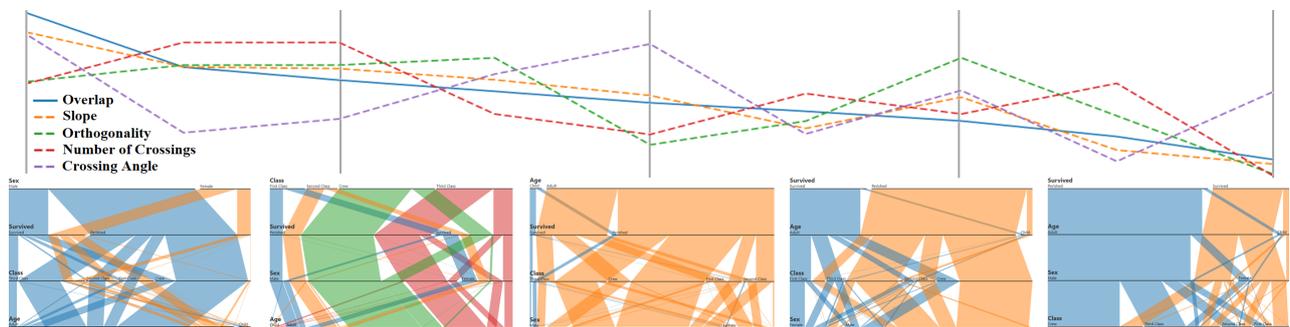


# ParSetgnostics: Quality Metrics for Parallel Sets

Frederik L. Dennig<sup>1</sup> , Maximilian T. Fischer<sup>1</sup> , Michael Blumenschein<sup>1</sup> , Johannes Fuchs<sup>1</sup> ,  
Daniel A. Keim<sup>1</sup> , and Evanthia Dimara<sup>1,2</sup> 

<sup>1</sup>University of Konstanz, Germany

<sup>2</sup>Utrecht University, Netherlands



**Figure 1:** The layout of dimensions and categories significantly influences the appearance of Parallel Sets. We show five different arrangements of the Titanic dataset [Daw95], ordered according to the Overlap metric, showing decreased clutter and improved readability from left to right. The lines represent the values for five metrics for each of the Parallel Sets visualizations, measuring different visual properties.

## Abstract

While there are many visualization techniques for exploring numeric data, only a few work with categorical data. One prominent example is Parallel Sets, showing data frequencies instead of data points - analogous to parallel coordinates for numerical data. As nominal data does not have an intrinsic order, the design of Parallel Sets is sensitive to visual clutter due to overlaps, crossings, and subdivision of ribbons hindering readability and pattern detection. In this paper, we propose a set of quality metrics, called ParSetgnostics (Parallel Sets diagnostics), which aim to improve Parallel Sets by reducing clutter. These quality metrics quantify important properties of Parallel Sets such as overlap, orthogonality, ribbon width variance, and mutual information to optimize the category and dimension ordering. By conducting a systematic correlation analysis between the individual metrics, we ensure their distinctiveness. Further, we evaluate the clutter reduction effect of ParSetgnostics by reconstructing six datasets from previous publications using Parallel Sets measuring and comparing their respective properties. Our results show that ParSetgnostics facilitates multi-dimensional analysis of categorical data by automatically providing optimized Parallel Set designs with a clutter reduction of up to 81% compared to the originally proposed Parallel Sets visualizations.

## CCS Concepts

• **Human-centered computing** → Visualization design and evaluation methods;

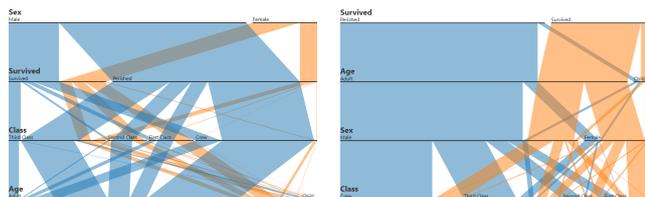
## 1. Introduction

Nominal data is an inherent data type in many real-world datasets. Examples include business intelligence, when assigning personnel to tasks and resources, or inventory data, when describing product qualities like color. However, most multi-dimensional visualization techniques, such as scatterplot matrices [Har75, Cle86], parallel coordinates [Ins85], and projections [CD18], are designed for numerical data, where data values come with a meaningful scale or ordering. In contrast, nominal data does not have an intrinsic ordering

or distance between the values. Instead, it describes properties in name only, requiring context for analysis. Frequency-based visualizations [Hof00, WLHS01, SB03] are a possible solution mapping categorical variables to their corresponding frequencies. Yet, for most techniques, the frequency information is often not visible or imposes a hierarchical structure. On the other hand, solutions that treat dimensions independently [RRBW03, TM03, JS98], mapping categories to numbers, follow a continuous design model which deviates from the discrete mental user model of the data [KBH06]. Parallel Sets visualization (ParSets) is a hybrid solution combin-

ing the strengths of frequency-based designs with the independent treatment of dimensions, which is essential for multi-dimensional analysis of categorical data [BKH05, Kos10].

To support multi-dimensional analysis of categorical data, Parallel Sets appropriate the layout of parallel coordinates [Ins85]. They replace the polylines representing numerical data points with parallelograms, called ribbons, representing the size and the frequency of the categories. Parallel Sets serve as an interaction framework used in various fields that require user-driven analysis of heterogeneous and multi-dimensional categorical data. Compared to other visualization types, Parallel Sets offer fewer degrees of freedom with respect to design considerations, making them a compelling solution for the challenging representation of categorical data. In contrast, Sankey diagrams [KS98] exhibit more degrees of freedom, such as placing dimension axes or sections of them freely on the chart, while stacked bars can show the same data without the explicit links between the individual values of each dimension. However, as nominal data does not have an intrinsic order, the readability of Parallel Sets depends on the chosen ordering of dimensions, as well as the ordering of categories within each dimension. Certain dimension and category orderings are more challenging to read than others. Figure 2 shows two Parallel Sets of the same data. The left Parallel Sets appears harder to read due to the high degree of clutter. On the right, an alternative reordering with minimized ribbon overlap has reduced clutter and is thus easier to read.



**Figure 2:** Two Parallel Sets of the Titanic dataset [Daw95]. The right version has less ribbon overlap than the one on the left. It is also easier to read because of the reduced amount clutter.

Identifying the optimal data representation of Parallel Sets can be challenging. Parallel coordinates can apply to categorical data. However, the frequency information is lost. For exploratory scenarios, choosing an adequate Parallel Sets configuration for the dataset is key to the understanding and knowledge gained in the process. Manual reordering of dimensions is not always feasible due to the large set of possible dimensions and category orderings. We note that the number of possible configurations exceeds those of parallel coordinates because the order of categories can be chosen freely. There are  $|C_d|!$  possible orderings of a dimension axis, where  $C_d$  are the dimension values of dimension  $d \in D$ . The dimension axis themselves can be reordered and allow for  $|D|!$  orderings. Thus, there are a total of  $|D|! \cdot \prod_{d \in D} |C_d|!$  possible Parallel Sets visualizations. Existing approaches focus on interaction [ZCYY19], which requires user interaction and suffers from summarization that loses information and imposes a biased first view [HD12] by reducing the dimensionality and number of categories. Automatic solutions to designing Parallel Sets do not sufficiently support data analysis in fully exploratory scenarios because they limit the dimensionality

of the displayed subsets [AHZ\*14]. Thus, these approaches often exclude possibly relevant information beforehand.

This paper contributes eight quality metrics. The metrics *Overlap*, *Slope*, *Orthogonality*, *Number of Crossings*, and *Crossing Angle* focus on the ordering of categories, while the metrics *Number of Ribbons*, *Ribbon Width Variance*, and *Mutual Information* focus on the ordering of dimensions. All metrics allow for the comparison and ranking of Parallel Sets to reduce clutter and improve their readability. To develop these metrics, we formalized the geometric properties of Parallel Sets. Additionally, we discuss the parameters of Parallel Sets in the context of readability. We evaluate our approach by applying our technique to six datasets from previous publications, showing that ParSetgnostics improve their readability. To make the acquired knowledge accessible while supporting the creation of optimized Parallel Sets, we provide the *ParSetgnostics Explorer* (dennig.dbvis.de/parsetgnostics). For reproducibility, we make all our statistical analysis, results, and source code available at [osf.io/rwhf5](https://osf.io/rwhf5). With this work, we hope to improve categorical data visualization, especially for exploratory tasks.

