CatNetVis: Semantic Visual Exploration of Categorical High-Dimensional Data with Force-Directed Graph Layouts

Michael Thane[1], Kai M. Blum[1] and Dirk J. Lehmann[1]

[1]Institute for Information Engineering, Ostfalia University of Applied Sciences,
Brunswick-Wolfenbuttel, Germany

Abstract

We introduce CatNetVis, a novel method of representing semantical relations in categorical high-dimensional data. Traditional methods provide insights into many aspects of visual exploration of data. However, most of them lack information on relations in between categories or even clusters of categories. The force-directed network layout utilized by CatNetVis enables a lightweight approach in order to explore such semantical relations. The connections within the network are perceived as an intuitive metaphor for clusters of connections/relations in categorical data denoted as communities. While the user interacts, visual encodings such as information about the entropy and frequencies allow a fast perception of relation between categories and its frequencies, respectively. We illustrate how CatNetVis performs as an effective addition to traditional methods by demonstrating the method on an example data sets and comparing it to conventional methods.

1. Introduction

Drawing conclusions is a vital task in data analysis and involves understanding semantic relations within categorical data. Such data is part of many domains and consists of nominal, ordinal or grouped data, such as medical records of patients or attributes of phenotypes for animals/plants. Visually analysing different distributions of dimensions and their mutual relations can be tedious, as the number of combinations (of/and relations) grows exponentially with the number of dimensions. Over the last decades, several visualization methods have been developed to address this issue of visually analysing high-dimensional categorical data. We describe categorical data using the terms dimension and category. A dimension is an attribute/feature of our high-dimensional data set, and a category is a set of indices which has a certain value.

visualization of such data usually involves two main aspects: (i) visualization of frequencies and (ii) relation of categories. Unfortunately, when visualizing relations in high-dimensional categorical data, clutter is introduced immediately as different categories might overlap [AHZ*14], making this visualization issue even harder to address. This is why frequently used approaches, such as *Mosaic Plots* [HK81] (MPS) or *Parallel Set Plots* (PSP) [KBH06], are limited to a smaller number of dimensions. Beyond that, visualizing a large amount of semantic relations between high-dimensional data at once needs a suitable layout, appropriate aggregation schemes, and intuitive interaction methods for the analyst.

Moreover, from the best of our knowledge, the visual analysis of semantic relations in between cluster/groups of categories are usually not supported, even though a semantical consideration might give new insights in the underlying data domain for an interested domain expert. This motivates us to introduce *CatNetVis*, an interactive visualization scheme that support to identify semantic relations amongst categories & communities (clusters) of categories, respectively. In our scheme, categories themselves are mutually aggregated and grouped by a similarity remove to work out the intrinsic and inherent semantic structure of the categorical data. In order to visualize such a similarity structure (by using a proximity metaphor), our approach is inspired by force-directed graph layout [FR91] used in network analysis.

This paper is organized into seven sections. Section 2 covers related work on categorical data visualization, Section 3 explains the necessary preprocessing, and Section 4 describes the visualization tool, *CatNetVis* and the visual representation of information. In Section 5 we show an example of our visualization approach and compare it with conventional methods for visualizing categorical data using a data set about arthritis treatment [Fri01]. In Section 6 we use a data set with higher dimensionality and demonstrate the exploration capabilities on a demographic data set. In Section 7 we discuss our results and the limitations of our method and draw conclusions for future work.

2. Related Work

There are different approaches known to visually analyse categorical data. *Parallel Sets Plot* (PSP) is an established method that has been frequently applied in recent years [FJ11, TEL16, ZCYY19, DFB*21, ML22]. PSP displays individual categories on vertical lines, similar to *Parallel Coordinates Plot* (PCP) for continuous data [Ins85], as illustrated in Fig. 2 (B). While PSP uses the width of the lines to show the frequency of each category, PCP uses the actual position of the line to show the value of each variable. Therefore, PSP reveals patterns and associations on categorical data, while PCP reveals trends and outliers on continuous data. PSP can be ordered vertically by the values and horizontally by the dimensions which creates various ways of displaying the data.

Upton [Upt00] present a *Cobweb Diagram* for displaying categorical data by drawing a graph with connections between related categories of different dimensions. Instead of only showing the count of the marginal table Upton displays associations between the categories by the standardised residuals of the independence model. Apart from [Upt00], little attention has been paid to network layouts, which have the advantage of arranging the nodes in such a way that connected nodes are displayed close to each other.

