevant. Generally, two concepts are utilized: the new categorization can either contain all possible attribute combinations (specialization), or contain a subset of these combinations (generalization). The effect is that more data axes are combined into fewer display axes, thus showing the relevant data, but keeping the visual complexity low.

The final step is to have a closer look at interesting relationships and to get detailed information. Details are in fact filtered data records that are the output of the visual analysis. Usually, once the user has found out some interesting relations within the data, he or she wants to get back to the original data items and to see all the details, e.g., in a standard table view. Concerning the investigation of relationships, Parallel Sets offers two schemes: histograms and highlighting. Histograms show statistical parameters to analyze relations in detail, highlighting is realized as a mouse-over effect: all connections that pass through the box under the cursor (i.e., relations that include the corresponding data attribute) are emphasized by drawing them with higher saturation and in front of all other connections. This way, multi-dimensional relationships become visible as a starting point for further exploration.

Zooming into a particular category is also possible, which means for its axis, only this one category is displayed, and thus a lot more room is available for its connections to other axes. This is similar to the way zooming/drill-down is handled in InfoZoom [13].

E. Hierarchical Meta Data

Meta data provides the program with information about the data set. The most basic information are the names of the different data dimensions, as well as labels for each of the values.

The meta information is organized hierarchically, which allows the definition of groups of dimensions as well as categories that belong together (rightmost part of Figure 4). This is useful in many applications, e.g., customer surveys (questions that are related or even redundant to check for validity), bank account data (account types and groups of types), etc., or in cases where the data model is such that one kind of information is spread over several dimensions. In these applications, it is often necessary to use dimension composition to produce the data that is then used for the actual analysis.

The hierarchy also provides a means of storing knowledge, especially in the case of hierarchical categories. This information is not shown in the visualization, but the user can read it from the tree. It is also possible to expand and collapse subtrees containing categories, thereby automatically combining all the categories into one. Thus, the amount of detail can be changed quickly, and in terms the user is already working with.

Using dimension composition, the user can create custom dimensions either to support a particular analysis question, or to encode information that has been found. Knowledge is accumulated this way, and stored directly in the meta data.

F. Continuous Dimensions

Many data sets do not contain only nominal or only continuous axes, but a mix of both. Parallel Sets can deal with continuous axes and show them in a meaningful way (Figure 9). In order to do this though, the non-categorical axis is divided into bins, and thus transformed into a categorical dimension. This is necessary to maintain a consistent visual metaphor, and also for performance reasons.

The user can select the number of categories the numerical axis gets split into. Such a dimension is then visualized as an axis with triangular shapes pointing at the bins, instead of parallelograms (because all bins have the same width). A histogram is of course available to see the distribution of values on the axis.

In most cases, the user will be interested in certain value intervals, which make up interesting groups of data items (e.g., certain ranges in household income). By selecting value ranges in terms of whole bins, a new dimension can be created quickly that specifies these groups. This new dimension then acts like any other categorical dimension within the program.

Showing continuous axes as true parallel coordinate dimensions would of course be the most useful display of this data, and this will be included in a future version of the program. Doing this, however, means losing the elegant independence from the data set size, and therefore requires some additional engineering to maintain interactivity for large data sets.

G. Reducing Visual Clutter

When dimensions with many categories are visualized, the many intersections between ribbons connecting categories can make the display very busy (Figure 11). A second ("bundled") mode can be used to alleviate this: the connections for each category (except the active one) are drawn in parallel (Figure 10), yielding a much tidier display, but also making it harder to track connections over more dimensions.