## 2. Related Work

### 2.1. Improvements of Parallel Sets

Parallel Sets can be improved through visual approaches. These techniques change the representation of ribbons to make them easier to follow. A common visual method for improving the readability of Parallel Sets in this way is to curve the ribbons of Parallel Sets [RWH\*16]. Another technique is to draw ribbons with a fixed angle, called Common Angle Plots [HV13], yielding better readability. This technique addresses the effects of a class of perceptual illusions, called Müller-Lyer illusion [DS91, Gol14], where lines appear to have a different distance or length. Our approach differs from these techniques in that we propose a different layout of coordinate axes and categories. Techniques changing the representation of ribbons can be applied after our quality metrics have been used to determine a useful dimension- and category ordering, further improving the readability. There also exist a set of dimension ordering strategies for parallel coordinates [BZP\*20], which can apply to Parallel Sets if modified. Parallel Sets can also be improved in a semi-automatic way, using machine learning or statistical methods. The interactive approach by Zhang et al. [ZCYY19] uses association rule mining to reduce the number of dimensions and categories, requiring user interaction. The approach by Alsakran et al. [AHZ\*14] changes the layout and ordering of dimension axes but restricts the dimensionality of the subgroups, i.e., ribbons, to two dimensions. This approach simultaneously uses mutual information [Sha48] to measure the dependence of two variables. Both techniques remove dimension information or data from the visualization. Our approach differs in that it does not remove any data and does not restrict the dimensionality of the displayed ribbons but tries to optimize a set of target properties.

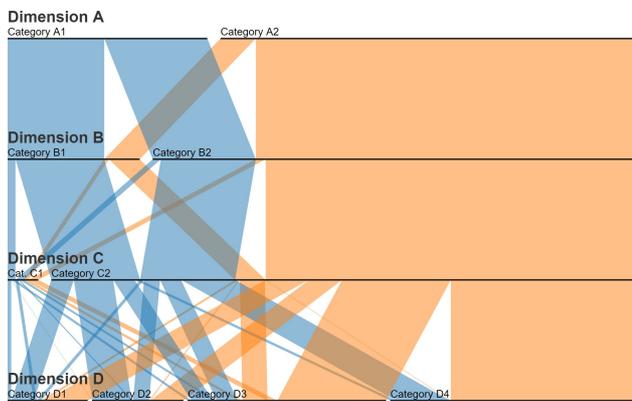
### 2.2. Quality Metrics for Visualization Techniques

Screen-space quality metrics describe a set of metrics specifically designed metrics or features that measure the quality of visualization and can be used to optimize them for readability or quantify

the appearance of specific patterns [BBK\*18]. They do not remove any information from the visualization. They rather measure properties of the visualization, which can be used to compare and rank them. Examples of those approaches are: Magnostics for matrix visualizations [BBH\*17], Scagnostics for scatterplots [WAG05], Pargnostics for parallel coordinate plots [DK10], Visualgnostics projections of high-dimensional data [LKZ\*15], and Pixgnostics for pixel-based visualizations [SSK06]. We contribute to this area of information visualization by providing a set of eight metrics for the quantification of visual properties of Parallel Sets. In this way, we improve the quality of Parallel Sets without performing any sampling or dimensionality reduction of the underlying data.

### 3. Parameters of Parallel Sets

In this section, we provide the necessary definitions to describe the properties of Parallel Sets formally. We also discuss the parameters of Parallel Sets in light of semi-automatic and fully automatic reordering of dimensions and categories.



**Figure 3:** A Parallel Sets visualization showing a generic example with four dimensions (A-D) and their respective categories (of cardinality two for dimensions A-C and four for dimension D).

#### 3.1. Background

Parallel Sets are a visualization type for categorical data. An example of a Parallel Sets visualization is shown in Figure 3. Parallel Sets show flow-paths that divide the flow into smaller and smaller subsets at each dimension if a dimension splits the subset into multiple categories. This introduces a direction or flow, in the case of Figure 3 from top to bottom, while also increasing granularity with each dimension axis splitting the dataset into smaller subsets. Every dimension is represented by an axis and a set of ribbons. Each ribbon represents a subset defined by the categories above and the one category connected to the following dimension axis. Compared to parallel coordinates, the individual categories on the dimension axis are not discrete points. Instead, the axis and the width of the ribbon are proportional in size to their flow, i.e., the number of data items with the corresponding categories they represent. They can be compared to stacked bars. However, stacked bars

only display dimensions that can show the same data without the explicit links between. Sankey diagrams exhibit more degrees of freedom, such as placing dimension axes or sections of them freely on the chart.

#### 3.2. Definitions

This work aims to optimize a Parallel Sets visualization by ordering the dimensions and categories to conform better to the design considerations described in the following. We developed our metrics with the general idea of quality metrics for information visualization described by Behrisch et al. [BBK\*18] in mind. With the definition of a quality criterion (see Equation 1) provided in their work, the problem is described formally:

$$\arg \max_{\phi \in \Phi} \min q(\phi|D, U, T) \quad (1)$$

$D$  denotes the data,  $U$  the user, and  $T$  the task.  $\phi$  denotes a specific configuration of a visualization of the set of all possible configurations of a given visualization type  $\Phi$ .  $q$  describes a quality criterion and  $\arg \max/\min_{\phi \in \Phi}$  optimization strategy. In this work, we focus on defining quality criteria  $q$  for Parallel Sets visualizations, i.e., a set of objective functions (see Section 4). We test our quality metrics with six datasets, which in this definition corresponds to  $D$  (see Subsection 5.1). The metrics can be task and user-dependent. The user can choose which quality metrics he aims to minimize or maximize or even how to weight them. It is also possible to limit  $\Phi$  by choosing a set of constraints, e.g., filtering or sampling. In our work, we consider the task  $T$  to be an exploration task with no prior knowledge of the specifics of the dataset. The result is  $\phi$ , in our case, the configuration of a Parallel Sets visualization, defined by the order of dimensions  $\mathbf{D}$  and the order of categories of all dimensions  $C_D$  for all dimensions  $D \in \mathbf{D}$ .

The appearance of Parallel Sets depends on the ordering of the dimensions. We define  $\mathbf{D}$  as the ordered set of all dimensions of a purely categorical dataset:

$$\mathbf{D} := (D_1, D_2, \dots, D_i) \quad (2)$$

Similarly, we define the ordering of the category values  $C_{D_i}$  of a single dimension  $D_i \in \mathbf{D}$  as:

$$C_{D_i} := (C_{D_i}^1, C_{D_i}^2, \dots, C_{D_i}^j) \quad (3)$$

where  $C_{D_i}^j$  is a single category in a specific dimension  $D_i$ . This is consistent with the tree-like structure of Parallel Sets [Kos10], separating the dataset into smaller subsets while descending the tree levels, where each level represents a dimension axis  $D_i$ . Ribbons are representatives of edges between two levels, i.e., connections between two adjacent dimension axes  $D_n$  and  $D_{n+1}$ . Thus, we can define the possible ribbons  $R_n^*$  between two adjacent dimensions for  $n \in [1, |D| - 1]$  as:

$$R_n^* := \bigtimes_{i=1}^{n+1} C_{D_i} \quad (4)$$

Since  $R_n^*$  denotes all possible ribbons between two dimensions, it includes empty subsets. Parallel Sets do not visualize empty or non-existent subsets. Thus, we remove such ribbons from the list by verifying that at least one entry exists that belongs to a subset defined

by a ribbon  $r$ , i.e.,  $|r| > 0$ . This yields the set of all existing ribbons between two dimensions axes, which we define as:

$$R_n := \{r \mid r \in R_n^* \wedge |r| > 0\} \quad (5)$$

Finally, we can define the set of all existing ribbons  $\mathbf{R}$  and analogous the set of all possible ribbons  $\mathbf{R}^*$  as:

$$\mathbf{R} := \bigcup_{i=1}^{|D|-1} R_n \quad \mathbf{R}^* := \bigcup_{i=1}^{|D|-1} R_n^* \quad (6)$$

### 3.3. Parameter Space

In the next section, we will discuss the specific parameters and caveats of Parallel Sets related to the choice of the category and dimension ordering, dataset-dependent properties, and ribbon parameters. We will use those parameters to explain our metrics described in Section 4.