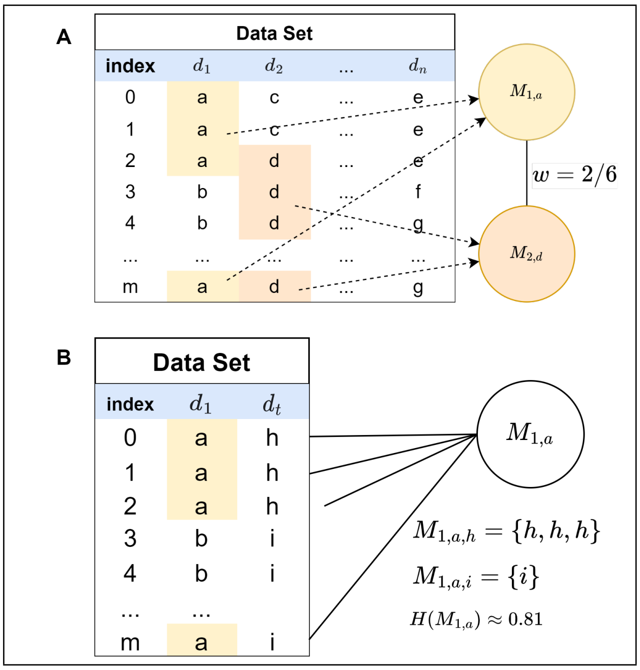
In that regard, Alsallakh et al. [AGMS11, AAMG12] introduce a technique called *Contingency Wheel*. The wheel visualizes categorical dimensions by displaying them in a circular network layout. The facets of dimensions include the histogram of frequencies from the contingency table. The visualization reveals associations between row and column categories and is capable of analysing contingency tables at a large scale while showing single categories characteristics. The approach is not capable of providing relation in between cluster of categories (=communities), as our approach is capable of.

Mosaic Plots (MP) [HK81] represents contingency tables with rectangles whose size is proportional to the frequencies. Thereby, each cell in the table is represented as a bin in the plot. The bin size is proportional to the number of cases in a cell. This allows a quick first impression of the contingency table. MPs are limited to a few dimensions [Fri94, Upt00] and are difficult to interpret for more than two dimensions.

An alternative approach by projecting categorical data in a 2D plane is multiple correspondence analysis (MCA) [Gre17], which can be seen as the categorical counterpart to principal component analysis (PCA) [Dun89] for continuous data. MCA uses a Burt table, that can be seen as similar to a covariance matrix for continuous data, in order to project data onto e.g. two dimensions which are the first two eigenvalues of the Burt table. The method is difficult to interpret for non-experts and does not show information on frequencies. Subsequently, our pre-steps will be explained.

3. Preprocessing

In the following, we explain the mathematical notation used in this work and guide through the steps of preprocessing which have to be performed before we are able to create a visual representation. We note a bold uppercase letter **D** a n x m data set of categorical data and a lowercase letter $d = (v_1, ..., v_m)^T$ a dimension of the data set. We handle mixed data by using quantile-based discretization of the continuous dimensions. Then, the data set **D** is set to be $\mathbf{D} = (d_1, d_2, ...d_n)$. Within a dimension d, each unique occurrence can be noted as a category of index k. Instances of a dimension d_i with the same value v_k are noted as category $M_{i,k}$. Our algorithm creates these sets of indices for all possible dimensions. In that regard, our scheme calculates similarities for each pair of categories from different dimensions $M_{i,k}$ and $M_{j,l}$ with $i \neq j$ and $k \neq l$ by using the Jaccard-Index as $\mathbf{J}(M_{i,k}, M_{j,l}) = \frac{|M_{i,k} \cap M_{j,l}|}{|M_{i,k} \cup M_{j,l}|}$.



index	d_1	d_2	...	d_n
0	a	c	...	e
1	a	c	...	e
2	a	d	...	e
3	b	d	...	f
4	b	d	...	g
...	...	...	...	...
m	a	d	...	g

index	d_1	d_t
0	a	h
1	a	h
2	a	h
3	b	i
4	b	i
...	...	
m	a	i

Figure 1: *(A) Building the graph G from the values of the dimensions, (B) Calculation of entropy $H(M_{i,k})$*

Based on this similarity, our scheme creates visual clusters of categories, in MCA sometimes also called point clouds. We denote them as *communities* (=clusters of categories). We aim to represent these communities in a graph **G** which consists of vertices V, edges E and a weight function w, illustrated by Fig. 1(A).