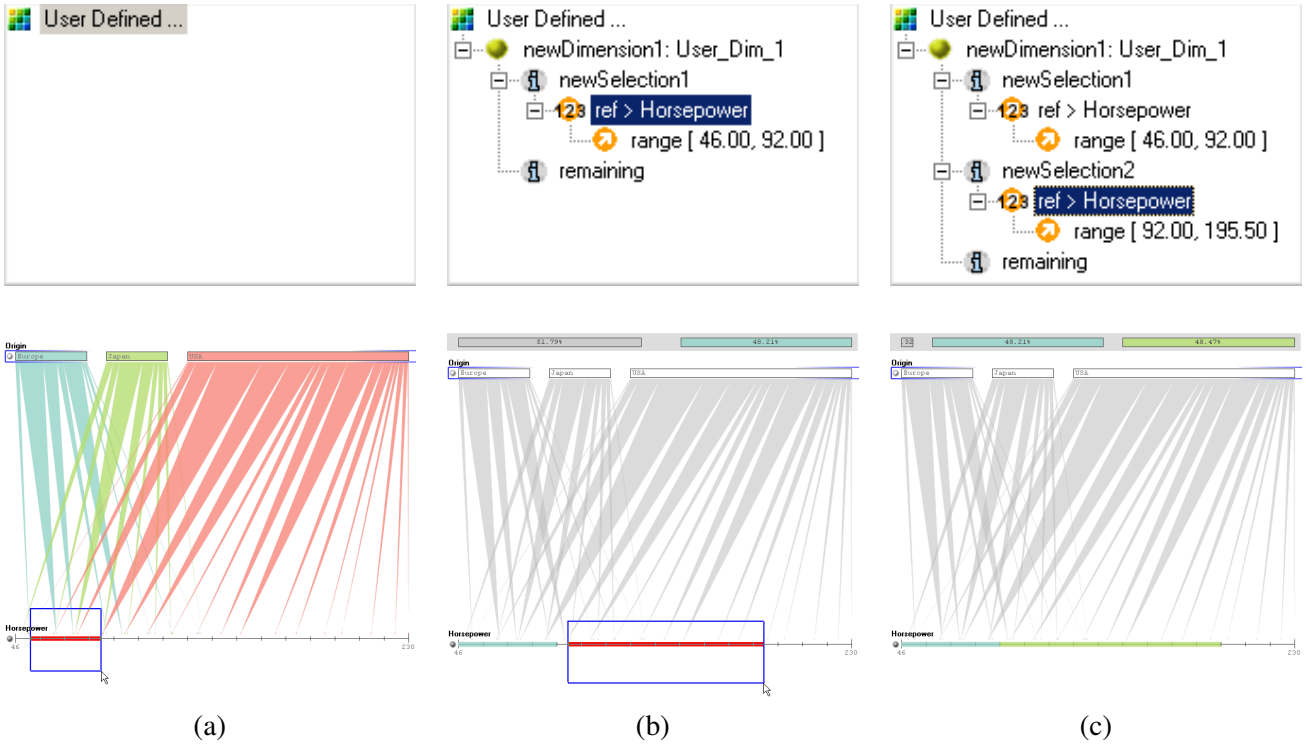

Fig. 9. Categorization of numerical dimensions. Initially, the user panel is empty and the user can start defining intervals on the axis (a). This interval is represented by the new dimension's category in the tree view and also in the visualization (b). The user can then refine the current category or create another one (c).

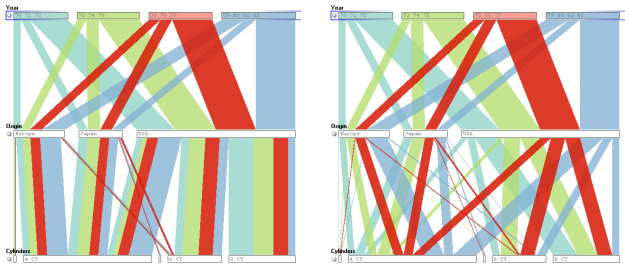

Fig. 10. Bundled mode. In bundled mode, the connections between each pair of non-active categories are parallel (left); in standard mode, the splitting up of connections is easier to see (right).

Also, it can happen that the connections between categories become so oblique that it is difficult to visually compare the represented frequencies. If this is the case, histograms can help, because they provide a very comprehensible visualization for frequency data that facilitates better comparison. In addition to the mode explained in Section III-C, relative frequencies can be displayed by this auxiliary plot: like in traditional histograms, the bars directly represent the frequency information.