**Selection of the first dimension:** The analysis task is *the* determining factor for the axes ordering. The first dimension and its categories determine the ribbon color, and thus the main aspects the analysis focuses on. In case there exists a formulated analysis question or hypothesis, we suggest determining this dimension beforehand or interactively. A partial ordering is possible. The user with domain knowledge can decide best which dimensions are more important than others. In the case of an exploratory scenario, we suggest a fully automatic approach, generating multiple clutter reduced and readability improved versions with different axes orderings to allow for an overview of the dataset. We suggest choosing the first dimension based on the dimension with the highest entropy for a fully automatic approach, thus focusing on the dimension with the most significant amount of information. Thus, it is a dimension with balanced category sizes. Dimensions with low entropy will contain more categories of less size, making them hard to read.

**Ordering of remaining dimensions:** The following axes split the ribbons into increasingly fine-grained subsets, each split according to a dimension's categories. With the increasing amount of ribbons, clutter is likely to increase. The strength of this effect is ultimately dependent on the dataset. We identified two effects on the ribbons linked to this parameter: the number of ribbons and the ribbon widths. Firstly, the number of ribbons should be kept as low as possible to avoid premature splitting into subsets. Secondly, the ribbon widths should be kept as large as possible to keep them easy to follow. This properties is also influenced by the slope of the ribbon, dependent on the ordering of categories.

In a fully automatic approach, the order can be determined by three strategies: (1) Order the dimensions by ascending number of categories, minimizing the number of ribbons. (2) Minimizing the ribbon width, lowering the number of thin ribbons, which are hard to perceive. (3) Ordering the dimension based on information-theoretic property, such as mutual information [Sha48].

**Ordering of categories:** While there is no natural order among nominal values and the order of categories on each dimension can be chosen freely [BKH05], not every category ordering is intuitive, useful, or supportive for exploratory or confirmatory data analysis.

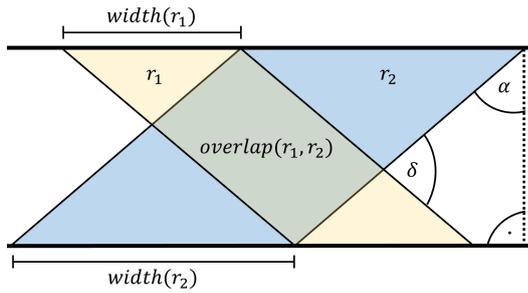
Some category orderings lead to a high degree of clutter by increasing the slope and overlap of ribbons. Therefore, the category ordering can be optimized such that the Parallel Sets visualization is readable and shows patterns inside the data, even with an increasing amount of ribbons caused by splits according to dimension axes. Since this parameter offers the most potential for improvement, five of the eight metrics we define are sensitive to category reordering and are designed to help analysts in their choice of dimension and category ordering. However, given that some categorical data is ordinal, e.g., time, the sequence is fixed by the inherent order and should not be changed.

**Impact of number and size of categories:** Dimensions with many categories split the data into many small ribbons that are hard to follow. Additionally, since the number of ribbons monotonously increases with every dimension axis, this leads to an increased number of ribbons in every following dimension. The data distribution is the determining factor, i.e., dimensions having a few categories of equal size, or the many small categories or a mixture thereof. The issue can be addressed by delaying splits yielding thin ribbons to later dimensions, i.e., prioritizing dimensions with large equal-sized categories. Such a dimension should be placed at the beginning of the dimension ordering.

**Influence of the distance between dimension axes:** A short distance increases the slope of diagonal ribbons, which increases the overlap of ribbons and clutter. Since ribbons are parallelograms, this reduces the perceived width [PDK\*19]. In contrast, an excessively large distance makes ribbons, especially thin ones, hard to follow since they are visually less prominent due to their small surface area. Additionally, it decreases the crossing angle of ribbons, which makes them also harder to follow [HHE08,WPCM02]. This parameter is ultimately dependent on the available screen-space and its aspect ratio. Four category ordering-dependent metrics, namely *Overlap*, *Slope*, *Orthogonality*, *Crossing Angle* are sensitive to this parameter. We fixed the distance between the dimensions for all our measurements.

**Impact of ribbon width and plot width:** The width of the ribbons is dependent on the available plot space. In the case of a vertical ribbon flow, it depends on the plot width. For a horizontal ribbon flow, it will depend on the plot height. The width of all ribbons remains relative, as with the number of ribbons, the ribbons width decreases. The plot size should be chosen accordingly. All ribbons, especially those representing small subsets, should have a large enough width such that they can be visually compared and easy to follow. With increasing plot size, the distance between the dimension axes also increases. Four of our category ordering-dependent metrics are sensitive to this parameter. Thus, we also choose a constant plot size for all our measurements.

**Selection of ribbon colors:** The ribbon color is not considered by our metrics. However, we suggest choosing colors according to common criteria, i.e., easy to differentiate colors [MJSK15, BHH03]. Since the number of colors is equal to the number of categories of the first dimension, it is beneficial to reduce the number of colors by selecting a dimension with a low number of categories that is still pertaining to the analysis question. In exploratory



**Figure 4:** This figure shows the geometric variables required for our metrics. It shows two ribbons  $r_1$  and  $r_2$  between two dimension axes.  $overlap(r_1, r_2)$  defines as their shared area. The angle  $\alpha$  denotes the slope of a ribbon. An orthogonal ribbon has a slope of  $\alpha = 0$ .  $\delta$  describes the crossing angle of  $r_1$  and  $r_2$ . The width of a ribbon is the distance of the intersections with a dimension axis.

tasks, we suggest a dimension with a category count no larger than nine based on Miller's Law [Mil56]. Parallel Sets are intrinsically "2.5D," meaning that the ribbons can have an ordering along the depth direction. The typical solution to avoid occlusion is to use transparency to show the path and area of overlapping ribbons. In this case, the colors of ribbons need to be chosen such that the mixtures of colors produce a distinguishable color that still implies which ribbons are crossing. If no transparency is used, we suggest ordering the ribbon, such that the thinner ribbon is on top of a thicker one to minimize occlusion.

#### 4. Metrics for Parallel Sets

This section describes and discusses a set of eight quality metrics that measure different properties of Parallel Sets. These properties are dependent on the dimension and category ordering. These properties are either desirable or undesirable, and thus, our metrics can be used to compare Parallel Sets and help adjust them to be more readable and interpretable. For explanation and comparability, we use the Titanic dataset [Daw95] to show-case their effects.

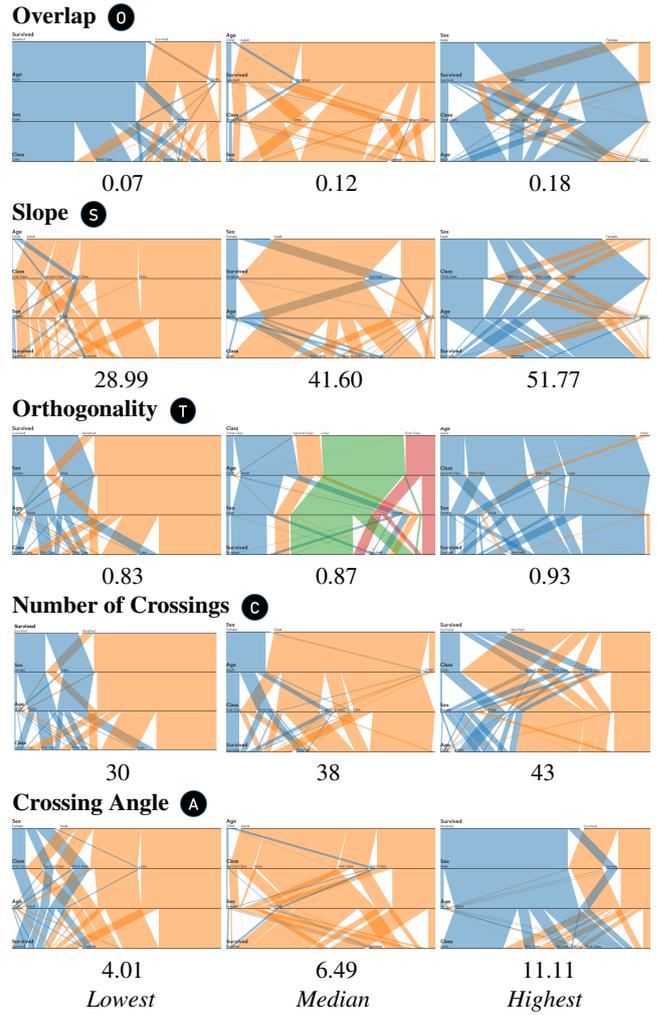
##### 4.1. Category Ordering-dependent Metrics

We present five category ordering-dependent metrics, which means that they are sensitive to the reordering of dimensions and individual categories of a dimension. Small changes in the order of categories can already have a large impact on the appearance of a Parallel Sets visualization. Three category ordering-dependent metrics consider the relationships of pairs of ribbons between two dimension axes. We describe this set as follows:

$$P_i := \{(r_1, r_2) | (r_1, r_2) \in R_i \times R_i \wedge r_1 \neq r_2\} \quad (7)$$

The set  $P_i$  describes all possible pairs of ribbons between the dimensions  $D_i$  and  $D_{i+1}$  and is required for the *Overlap* and *Number of Crossings* and *Crossing Angle* metrics.