The categories $M_{i,k}$ resemble the vertices of a graph $\mathbf{G} = (V, E, W)$, where $V = \{M_{i,k} \forall d_i \in D, \forall v_k \in d_i\}$. The edges of that graph are drawn between all categories that are not from the same dimension: $E = \{e_{i,j,k,l} = (M_{i,k}, M_{j,l}) | \forall d_i, d_j \in \mathbf{D}, \forall c_k \in d_i, c_l \in d_j, d_i \neq d_j\}$. The weight of each edge is given by the Jaccard-Index $\mathbf{J}(M_{i,k}, M_{j,l})$ between the categories $M_{i,k}$ and $M_{j,l}$.

Our approach visualizes the distribution of a certain class dimension in the data set. This gives an additional insight on how certain dimensions are predictors of the class. Because all the dimensions e categorical, the class dimension can be any of the dimensions. Given a category $M_{i,k}$ and a class dimension d_t, the entropy of the frequencies for the instances belonging to the class is calculated as $\mathbf{H}(M_{i,k}) = -\sum_{v_i \in d_t} \frac{|M_{i,k,t}|}{|M_{i,k}|} \cdot log_{|d_t|}(\frac{|M_{i,k,t}|}{|M_{i,k}|})$ with $M_{i,k,t}$ being a category of the attribute belonging to a class label t. Fig. 1 (B) shows an example on how the entropy is calculated for the value a of dimensions d_1.

4. Visual Representation and Interaction

For the positioning of the nodes based on their Jaccard Similarity we use a force-directed graph layout. The algorithm begins with

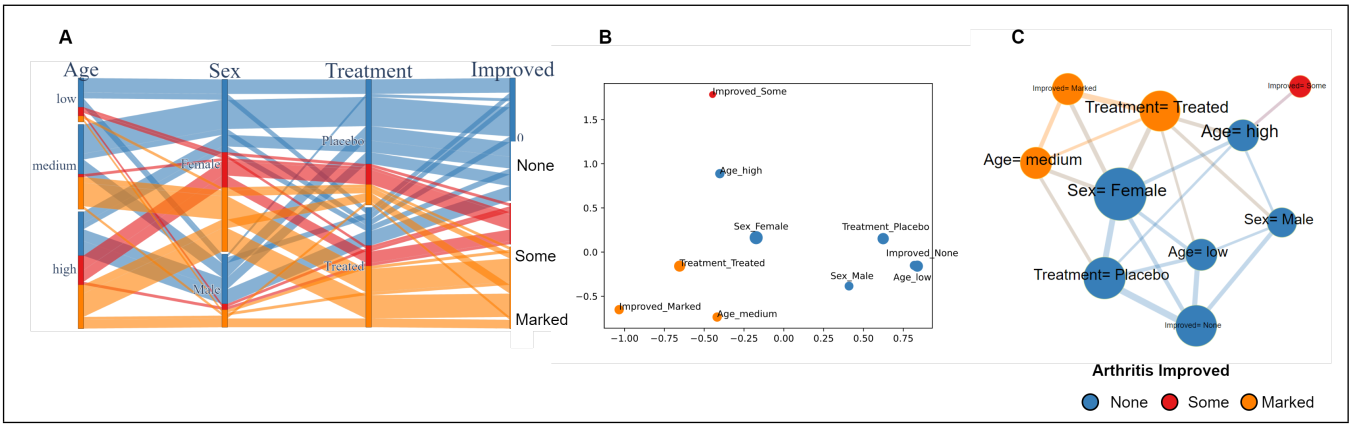


Figure 2: *Comparison of PSP, MCA and CatNetVis using the arthritis data set [Fri01] (A) PSP (B) MCA (C) CatNetVis, The colour shows the improvement of arthritis, blue: None, red: Some, orange: Marked.*

initializing the nodes at random positions, then calculate the forces and updates the positions iteratively until certain criteria (e.g. minimum y is reached. We use the D3.js implementation which uses Verlet integration [Ver67], Barnes-Hut approximation [BH86] and a constrained graph layout [Dwy09]. The algorithm uses a pseudo-gravity to keep the unconnected sub-graphs from getting repelled outside the layout. The information of our visualization is encoded as follows:

- Position: calculated iteratively with the force-directed graph layout. The position of the nodes is merely a result of that and does not result from either MCA, MDS or other projection methods.
- Colour: The colour of a certain node is given by the colour code of the majority class, the class with the most entries
- Size: The size of the nodes is proportional to the number of elements that belong to the value v_{ij}.
- Label: The label shows the value $d_i = v_{ij}$ with its corresponding dimension d_i. The higher the entropy of the node, the higher the font size of the label
- Edge colour: The colour of the edges is a $50:50$ mixture of the source and the target node
- Edge line width: The line width of the edge is proportional to the Jaccard Similarity between the source node and the target node
- Edge line length: Is a result of the force layout and does not encode any information itself

Since the graph is fully connected, the vertices assemble in the middle of the layout, as they all have attracted forces that hold them together. This makes the categories and their relationships harder to read, as edges block the sight. In order to prevent cluttering, it is possible to filter edges based on their weights. The idea behind this is, that vertices that have a low entropy are more reliable when drawing conclusions based on the associations. On the other side, it can help to filter for high entropy values which show vertices with more complex relationships amongst each other. [AHZ*14] To keep track of the impact by a single category it is possible to filter based on the relative size within the dataset. Thus, categories with a small or a high number of instances can be hidden to get a better insight on major or minor categories within the data set.

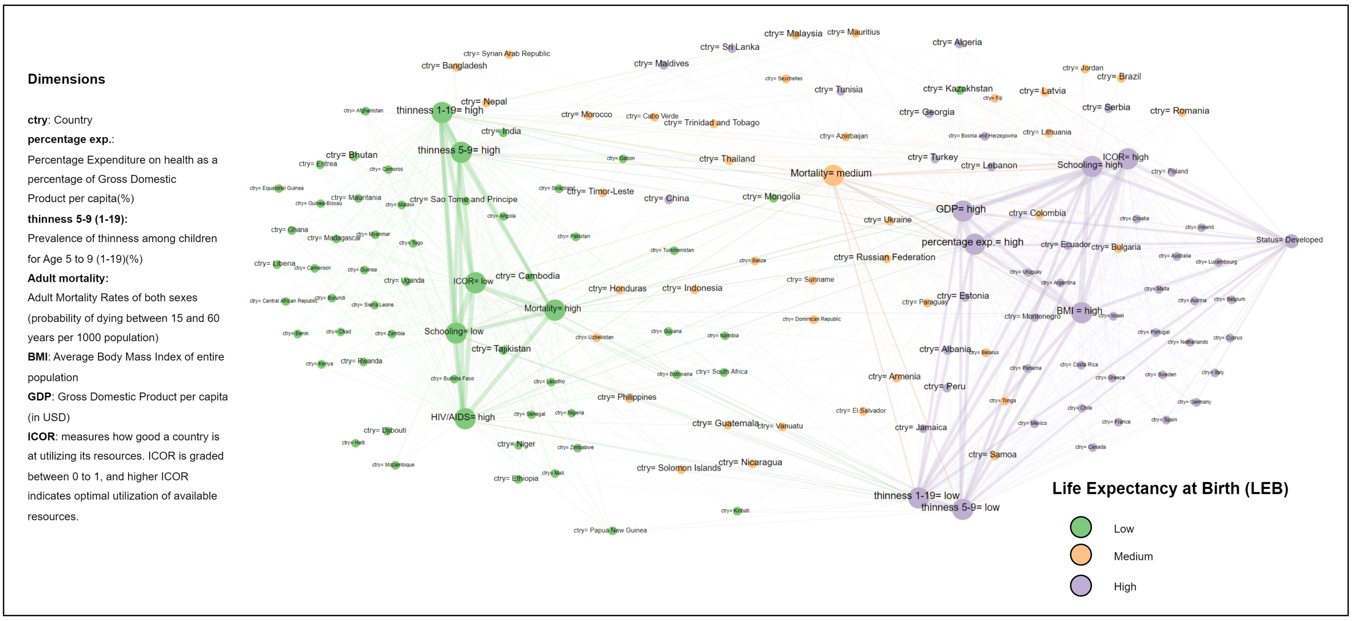
5. Comparison with Multi Correspondence Analysis and Parallel Sets Plot

In Fig. 2 we used an arthritis medical data set [Fri01] which consists of 84 patients with 4 dimensions, *Treatment*, *Sex*, *Age*, *Improved*. The *Treatment* dimensions have the values *Treatment* and *Placebo* and the *Improved* dimension has the values *None*, *Some* and *Marked*. *None* means that the patient did not have any change in the arthritis symptoms, *Some* indicates some improvement and the patients who belong to the *Marked* group show a significant improvement in their symptoms.