All interactions are also animated, helping the user understand the results of actions. This is especially important when the display is abstract, since the user does not have a good way of orientation within the data. Animation also helps solve the problem of change blindness [25], where users cannot tell how (or even if) the display changed.

H. Interaction Support

While the user has to be able to work with the categories and dimensions directly, we found that there were several common interaction patterns that warranted automated support [17].

Creating the complete set of combinations (cross product) between two or more dimensions is one of these tasks. For two dimensions with m and n categories, respectively, the user has to click $2mn$ times. Our prototype can perform this action for the user, creating a new "user-defined" dimension, and naming the new categories after all the original categories that they were created from.

Folding a number of related binary dimensions into one that represents all of them is another task that is especially common in customer survey data. This works differently than the cross product, because the program has to understand which of the categories means "yes", and there can be overlaps between the generated categories, requiring a hierarchical approach.

The user is also often interested in sorting the categories by either absolute size or according to the degree

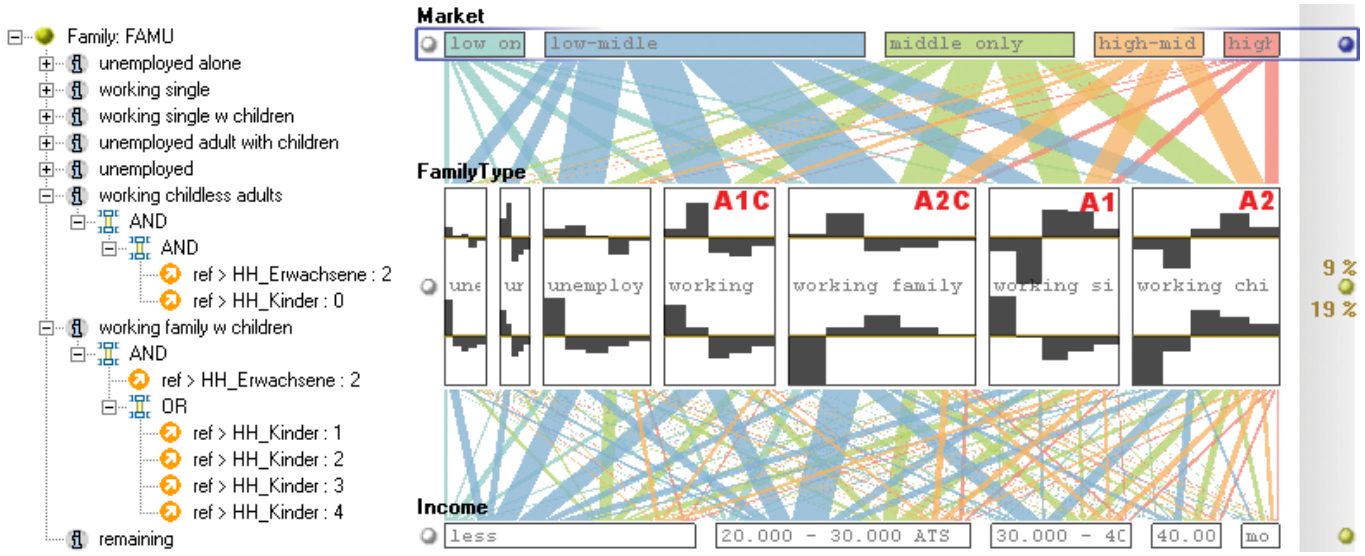

Fig. 11. Left: The meta information for the user-defined dimension *household type*; the categories are built out of available dimensions (e.g., *working childless adults* is a Boolean combination of the categories *two* (number of adults) and *unknown or none* (number of children); because the meta information is processed top-down, all unemployed persons are already classified by the above category. Right: the type of household is shown in relation to the income and the user-defined favorite supermarket dimension: households with two adults (*A2C* and *A2*) are more likely to have a higher income than ones with one adult; the choice of supermarket depends on the presence of children in the household (*HH_Kinder* is the number of children, and *HH_Erwachsene* the number of adults in a household).

of independence. Our prototype can also perform this task for the user.

While these tasks are computationally very simple, they save a lot of work, and thus improve the usability of the system considerably.