**Overlap**  $\odot$  measures the overlapping area of all ribbons. A high overlap is indicative of clutter since overlapping areas are harder to interpret, since crossing ribbons are harder to follow [HHE08,



**Figure 5:** We show three Parallel Sets visualizations for each of the five category ordering-dependent metrics: *Overlap*, *Slope*, *Number of Crossings*, and *Crossing Angle*. We show the Parallel Sets corresponding to the lowest, median, and highest metric value. Lower values signify less clutter and thus improved readability, presenting a good starting point for exploratory data analysis.

WPCM02]. Furthermore, there is a connection to the slope of a ribbon as only sloped ribbons contribute to overlap. The overlap is especially high if large subsets overlap in their ribbon representation. We formally describe this metric in Equation 8.

$$OVERLAP := \frac{1}{A} \sum_{i=1}^{|D|-1} \sum_{(r_1, r_2) \in P_i} overlap(r_1, r_2) \quad (8)$$

The set of tuples  $P_i$  defines all possible pairs ribbon between two neighboring dimension axes the Parallel Sets.  $A$  denotes the area of the Parallel Sets visualization. The factor  $\frac{1}{A}$  allows for the comparability of different Parallel Sets visualizations on different resolutions.  $overlap(r_1, r_2)$  with  $r_1, r_2 \in \mathbf{R}$  defines the overlapping area of two ribbons as described in Figure 4. The examples shown in Fig-

Figure 5 show the effects of reducing the overlap of ribbons yielding a Parallel Sets visualization with a low degree of clutter.

**Slope**  $\textcircled{S}$  measures the average slope of all ribbons. A low average slope is preferable since ribbons that have a high angle to the dimension axes are easier to follow [HHE08, WPCM02]. This is grounded in the area preserving geometrical properties of parallelograms. Highly sloped ribbons get thinner and longer [PDK\*19]. Only sloped ribbons contribute to overlap. The *Slope* metric differs from the *Overlap* metric in that it is not affected by the ribbon width, meaning that the *Slope* metric is not weighting the slope by the size of the subset that the ribbon represents. We formally describe this metric in Equation 9.

$$\text{SLOPE} := \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \alpha(r) \quad (9)$$

In this equation, the slope of a ribbon is denoted by angle  $\alpha$ , which is geometrically defined as depicted in Figure 4. The effects of minimizing the *Slope* metric can be observed in Figure 5. A low average slope reduces clutter, while high *Slope* introduces a noticeable zigzag pattern which is hard to interpret.

**Orthogonality**  $\textcircled{T}$  leverages the concept to the *Slope* metric but explicitly focuses on the orthogonality of ribbons. This focus restricts the layout of ribbons to enforce a close to a perpendicular angle to the dimension axis. This property increases readability [HHE08, WPCM02]. It measures the average number of ribbons with a slope  $\alpha$  smaller than a threshold value  $\tau$ . We formally describe this metric in Equation 10.

$$\text{ORTHOGONALITY} := \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \begin{cases} 1 & \text{if } \alpha(r) > \tau \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

A group of ribbons that is perpendicular to the dimension axes shows a categorical correlation. Therefore, we choose  $\tau = 0$ . However,  $\tau$  can be chosen with respect to the target orthogonality, such that slightly sloped ribbons are also considered. In Figure 5, we can see that enforcement of perpendicular ribbons, forming rectangles, reduces clutter. In the example of the Titanic dataset [Daw95] it improves the Parallel Sets visualization even more than the *Slope* metric, significantly differing from it.

**Number of Crossings**  $\textcircled{C}$  measures the number of ribbon crossings. This metric is analogous to the *Number of Line Crossings* metric of the Pargnostics [DK10] metric set for parallel coordinates. A high number of crossing produces similar patterns like dissimilarity orderings for parallel coordinates, which can be used to detect patterns [BZP\*20]. In Parallel Sets visualizations a high degree of ribbon crossings can lead to visual clutter, making ribbons hard to follow. This effect has been observed for parallel coordinates [ED06]. Thus, a very high and very low value for *Number of Crossings* can indicate an interesting Parallel Sets for exploratory analysis. The value  $C$  in Equation 11 describes the absolute number of crossings.

$$C := \sum_{i=1}^{|D|-1} \sum_{(r_1, r_2) \in P_i} \begin{cases} 1 & \text{if } \text{overlap}(r_1, r_2) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

We formally describe this metric in Equation 12, which provides a relative number of crossing proportional to the number of ribbons contained in a Parallel Sets visualization.

$$\text{CROSSINGS} := \frac{C}{|\mathbf{R}|} \quad (12)$$

The examples depicted in Figure 5 show that a minimization of the number of crossings progressively reduces the amount of clutter. A Parallel Sets visualization with a maximum number of is likely to exhibit zigzag patterns.

**Crossing Angle**  $\textcircled{A}$  quantifies the average crossing angle of crossing ribbons of a Parallel Sets visualization. This metric is motivated by the *Angels of Crossing* metric of the Pargnostics [DK10] metric set for parallel coordinates. A very high or very low angle of crossing benefits the readability of the Parallel Sets visualization. Ribbons crossing at a flat angle are hard to follow compared to ribbons crossing at close to right angles. This effect has already been observed for lines [HHE08, WPCM02]. We formally describe this metric in Equation 13.

$$\text{CROSSINGANGLE} := \frac{1}{C} \sum_{i=1}^{|D|-1} \sum_{(r_1, r_2) \in P_i} \delta(r_1, r_2) \quad (13)$$

In this equation, the crossing angle of two ribbons is denoted by angle  $\delta$ . The factor  $\frac{1}{C}$  based on Equation 12 provides a value relative to the total number of crossings. The concept of a crossing angle and how it is described by  $\delta$  is depicted in Figure 4. In the examples shown in in Figure 5 this metric offers Parallel Sets visualizations with a low amount of clutter for a high and low value, while the median exhibits a zigzag pattern and clutter. In general, a high crossing angle is preferred since it supports readability.

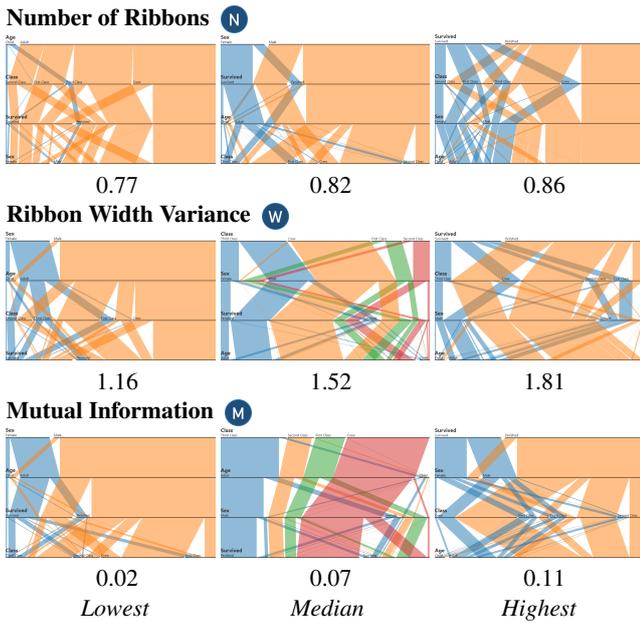
## 4.2. Dimension Ordering-dependent Metrics

This section describes three dimension ordering-dependent metrics, which means that they are only sensitive to the reordering of dimensions and are not affected by changes in the order of categories of any dimension axes. These metrics can be used to limit the search space by fixing the order of dimension axes.