In Fig. 2 (A) we can observe a *PSP* of the dataset which has the class dimension *Improved* as the last axis. The *None* group is mostly associated with low or medium age and placebo treatment. When comparing male and female gender one can see that amongst the male individuals the majority has of them showed no improvement. The *Some* group consist mostly of high aged female individuals where a majority received the placebo treatment. The *Marked* group consists of individuals with medium or high age, which are mostly female and are treated. In Fig. 2 (B) we display the MCA plot of the data set. For easy comparison we used a similar encoding like in the *CatNetVis* approach. The colour of the nodes are the majority class value for the category and the size is proportional to the number of elements. The individuals that are treated also have a marked outcome regarding their symptoms which indicates that the treatment was effective. The group that received the placebo showed no effect, because the value of *Improved_None* is close to *Improved_Placebo*. In Fig. 2 (C) we show *CatNetVis* approach of the same data set, where the edges are filtered with a weight higher than 0.2. This results in a graph that is no longer fully connected. We can observe that the topology of the node positions is similar to the MCA in Fig. 2 (B). The values *Improved=Marked*, *Treatment=Treated* and *Age=Medium* are close to each other as well as the three values *Improved=None*, *Treatment=Placebo*, *Age=low*, *Sex=Male*. In both (A) and (C) the value *Sex=Female* is centered in both layouts and the *Improved=Some* group is a bit far off from the rest of the values and its nearest neighbour is *Age=high*.

6. Case Study: Life Expectancy in Different Countries

Next we evaluated how the MCA projection and *CatNetVis* compare when there is a higher number of dimensions. We excluded the


Dimensions

ctry: Country
percentage exp.: Percentage Expenditure on health as a percentage of Gross Domestic Product per capita(%)
thinness 5-9 (1-19): Prevalence of thinness among children for Age 5 to 9 (1-19)(%)
Adult mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
BMI: Average Body Mass Index of entire population
GDP: Gross Domestic Product per capita (in USD)
ICOR: measures how good a country is at utilizing its resources. ICOR is graded between 0 to 1, and higher ICOR indicates optimal utilization of available resources.

Figure 3: *CatNetVis using the life expectancy data set of the WHO [Wor21] filtered with $H(M_{i,k} < 0.8)$.*

PSP from this comparison as it is not applicable for a high number of dimensions, because of the visual clutter introduced by many line intersections. In Fig. 3 we show the CatNetVis plot. We used the life expectancy data set from the WHO [Wor21] in order to test the approach for a multidimensional data set on demographic data of 93 different countries. The data set consists of 22 dimensions, where 2 are categorical dimensions and 20 numeric dimensions. We filtered nodes of entropy below 0.8 and visualized the data set using *CatNetVis*. We can see that two main communities are forming. On the left appear the development countries, which are associated with low schooling, high mortality and starvation (thinness 5-9 (1-19) = high). Also HIV/AIDS seems to be prevalent amongst these countries. On the right we can see the development countries which are connected to the categories *GDP=high*, *Schooling=high* or *ICOR=high* and *Status=Developed*. The connections between the categories of the countries is helping on understanding the relationship between the demographic attributes of the countries.

7. Discussion and Conclusions

From our layout we can gather information about the overall frequencies of different categories, how the class dimension is distributed and look at the relations of categories in detail. The force directed graph drawing allows detecting communities of connected categories, since similar categories attract each other. Clutter within the visualization is reduced because of automatic ordering, filtering options and hiding of the edges.

We applied our method to different example data sets. In Fig. 2 we compared the commonly used PSP approach with the MCA and CatNetVis. Even though PSP [KBH06] are a good way of visualizing the bifurcation of categories in different dimensions, they do not show the direct relationship between the categorical dimensions. In our additional material (Figure 1) we show how our method can be useful for classifying poisonous/edible mushrooms.

In PSP, both the order of the vertical placement of the categories and the horizontal placement of the dimensions influence the effectiveness of the algorithm. The visualization can become cluttered when there are too many values in each dimension [KBH06], and it can become difficult to interpret when there are too many dimensions, because of the cognitive effort of following the different lines. For higher dimensionality the large number of polylines will result in individual categories [AHZ*14]. Often, dimension reduction is necessary in order to use PSP.