I. Performance

The analysis presented here was performed on standard consumer hardware, where the program runs at interactive speed. The construction of the view is accomplished immediately, as long as the number of displayed categories is limited to approximately twenty to thirty (the delay is directly proportional to the number of categories that have to be aggregated). Once the categorical data is transformed into the frequency information, interactions such as reordering of dimensions and categories, highlighting, and the animation between rendering modes (Section III-G), and changing of the active dimension happen without noticeable delay.

IV. Case Study

To test the usefulness of Parallel Sets in a realistic environments, we performed two case studies using our research prototype.

A. Customer Survey Data

We cooperated with a large, multi-national company in analyzing a customer relationship management (CRM) data set. This data set consists of 99 dimensions and contains information about 93,872 Austrian households.

The data contains information about people's living standards, shopping habits, pet care, etc. In addition to being high-dimensional and categorical, this data set also contains a considerable number of unknown values, as well as groups of dimensions that can be organized hierarchically.

The data fields directly reflect the layout of the questions on the survey forms, which are not particularly well suited for analysis, though. For instance, the data has separate dimensions for the number of adults in the household (*unknown*, *one*, and *two*), the number of children (*unknown or none*, *one*, *two*, *three*, and *more*), and employment (*unknown*, *unemployed*, *half-day*, and *full-time*). Combining these into one *household type* (Figure 11) reduces three dimensions to one, and also presents the data in a way that analysts are used to – especially when comparing them to the so-called *golden household*.

Supermarkets were grouped into five categories, thus reducing 16 dimensions (one for each candidate for the favorite supermarket) to one.

Income, Family Type, Supermarkets. Figure 11 shows three dimensions that are of particular interest to analysts: household types, income, and favorite supermarkets. The histograms show the frequency distribution of the market and income classes relative to the house-

hold types. Because similar histograms mean similar dependencies between the particular household types and the neighboring dimension, the top histograms show that households with children are equally distributed compared to the types without children concerning their favorite supermarkets (similar top histograms for the categories labeled *A1C* and *A2C*). The histograms also reveal that people living in households with children are more likely to buy their goods in low and middle class supermarkets in contrast to households without children.

Whether there are two adults living in the household, or only a single parent, does not seem to make a difference. What does make a difference is the total income, which is of course generally higher when there are two adults in the household, regardless of the number of children (similar histograms for the categories *A2C* and *A2*).

Detergents. Another question was wether there was a preference for different types of detergents by different types of households. This is particularly interesting when planning cross-marketing campaigns, to bundle the right "new" product with the detergent for a particular target group. The favorite detergents are again grouped into a small number of categories (*powdery*, *liquid*, *compact*, *normal*).

Figure 1 shows the relations between household types, washing agent types, residence, and income. The absolute frequencies are represented by the connections that are displayed using bundled mode (Section III-G). Concerning the detergent types, the histograms for the categories *tabs* and *liquid* are very similar and state: both types are more frequently bought by households with no children and by people that live in larger cities (similar distributions of the top and bottom histogram for the left and right type of washing agent). Also, the latter fact is correlated with having a higher income. People buying liquid washing agents or tabs are more likely to live in urban areas and have a higher income (and education, not shown) as compared to others living in the countryside.

The behavior found here can be explained by two factors: higher income in urban areas, and larger packaging sizes for washing powders. People living in Austrian cities are much more likely to have to carry their groceries over a longer distance (to the car or home), whereas people living in the countryside do all their shopping by car, anyway. Lower income and higher use of washing agents (children), combined with the relative ease of transportation make washing powders much more interesting to the rural population. Liquid washing agents and tabs are available in smaller sizes and are also more convenient to use – but more expensive. They are thus more attractive to people living in cities with more expendable income.