**Number of Ribbons**  $\textcircled{N}$  measures the number of ribbons. The number of ribbons determine the number of ribbon splits according to the categories of dimension axes. In general, a low number of splits is preferable since a high number of ribbons increase the likelihood of sloped and overlapping ribbons. Furthermore, splits reduce the ribbon width, creating thin ribbons, which are hard to follow. Thus, splits into subcategories should be avoided and only occur where the analysis question requires it. The only exception is when the analyst wants to determine the number of subsets created by a specific category or dimension.

$$\text{RIBBONS} := \frac{|\mathbf{R}|}{|\mathbf{R}^*|} \quad (14)$$

The equation measures the ratio of all exiting ribbons to all possible ribbons, allowing for comparability between different dimension orderings. The effects of minimizing the number of ribbons is shown in Figure 6. A low amount of ribbons reduces clutter.



**Figure 6:** We show three Parallel Sets visualizations for each of the three dimension ordering-dependent metrics: Number of Ribbons, Ribbon Width Variance, and Mutual Information. The results show a reduction of clutter for a reordering of the dimensions, which can serve as a basis for further improvements.

**Ribbon Width Variance**  $\textcircled{W}$  measures the variance of ribbon widths. A low ribbon width variance is preferable, splits that create very small categories should be delayed. Very broad ribbons hide smaller ones. We calculate the standard deviation  $\sigma$  of the ribbon widths, allowing for comparability of different Parallel Sets. To avoid absolute widths, we define  $\text{maxWidth} = \max(\{\text{width}(r) \mid r \in \mathbf{R}\})$ , which we use to normalize the ribbon widths. We formally describe this metric in Equation 15.

$$\text{WIDTHVARIANCE} := \sigma(\{\text{width}(r)/\text{maxWidth} \mid r \in \mathbf{R}\}) \quad (15)$$

The effect is shown in Figure 6. We found that a ribbon with variance can reduce clutter of Parallel Sets, showing that a uniform ribbon width improves readability.

**Mutual Information**  $\textcircled{M}$  measures the average mutual information of neighboring dimension axes. It was first proposed by Shannon [Sha48]. Mutual information measures the dependence between two variables, in the case of Parallel Sets, two neighboring dimensions. It measures the amount of information gained about one variable by observing another variable. Mutual information is formally defined as:

$$\text{MUTUALINFO} := \frac{1}{|D|-1} \sum_{i=1}^{|D|-1} I(C_{D_i}, C_{D_{i+1}}) \quad (16)$$

where  $I(X, Y) := \sum_{x \in X} \sum_{y \in Y} p(x_i, y_i) \log_2 \frac{p(x_i, y_i)}{p(x_i)p(y_i)}$

In this equation,  $p(x, y)$  is the probability of the values  $x \in X$  and  $y \in Y$  occurring together. Since mutual information only measures

the distribution of categories between two dimensions without considering the category ordering, it does not change by reordering categories. Thus, it can only be used to determine an ordering of the dimensions axes. It is used by Dasgupta and Kosara [DK10] in the reordering of parallel coordinate axes and by Alsakran et al. [AHZ\*14] where it is combined with binning or dimensionality reduction. In Figure 6, this metric shows an improvement of readability for high and low values. In general, it should be maximized to improve visualizations.

### 4.3. Combining Quality Metrics

Our metrics can be combined since they measure different aspects of Parallel Sets. Two or more metrics can be minimized or maximized simultaneously, or they can be optimized successively. This especially applies to the combination of a dimension ordering dependent-metric and a category ordering-dependent metric.

The order of categories of an axis in Parallel Sets is the most flexible parameter. Therefore, we are free to maximize or minimize the category ordering for one or multiple of the category ordering-dependent metrics, each reducing different artifacts. They can also be combined in frameworks for the weighting of features [PST\*17].

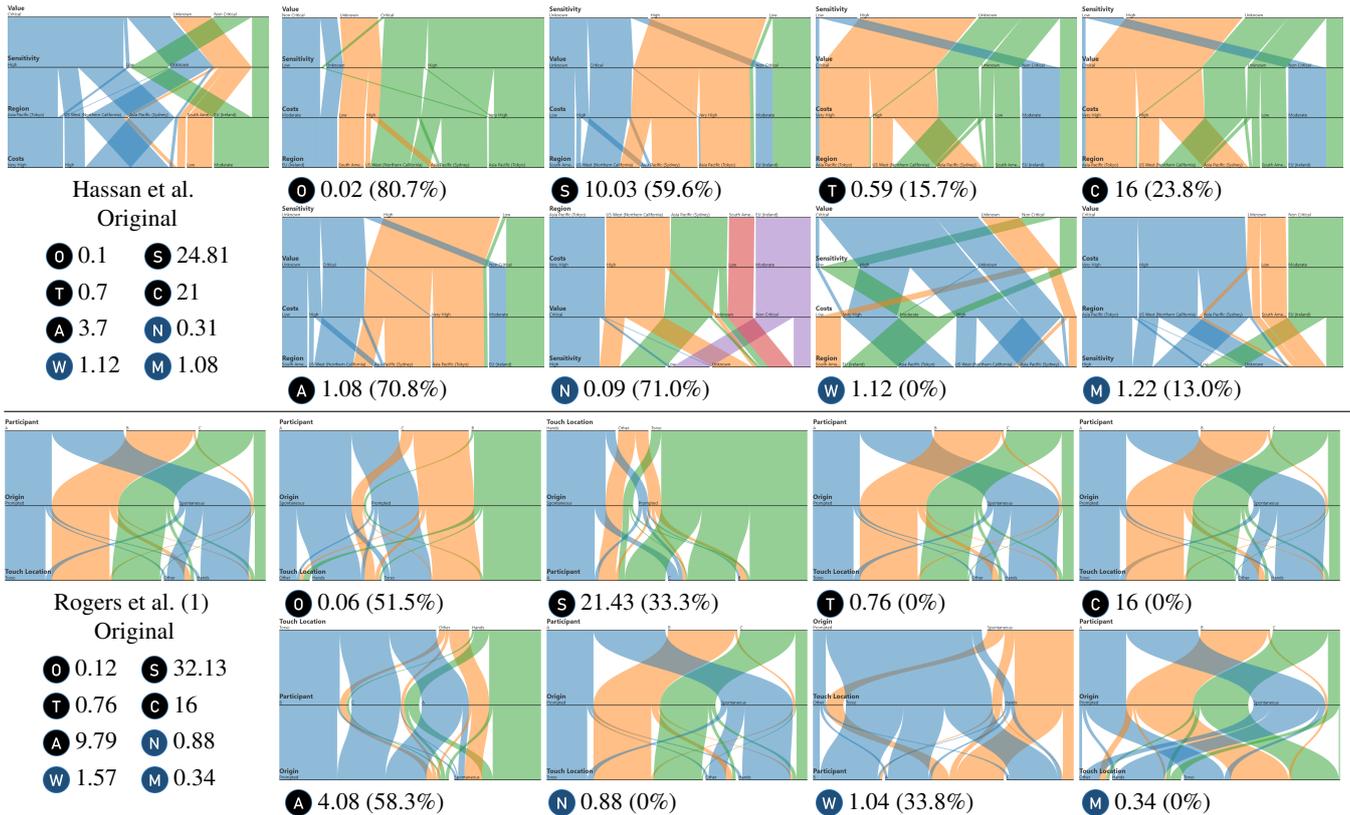
The ordering of dimensions is not as flexible as the ordering of categories. The reasons are: (1) The number of dimensions is usually lower than the number of categories. (2) The categories of the first dimension axis determine the ribbon colors, and thus, the primary target of analysis. (3) All remaining axes split the ribbons into finer and finer subcategories according to their ordering. We suggest minimizing the number of ribbons to reduce the possibility of crossings and overlap. However, this may lead to thin ribbons in the visualization. Alternatively, we propose to reduce the ribbon width variance to avoid excessively thin or broad ribbons, which does not enforce the minimum amount of ribbons. The mutual information metric tries to place related dimensions close to each other, independent of ribbon sizes. We propose the use of those types of metrics as a filtering step.

## 5. Evaluation

To show the effectiveness of our approach, we perform a quantitative evaluation based on visualizations used in previous publications. We perform single-metric and multi-metric optimizations of the Parallel Sets visualizations and conduct a correlation analysis to validate the distinctiveness of our metrics.