Unlike PSP our approach has no need for ordering of the dimensions and categories. We use a single view with interaction possibilities such as zooming and hovering. It has been shown by Fernstadt et al. [FJ11] that single view approaches are better for analysis of high-dimensional categorical data. In MPs, it is necessary to use legend information in order to understand relationships between different categories. In our approach categories are directly annotated to vertices which makes their exploration easy. The MCA method is a good way of analysing the global relations of values in categorical data. It gives the analyst an overview about the high-dimensional data set, but it has similar drawbacks as the PCA. The MCA is sensitive to outliers and assumes linear relationships amongst the dimensions. Also, there are effects of distortion, also known as the horseshoe effect which are artefacts of the method that influence readability. This also leads to overlapping labels of the categories. This problem is reduced in CatNetVis due to the repelling forces. For reference, we show an MCA plot in the additional material (Figure 2). One limitation of our method is that the edges might introduce visual clutter, which can be addressed by using edge bundling techniques. In addition, our scheme also provides to hide the edges if the users want it.

In the future, we are going to extend the approach and link MCA or PSP to CatNetVis. Also, we want to use interaction methods such as brushing in order to select and visualize groups of categories and evaluate graph clustering methods to detect communities.

References

[AAMG12] ALSALLAKH B., AIGNER W., MIKSCH S., GRÖLLER M. E.: Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *IEEE Transactions on Visualization and Computer Graphics* (2012). 2

[AGMS11] ALSALLAKH B., GROELLER E., MIKSCH S., SUNTINGER M.: Contingency Wheel: Visual Analysis of Large Contingency Tables. In *EuroVA 2011: International Workshop on Visual Analytics* (2011), Miksch S., Santucci G., (Eds.). 2

[AHZ*14] ALSAKRAN J., HUANG X., ZHAO Y., YANG J., FAST K.: Using entropy-related measures in categorical data visualization. In *2014 IEEE Pacific Visualization Symposium* (2014). 1, 3, 4

[BH86] BARNES J., HUT P.: A hierarchical o (n log n) force-calculation algorithm. *nature 324*, 6096 (1986), 446–449. 3

[DFB*21] DENNIG F. L., FISCHER M. T., BLUMENSCHEIN M., FUCHS J., KEIM D. A., DIMARA E.: Parsetgnostics: Quality metrics for parallel sets. *Computer Graphics Forum* (2021). 2

[Dun89] DUNTEMAN G. H.: *Principal components analysis*. No. 69. Sage, 1989. 2

[Dwy09] DWYER T.: Scalable, versatile and simple constrained graph layout. *Comput. Graph. Forum 28* (06 2009), 991–998. 3

[FJ11] FERNSTAD S. J., JOHANSSON J.: A task based performance evaluation of visualization approaches for categorical data analysis. In *2011 15th International Conference on Information Visualisation* (2011). 2, 4

[FR91] FRUCHTERMAN T. M., REINGOLD E. M.: Graph drawing by force-directed placement. *Software: Practice and experience* (1991). 1

[Fri94] FRIENDLY M.: Mosaic displays for multi-way contingency tables. *Journal of The American Statistical Association* (1994). 2

[Fri01] FRIENDLY M.: Visualizing categorical data. *SAS Institute* (2001). 1, 3

[Gre17] GREENACRE M.: *Correspondence analysis in practice*. chapman and hall/crc, 2017. 2

[HK81] HARTIGAN J. A., KLEINER B.: Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (1981). 1, 2

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer* (1985). 2

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics* (2006). 1, 4

[ML22] MATUTE J., LINSEN L.: Evaluating data-type heterogeneity in interactive visual analyses with parallel axes. *Computer Graphics Forum* (2022). 2

[TEL16] TUOR R., EVÉQUOZ F., LALANNE D.: Parallel bubbles: Categorical data visualization in parallel coordinates. In *Actes de La 28ième Conference Francophone Sur l'Interaction Homme-Machine* (2016), Association for Computing Machinery. 2

[Upt00] UPTON G.: Cobweb diagrams for multi-way contingency tables. *Journal of the Royal Statistical Society Series D* (2000). 2

[Ver67] VERLET L.: Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.* (Jul 1967). 3

[Wor21] WORLD HEALTH ORGANIZATION: Global health observatory data repository, 2021. Accessed on 2021-11-29. URL: http://www.who.int/gho/en/. 4

[ZCYY19] ZHANG C., CHEN Y., YANG J., YIN Z.: An association rule based approach to reducing visual clutter in parallel sets. *Visual Informatics* (2019). Proceedings of PacificVAST 2019. 2

Proceedings published by Eurographics - The European Association for Computer Graphics.