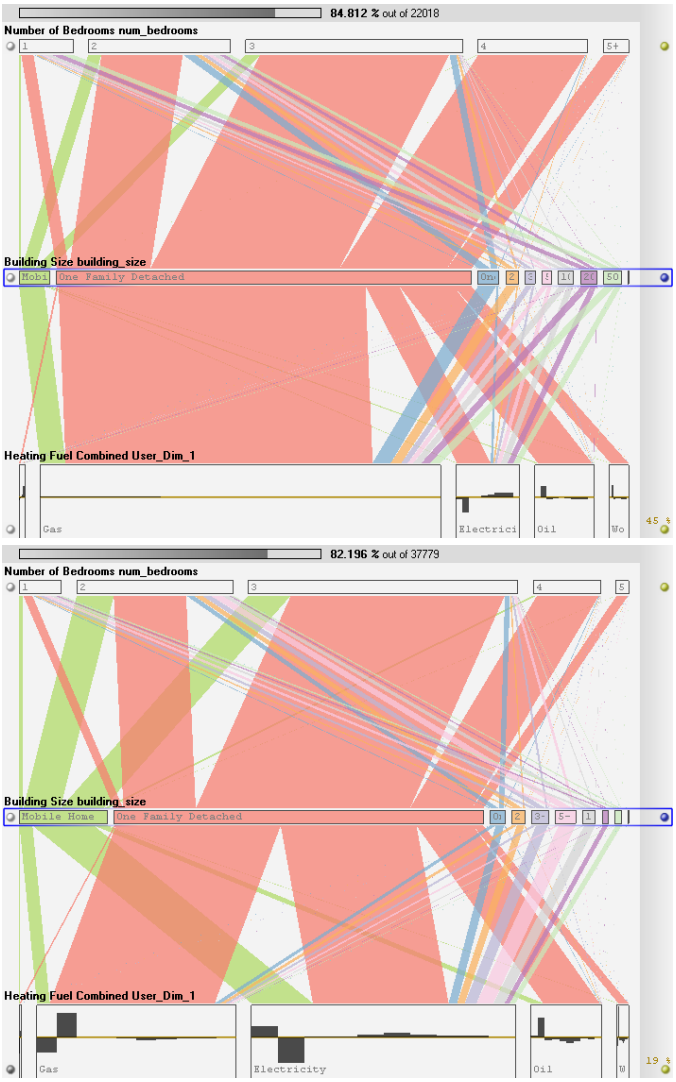

Fig. 12. Comparing housing data from Minnesota (top) and North Carolina (bottom). People in colder climates prefer gas as heating fuel over electricity, and are also less fond of mobile homes. Houses in the south have slightly fewer bedrooms on average, but the three bedroom one family house is even more common than in the north.

B. Housing Data

In a second example (Figure 12), we compare housing data from two states of the US: North Carolina (NC) and Minnesota (MN). This data was extracted from the publicly available 1% sample of the 2000 census data set [26].

A new dimension was defined to group the different heating fuels into more useful categories: natural gas (no matter if bottled, from a tank, or from a pipe), electricity, oil/kerosene, wood, and others (including the original *other* category as well as solar and wind). All *unknown* values have been removed from the display to improve

readability. And even though NC has about 70% more households, the relative number of unknown values is almost identical in the two states.

The different preferences for heating fuels are immediately visible. While 39% of households in NC use electricity for heating, only 9% do in MN. The difference goes entirely to gas, even taking away 1% from oil, and 2% from other fuels.

Another difference that can be observed easily is the much lower preference for mobile homes in MN (5%) when compared to NC (15%). Mobile homes also tend to be smaller in MN, with the majority having two bedrooms, instead of the rather equal distribution between two and three bedrooms in NC.

These differences can of course be explained with the different climates in the two states, with heating being much less expensive using gas, and also mobile homes being less suited for long, cold winters. And even though one family houses are by far the most popular in both states (61% in NC, 69% in MN), the "classical" three bedroom house has a much more pronounced majority in NC than MN – at the cost of larger houses. The values for two- and more apartment houses are very similar in both states, and only represent a clear minority in both cases.

The relative histograms in the lower part of the two images show how over- or underrepresented different heating fuels are for different housing sizes. It is easy to see that mobile homes have more than their fair share of electrical heating in NC, but are underrepresented in this category in MN. The use of gas is practically completely independent of the type of home in MN, while one-family houses have a clear preference for this fuel in NC. Wood (the right-most fuel category) is underrepresented in mobile homes in NC, but not in MN. The remaining categories share very similar patterns between the states.