### 5.1. Reconstruction of Datasets from Parallel Set Visualizations

To evaluate our approach, we performed a literature search with the terms “Parallel Sets” and “ParSets”. Additionally, we performed a forward search on the foundational publication on Parallel Sets by Bendix et al. [BKH05], and Kosara et al. [KBH06]. Both searches were performed using the digital libraries of ACM, IEEE, and Eurographics. This yields a set of five publications using Parallel Sets listed in Table 1. The Titanic dataset is available online [Daw95]. We reconstructed the other remaining five datasets manually. To this end, we measured the width of the ribbons in the lowest level



**Figure 7:** We show the optimization results for the visualizations provided by Hassan et al. [HP14] and Rogers et al. (1) [RWH\*16] with curved ribbons. The original Parallel Sets visualizations are shown on the left with their metric values. All single-metric optimizations are shown towards the right with the percent improvement compared to the original below. In both cases, all category ordering-dependent metrics have lower clutter. All metrics are lower in comparison to the original Parallel Sets visualization. For the visualization by Hassan et al., the Ribbon Width Variance metric yields the worst result. For Rogers et al. (1), it is the Number of Ribbons metric that performs worst.

to get the size of the subset and traced the ribbon from top to bottom to determine the categories determining the subset. After a visual inspection, we estimate the reconstructions to be mostly accurate. The most challenging to reconstruct was the dataset by Koh et al. [KSDK11] since it contains many small ribbons only one to five pixels wide and a high slope. We choose these datasets because they have published Parallel Sets, implying that they are suitable targets for comparison. To determine the optimized values of our metrics, we calculated all metrics for all possible configurations.

Source	Domain/Desc.	Type	Dim.
[HP14]	Cyber-security	Application	4
[KSDK11]	Property sales	Design-study	3
[KBH06]	Titanic datas. [Daw95]	Technique	4
[RWH*16]	HCI study; 2 datasets	Application	3
[SDB*19]	Linguistics	Application	2

**Table 1:** We found five papers from different domains using Parallel Sets yielding six datasets for our evaluation of ParSetgnostics.

## 5.2. Single-metric Optimization

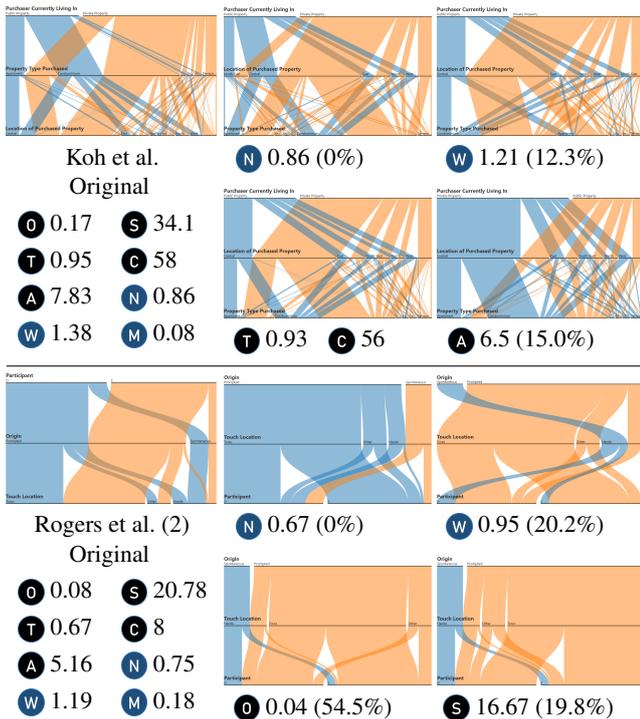
To show the usefulness of each metric, we perform a single-metric optimization on two visualizations using visualizations provided by Hassan et al. [HP14] and Rogers et al. [RWH\*16].

**Hassan et al.** In Figure 7 (top), we perform an optimization using all metrics individually on the Parallel Sets published by Hassan et al. [HP14]. This visualization aims to analyze the security and cost of data storage, determining the location where data storage should be bought with a high security level. We can see that all category ordering-dependent metrics produce visualizations with lower clutter. The *Overlap* **O** metric reduces the overlap of ribbons by 80.7% compared to the original. If we assume the overlap as an objective measure of clutter [ZBD\*18] the *Slope* **S** and *Crossing Angle* **A** metric reduce overlap by 70.8%. These metrics improve by 59.6% and 70.8%. The dimension ordering-dependent metrics reduce clutter as well, with the exception of the *Ribbon Width Variance* **W** metric. All metrics are lower in comparison to the original Parallel Sets, showing that the original visualization was not optimized according to any property of the Parallel Sets. We note that *Slope* **S** and *Crossing Angle* **A** create the same visualizations, as well as *Orthogonality* **T** and *Number of Crossings* **N**.

**Rogers et al. (1)** We perform an optimization using all metrics individually on the Parallel Sets published by Rogers et al. [RWH\*16] showing the more complex dataset of this publication's datasets with curved ribbons. We determine the angles of the ribbons based on the underlying straight ribbons. The task for this visualization is to present the result of a Human-Computer Interaction (HCI) study. The optimization results are shown in Figure 7 (bottom). *Orthogonality* (T), *Number of Crossings* (C), and *Number of Ribbons* (N) are already optimized in the original visualization. Thus, there is no improvement by these metrics. We can see that the *Slope* (S) metric produces large contiguous ribbons and focuses the smaller ribbons in the center. Considering the overlap as a measure of the degree of clutter [ZBD\*18], the *Slope* (S), metric reduces clutter by 50.9% and the *Crossing Angle* (A) metric by 45.1%. The dimension ordering-dependent metrics yield the same ordering for *Number of Ribbons* (N) and *Mutual Information* (M) than the original visualization and thus optimal in those aspects.

### 5.3. Multi-metric Optimization

We evaluate the multi-metric optimization capabilities by selecting the dimension ordering that two out of three dimension ordering-dependent metrics agree on. Based on this ordering, we choose a Parallel Sets visualization according to the metric that improved the most compared to the original visualization.



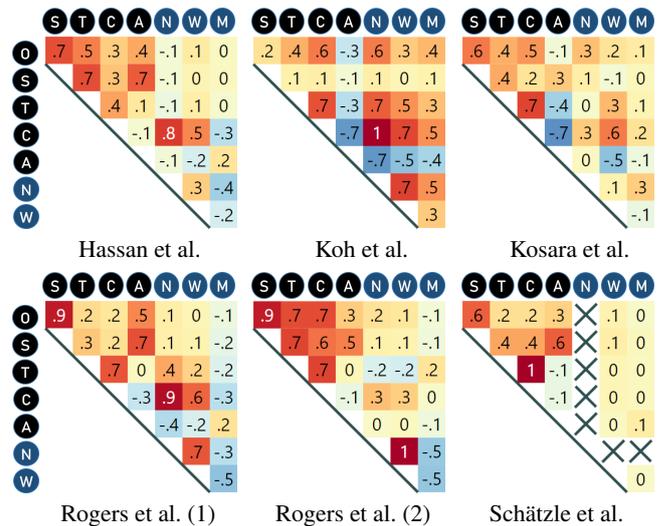
**Figure 8:** At the top, we optimize the dataset supplied by Koh et al. [KSDK11]. On the bottom, we apply our metrics to improve the second visualization of Rogers et al. [RWH\*16]. In both cases, the optimization is based on the dimensions ordering derived from the *Number of Ribbons* and *Ribbon Width Variance* metrics.

**Koh et al.** We perform this optimization using the visualization in the publication by Koh et al. [KSDK11] dealing with property sales analysis. Each step is shown in Figure 8 (top). First, we analyze the dimension ordering. The *Number of Ribbons* (N) and *Ribbon Width Variance* (W) yield the same dimension ordering, while *Ribbon Width Variance* (W) is reduced by 12.4%. For the category ordering we fix the dimension ordering accordingly. We apply all category ordering-dependent metrics to the visualization. We can observe that the *Orthogonality* (T) and *Number of Crossings* (C) yield the identical visualization. By assessing the *Crossing Angle* (A), reducing its value by 15.0% and choosing the overlap as an objective measure for clutter [ZBD\*18] we can see that clutter is reduced by 0.8%. Observing the result, we can also see a cleared-up top level compared to the original.