V. Lessons Learned, Future Work

Overall, the response to Parallel Sets has been very positive. In addition to the case studies presented above, we have demonstrated the program to a considerable number of people from various application areas (finance, customer relations, communication, etc.), many of which had little or no prior knowledge of visualization. We found that the (potential) users of the program were immediately able to pick out relationships, and ask questions based on those findings.

While the basic interaction is very simple for users familiar with graphical user interfaces, the more advanced concepts like dimension composition and the degree of independence between axes was considerably less intuitive. Gaining new insights from using our tool certainly requires an understanding of basic statistics, and also the willingness to understand complex relationships. This is not necessarily a shortcoming of Parallel Sets, but rather a necessity when trying to find meaningful information in complex data sets. The graphical nature of visualization enables the user to experiment, and use prior knowledge of the data to explore the meanings of these concepts.

When there are many categories, or categories of very different size, it can be hard to see and compare them. This problem also affects histograms, which need some horizontal space. To alleviate this problem, the program allows the user to trade vertical for horizontal space, by changing the aspect ratios of all the category rectangles. This helps, but only up to a certain level, and also interferes with the ribbons that connect the dimensions. A distortion-oriented technique for quickly zooming into a set of categories appears to be a good solution, but has not yet been implemented. Providing histograms as a tooltip that does not depend on the size of the category is another option that we want to explore.

More histograms and statistical information need to be included. The current ones mostly served to make the users want more information directly in the display, showing different relationships and statistical measures numerically: a typical question when pointing out a relationship was "How many people does this apply to?". A related issue is finding outliers, since currently big trends are emphasized.

Navigating between and selecting dimensions is not supported by visualization, and can be rather demanding (especially in the case of high-dimensional data sets). We are therefore working on support that will show the user which dimensions may be interesting to look at more closely. This support can be both purely visual (e.g., using histograms or value and relationship displays [27]), or based on statistics and data mining.

VI. Conclusions

Categorical data plays a key role in many real-world data sets, yet has not been adequately addressed in visualization so far. By providing a visual metaphor as well as a set of interactions that allow the user to efficiently work with complex data, Parallel Sets fill an important gap.

Parallel Sets not only display data, but allow the user to create new dimensions and thus use his or her knowledge effectively when working with data. The extensive use of meta data leads to the visualization being presented in the way the user already thinks about the data, and thus makes efficient work possible.

By combining interaction with semiautomatic features, and providing the ability to store and directly use find-

ings by means of meta data, the user is able to quickly and accurately analyze large and complex data sets.

VII. Acknowledgements

This work was done in the scope of the basic research on visualization (http://www.VRVis.at/vis/) at the VRVis Research Center in Vienna, Austria, which is funded by the Austrian research program *Kplus*.

We would like to thank the reviewers of both InfoVis and TVCG for their thoughtful comments.

References

[1] F. Bendix, R. Kosara, and H. Hauser, "Parallel sets: Visual analysis of categorical data," in *Proceedings IEEE Information Visualization*. IEEE CS Press, 2005, pp. 133–140.

[2] J. J. Thomas and K. A. Cook, Eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.

[3] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proceedings IEEE Visualization*. IEEE CS Press, 1990, pp. 361–378.

[4] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang, "Mapping nominal values to numbers for effective visualization," in *Proceedings IEEE Information Visualization*. IEEE CS Press, 2003, pp. 80–95.

[5] S. T. Teoh and K.-L. Ma, "PaintingClass: Interactive construction, visualization and exploration of decision trees," in *Proceedings Knowledge Discovery and Data Mining*. ACM Press, 2003, pp. 667–672.

[6] D. F. Jerding and J. T. Stasko, "The information mural: A technique for displaying and navigating large information spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 4, no. 3, pp. 257–271, July–Sept. 1998.

[7] J. A. Hartigan and B. Kleiner, "Mosaics for contingency tables," *Proceedings Symposium on the Interface*, pp. 268–273, 1981.

[8] M. Friendly, "Visualizing categorical data: Data, stories and pictures," *SAS User Group Conference*, pp. 190–200, 1992.

[9] J. LeBlanc, M. O. Ward, and N. Wittels, "Exploring n-dimensional databases," in *Proceedings Visualization*. IEEE CS Press, 1990, pp. 230–237.

[10] M. Theus, H. Hofmann, B. Siegl, and A. Unwin, "MANET: Extensions to interactive statistical graphics for missing values," in *New Techniques and Technologies for Statistics II*. IOS Press Amsterdam, 1997, pp. 247–259.

[11] H. Hoffmann, "Exploring categorical data: Interactive mosaic plots," *Metrika*, pp. 11–26, 2000.

[12] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton, "Parallel bargrams for consumerbased information exploration and choice," in *Proceedings ACM User Interface Software and Technology*. ACM Press, 2001, pp. 51–60.

[13] M. Spenke and C. Beilken, "Visualization of trees as highly compressed tables with InfoZoom," in *Proceedings IEEE Information Visualization*. IEEE CS Press, 2003, pp. 122–123.

[14] D. Brodbeck and L. Girardin, "Visualization of large-scale customer satisfaction surveys using a parallel coordinate tree," in *Proceedings Information Visualization*, 2003, pp. 197 – 201.

[15] "Titanic data set (statlib)." [Online]. Available: http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html

[16] S. K. Card, J. D. Mackinlay, and B. Shneiderman, "Using vision to think," *Readings in Information Visualization: Using Vision to Think*, pp. 579–581, 1999.

[17] D. A. Keim, "Information visualization and visual data mining," *Transactions on Visualization and Computer Graphics*, vol. 7, no. 1, pp. 100–107, 2002.

[18] "Electronic Statistics Textbook." [Online]. Available: http://www.statsoft.com/textbook/stathome.html

[19] "ColorBrewer." [Online]. Available: http://colorbrewer.org/

[20] H. Hauser, F. Ledermann, and H. Doleisch, "Angular brushing of extended parallel coordinates," in *Proceedings Information Visualization*. IEEE CS Press, 2002, pp. 127–130.

[21] A. Agresti, *An Introduction to Categorical Data Analysis*. Wiley & Sons, 1996.

[22] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings Visual Languages*. IEEE CS Press, 1996, pp. 336–343.

[23] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press Professional, Inc., 1990.

[24] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang, "Visual hierarchical dimension reduction for exploration of high dimensional datasets," in *Proceedings Joint IEEE TCVG - EUROGRAPHICS Symposium on Visualization*. Eurographics Association, 2003, pp. 19–28.

[25] L. Nowell, E. Hetzler, and T. Tanasse, "Change blindness in information visualization: A case study," in *Proceedings IEEE Information Visualization*. IEEE CS Press, 2001, pp. 15–22.

[26] U. Cvek, M. Trutschl, and M. Wattenberg, "IEEE InfoVis 2006 Contest, US 2000 Census Data." [Online]. Available: http://sun.cs.lsus.edu/iv06/

[27] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner, "Value and relation display for interactive exploration of high dimensional datasets," in *Proceedings Information Visualization*. IEEE CS Press, 2004, pp. 73–80.

Robert Kosara is an Assistant Professor at the University of North Carolina at Charlotte. He received his Ph.D. in computer science from Vienna University of Technology in 2001. His research interests include information visualization, interaction, and the use of artistic and psychological principles for visual analysis. Robert can be contacted at rkosara@uncc.edu.

Fabian Bendix received his Master's degree in computer science from Vienna University of Technology in 2005. He now works for the company Inte:Ligand, where he develops visualization software for computer-supported pharmaceutical research. Fabian Bendix can be contacted at bendix@inteligand.com.

Helwig Hauser is the scientific director of the VRVis Research Center in Vienna, Austria. He graduated from Vienna University of Technology in 1998 with a PhD in computer science. His primary interests include the interactive visual analysis of large and complex data sets as well as classical scientific and information visualization. Helwig Hauser can be contacted at Hauser@VRVis.at.