**Rogers et al. (2)** The visualization presented by Rogers et al. [RWH\*16] describes the result of an HCI study with curved ribbons. We determine the ribbon angles based on the underlying straight ribbon. The steps are shown in Figure 8 (bottom). The *Number of Ribbons* (N) and *Ribbon Width Variance* (W) provide the same dimension ordering. Thus, we only consider layout with this ordering. To determine the order of categories, which influences the appearance of the ribbons. We find that optimum of *Overlap* (O) and *Slope* (S) have the dimension ordering as suggests by the dimension ordering-dependent metrics. The visualization suggested by the *Slope* (S) metric reduces the clutter by 53.2% considering overlap as an objective measure [ZBD\*18]. This visualization focuses all splits and crossings on one the left half of the visualization.

### 5.4. Correlation Analysis

In order to evaluate that our metrics measure different properties of Parallel Sets we performed a Pearson correlation analysis [Kir08]



**Figure 9:** Results of the correlation analysis of the metrics for all reconstructed datasets. We found that no metric correlates with any other metric for all analyzed datasets. This shows that all metrics are independent and measure distinct properties.

of the metrics. We calculated the value of all metrics for all dimension and category layout for all available datasets. The results of the analysis are summarized in Figure 9. The metric *Crossing Angle* (A) shows a weak negative correlation for the Koh et al. [KSDK11] dataset and *Mutual Information* (M) shows a weak negative correlation for the Rogers et al. (1) [RWH\*16] dataset. The *Number of Ribbons* (N) metric could not be analyzed for the data by Schätzle et al. [SDB\*19] because it only has two dimensions and thus a fixed number of ribbons for all configurations. The correlation analysis shows that the correlations between metrics is dependent on the dataset. This is shown by the differing Pearson correlations. Figure 9 provides the correlations between the metrics for all datasets. We found that no metric correlates with any other metric for all analyzed datasets. This shows that all metrics are independent and measure distinct properties, and are mutually independent.

## 6. Discussion

The calculation of all quality metrics is dependent on the number of ribbons of a Parallel Sets visualization. All metrics are described in terms of vector graphics. Our metrics can be applied before the ribbons are curved since the straight ribbons approximate the properties of the curved ribbons. All dimension ordering-dependent metrics are directly applicable since they are not dependent on the ribbon shape. All category ordering-dependent metrics, except the *Number of Crossings* metric, will provide an approximate result, which can improve the visualization. All quality metrics, except the angle-related metrics (i.e., *Slope*, *Orthogonality*, and *Crossing Angle*) can be applied to Common Angle Plots directly since they enforce the angle a ribbon has in-between two dimension axes. Our metrics can be used to measure the quality increase or decrease in cases where the underlying data changes. This is also true for streaming scenarios, where new categories might be encountered. However, determining an optimal ordering of dimensions and categories would require a more efficient optimization strategy, other than calculating the metrics for all possible configurations. The metrics are calculated reasonably fast, such that in an interactive design process, they can be used to compare and rank different manually created Parallel Sets visualizations instantly. Our correlation analysis shows that all metrics measure distinct properties and thus are mutually independent. We derive the set of metrics from our discussion on parameters of Parallel Sets related to the choice of the category and dimension ordering, dataset-dependent properties, and ribbon parameters. Our metrics address all parameters, and thus, we argue for completeness in terms of geometric properties. We plan a user study as an additional validation of completeness.

### 6.1. Guidelines

We found the following design guidelines for the layout of dimensions and categories of Parallel Sets visualizations. (1) Choose the first dimension according to the analysis question or well-known categories. In exploratory tasks, choose a dimension with a category count no larger than nine. We suggest following Miller's Law, which states to limit the number of shown items to seven plus or minus two [Mil56]. We also suggest choosing a dimension with a high entropy leading to equal-sized categories. (2) Filter the set of all configurations by dimension ordering-dependent metrics. These

metrics can be used in a voting system as we do in Subsection 5.3. (3) Minimize/Maximize a category ordering-dependent metric. In our experiments, we found some suggestions: Parallel Sets with a low number of ribbon splits, i.e., a low number of ribbons in the lower levels of Parallel Sets show better results when optimized with the *Overlap* and *Slope* metrics. Parallel Sets with a high number of ribbons are optimized with the *Orthogonality*, *Number of Crossings* and *Crossing Angle*. Curved ribbons are easier to read. This is based on the fact that curved lines have a larger crossing angle, which makes lines easier to follow [HHE08, WPCM02].

### 6.2. Limitations and Future Work

Our metrics quantify the visual appearance of Parallel Sets. They do not provide a reordering strategy. The next step is to assess the properties of our metrics and derive a reordering algorithm. Another possible direction is an extension towards local metric descriptors since our metrics only describe Parallel Sets globally. We plan to study the connection between specific metrics with general tasks and data set characteristics through a user-study. A user study would also verify whether the set of metrics is exhaustive. This work does not describe an efficient strategy to determine the minimum and maximum value of a metric. Additionally, we plan to study the effects of the metrics in the interactive design of Parallel Sets suggesting and validating user choices. One drawback of our approach is that the metrics need to be recalculated if the aspect ratio of the plot changes. In the case of simple zooming with a fixed aspect ratio, the values can be reused. Our quality metrics could potentially be transferred to the quantification of properties of Sankey diagrams since many desirable properties of Parallel Sets are also desirable for Sankey diagrams, e.g., a low overlap of bands.

## 7. Conclusion

Determining a useful dimension and category ordering for Parallel Sets is challenging. We propose a set of eight distinct quality metrics for Parallel Sets, called *ParSetgnostics*. They provide a new model for quantifying properties of Parallel Sets visualizations, which can be used as a quality criterion as described by Behrisch et al. [BBK\*18]. Our metrics allow us to improve the readability of Parallel Sets visualizations by optimizing a specific metric or a combination thereof or even determining the presence of undesirable patterns. We argue for our metrics' effectiveness by applying them to Parallel Sets in previous publications, showing their applicability in a single- and multi-metric optimization approach. We perform a correlation analysis on all datasets and quality metrics combinations and validate that no metric correlates with any other metric for all datasets, showing each metric's distinctiveness. We published the results online where users can explore our results and test the quality metrics' properties interactively. Our work provides a more meaningful way to analyze categorical data with Parallel Sets, especially in exploratory scenarios.

### Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A03).

## References

- [AHZ\*14] ALSAKRAN J., HUANG X., ZHAO Y., YANG J., FAST K.: Using entropy-related measures in categorical data visualization. In *IEEE Pacific Visualization Symposium* (2014), pp. 81–88. doi:10.1109/PacificVis.2014.43.2,7
- [BBH\*17] BEHRISCH M., BACH B., HUND M., DELZ M., VON RÜDEN L., FEKETE J., SCHRECK T.: Magnostics: Image-based search of interesting matrix views for guided network exploration. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 31–40. doi:10.1109/TVCG.2016.2598467.3
- [BBK\*18] BEHRISCH M., BLUMENSCHNEIN M., KIM N. W., SHAO L., EL-ASSADY M., FUCHS J., SEEBACHER D., DIEHL A., BRANDES U., PFISTER H., SCHRECK T., WEISKOPF D., KEIM D. A.: Quality metrics for information visualization. *Computer Graphics Forum* 37, 3 (2018), 625–662. doi:10.1111/cgf.13446.3,10
- [BHH03] BREWER C. A., HATCHARD G. W., HARROWER M. A.: Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science* 30, 1 (Jan 2003), 5–32. doi:10.1559/152304003100010929.4
- [BKH05] BENDIX F., KOSARA R., HAUSER H.: Parallel sets: Visual analysis of categorical data. In *IEEE Symposium on Information Visualization* (2005), pp. 133–140. doi:10.1109/INFVIS.2005.1532139.2,4,7
- [BZP\*20] BLUMENSCHNEIN M., ZHANG X., POMERENKE D., KEIM D. A., FUCHS J.: Evaluating reordering strategies for cluster identification in parallel coordinates. *Computer Graphics Forum* 39, 3 (2020), 537–549. doi:10.1111/cgf.14000.2,6
- [CD18] CAVALLO M., DEMIRALP Ç.: A visual interaction framework for dimensionality reduction based data exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), Mandryk R. L., Hancock M., Perry M., Cox A. L., (Eds.), ACM, p. 635. doi:10.1145/3173574.3174209.1
- [Cle86] CLEVELAND W. S.: The elements of graphing data. *The American Journal of Psychology* 99, 3 (1986), 443–448. doi:10.2307/1422498.1
- [Daw95] DAWSON R. J. M.: The "unusual episode" data revisited, 1995. <http://jse.amstat.org/v3n3/datasets.dawson.html>, last accessed 2020-09-18. 1,2,5,6,7,8
- [DK10] DASGUPTA A., KOSARA R.: Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1017–1026. doi:10.1109/TVCG.2010.184.3,6,7
- [DS91] DAY R. H., STECHER E. J.: Sine of an illusion. *Perception* 20 (1991), 49–55. doi:10.1068/p200049.2
- [ED06] ELLIS G. P., DIX A. J.: Enabling automatic clutter reduction in parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 717–724. doi:10.1109/TVCG.2006.138.6
- [Gol14] GOLDSTEIN E. B.: *Sensation and Perception*, 9th ed. Wadsworth Cengage Learning, 2014. 2
- [Har75] HARTIGAN J. A.: Printer graphics for clustering. *Journal of Statistical Computation and Simulation* 4, 3 (1975), 187–213. doi:10.1080/00949657508810123.1
- [HD12] HEALEY C. G., DENNIS B. M.: Interest driven navigation in visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 10 (2012), 1744–1756. doi:10.1109/TVCG.2012.23.2
- [HHE08] HUANG W., HONG S., EADES P.: Effects of crossing angles. In *IEEE VGTC Pacific Visualization Symposium* (2008), IEEE Computer Society, pp. 41–46. doi:10.1109/PACIFICVIS.2008.4475457.4,5,6,10
- [Hof00] HOFMANN H.: Exploring categorical data: interactive mosaic plots. *Metrika* 51, 1 (July 2000), 11–26. doi:10.1007/s001840000041.1
- [HP14] HASSAN S., PERNUL G.: Efficiently managing the security and costs of big data storage using visual analytics. In *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services* (2014), pp. 180–184. doi:10.1145/2684200.2684333.8
- [HV13] HOFMANN H., VENDETTUOLI M.: Common angle plots as perception-true visualizations of categorical associations. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2297–2305. doi:10.1109/TVCG.2013.140.2
- [Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer* 1, 2 (August 1985), 69–91. doi:10.1007/bf01898350.1,2
- [JS98] JERDING D. F., STASKO J. T.: The information mural: A technique for displaying and navigating large information spaces. *IEEE Transactions on Visualization and Computer Graphics* 4, 3 (1998), 257–271. doi:10.1109/2945.722299.1
- [KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics* 12, 4 (2006), 558–568. doi:10.1109/TVCG.2006.76.1,7,8
- [Kir08] KIRCH W. (Ed.): *Pearson's Correlation Coefficient*. Springer Netherlands, Dordrecht, 2008, pp. 1090–1091. doi:10.1007/978-1-4020-5614-7\_2569.9
- [Kos10] KOSARA R.: Turning a table into a tree: Growing parallel sets into a purposeful project. In *Beautiful Visualization: Looking at Data through the Eyes of Experts* (2010), Steele J., Iliinsky N., (Eds.), O'Reilly Media, pp. 193–204. 2,3
- [KS98] KENNEDY A. B. W., SANKEY H. R.: The thermal efficiency of steam engines. *Minutes of the Proceedings of the Institution of Civil Engineers* 134 (1898), 278–312. doi:10.1680/imotp.1898.19100.2
- [KSDK11] KOH L. C., SLINGSBY A., DYKES J., KAM T. S.: Developing and applying a user-centered model for the design and implementation of information visualization tools. In *15th International Conference on Information Visualisation, IV 2011, London, United Kingdom* (2011), pp. 90–95. doi:10.1109/IV.2011.32.8,9,10
- [LKZ\*15] LEHMANN D. J., KEMMLER F., ZHYHALAVA T., KIRSCHKE M., THEISEL H.: Visualnostics: Visual guidance pictograms for analyzing projections of high-dimensional data. *Computer Graphics Forum* 34, 3 (2015), 291–300. doi:10.1111/cgf.12641.3
- [Mil56] MILLER G. A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* (1956), 81–97. 5,10
- [MJSK15] MITTELSTÄDT S., JÄCKLE D., STOFFEL F., KEIM D. A.: Colorcat: Guided design of colormaps for combined analysis tasks. In *17th Eurographics Conference on Visualization, EuroVis 2015 - Short Papers* (2015), Bertini E., Kennedy J., Puppo E., (Eds.), Eurographics Association, pp. 115–119. doi:10.2312/eurovisshort.20151135.4
- [PDK\*19] POMERENKE D., DENNIG F. L., KEIM D. A., FUCHS J., BLUMENSCHNEIN M.: Slope-dependent rendering of parallel coordinates to reduce density distortion and ghost clusters. In *30th IEEE Visualization Conference* (2019), IEEE, pp. 86–90. doi:10.1109/VISUAL.2019.8933706.4,6
- [PST\*17] PAJER S., STREIT M., TORSNEY-WEIR T., SPECHTENHAUSER F., MÖLLER T., PIRINGER H.: Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 611–620. doi:10.1109/TVCG.2016.2598589.7
- [RRBW03] ROSARIO G. E., RUNDENSTEINER E. A., BROWN D. C., WARD M. O.: Mapping nominal values to numbers for effective visualization. In *9th IEEE Symposium on Information Visualization* (2003), IEEE Computer Society, pp. 113–120. doi:10.1109/INFVIS.2003.1249016.1

- [RWH\*16] ROGERS K., WILES J., HEATH S., HENSBY K., TAUFATOFUA J.: Discovering patterns of touch: A case study for visualization-driven analysis in human-robot interaction. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (2016), pp. 499–500. doi:10.1109/HRI.2016.7451825. 2, 8, 9, 10
- [SB03] SPENKE M., BEILKEN C.: Visualization of Trees as Highly Compressed Tables with InfoZoom. In *Proceedings of the IEEE Symposium on Information Visualization* (2003), pp. 122–123. 1
- [SDB\*19] SCHÄTZLE C., DENNIG F. L., BLUMENSCHEN M., KEIM D. A., BUTT M.: Visualizing linguistic change as dimension interactions. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (Florence, Italy, August 2019), Association for Computational Linguistics, pp. 272–278. doi:10.18653/v1/W19-4734. 8, 10
- [Sha48] SHANNON C. E.: A mathematical theory of communication. *Bell System Technical Journal* 27, 3 (1948), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x. 2, 4, 7
- [SSK06] SCHNEIDEWIND J., SIPS M., KEIM D. A.: Pixnostics: Towards measuring the value of visualization. In *IEEE Symposium On Visual Analytics Science And Technology, IEEE VAST 2006, October 31-November 2, 2006, Baltimore, Maryland, USA* (2006), pp. 199–206. doi:10.1109/VAST.2006.261423. 3
- [TM03] TEOH S. T., MA K.: Paintingclass: interactive construction, visualization and exploration of decision trees. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003), Getoor L., Senator T. E., Domingos P. M., Faloutsos C., (Eds.), ACM, pp. 667–672. doi:10.1145/956750.956837. 1
- [WAG05] WILKINSON L., ANAND A., GROSSMAN R. L.: Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization* (2005), pp. 157–164. doi:10.1109/INFVIS.2005.1532142. 3
- [WLHS01] WITTENBURG K., LANNING T., HEINRICHS M., STANTON M.: Parallel bargrams for consumer-based information exploration and choice. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (2001), Marks J., Mynatt E. D., (Eds.), ACM, pp. 51–60. doi:10.1145/502348.502357. 1
- [WPCM02] WARE C., PURCHASE H. C., COLPOYS L., MCGILL M.: Cognitive measurements of graph aesthetics. *Information Visualization* 1, 2 (2002), 103–110. doi:10.1057/palgrave.ivs.9500013. 4, 5, 6, 10
- [ZBD\*18] ZARATE D. C., BODIC P. L., DWYER T., GANGE G., STUCKEY P. J.: Optimal sankey diagrams via integer programming. In *IEEE Pacific Visualization Symposium* (2018), IEEE Computer Society, pp. 135–139. doi:10.1109/PacificVis.2018.00025. 8, 9
- [ZCYY19] ZHANG C., CHEN Y., YANG J., YIN Z.: An association rule based approach to reducing visual clutter in parallel sets. *Visual Informatics* 3, 1 (2019), 48–57. doi:10.1016/j.visinf.2019.03.006. 2