Erschienen in: 2024 IEEE Visualization in Data Science, VDS 2024,
Proceedings. Piscataway, NJ: IEEE, 2024, S. 25-34. ISBN 979-8-3315-2843-0
https://doi.org/10.1109/vds63897.2024.00008

The Categorical Data Map: A Multidimensional Scaling-Based Approach

Frederik L. Dennig, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel A. Keim, and Maximilian T. Fischer

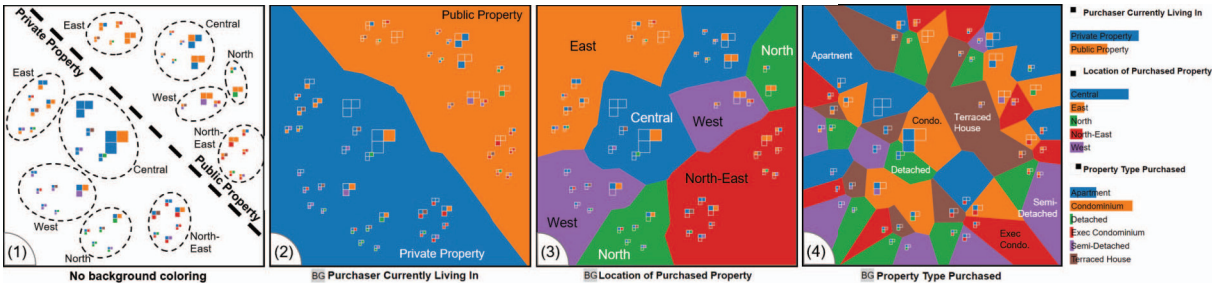


Fig. 1: The *Categorical Data Map* enables projection-based analysis of categorical data here exemplified by the *Property Sales* dataset [48] with *MDS* [80] using the *Jaccard coefficient* [39]: (1) shows 10 groups without layout enrichment. Our method reveals the patterns annotated in (1) in plots (2)-(4). (2) shows a clear separation between Private Property vs Public Property. (3) indicates boundaries and symmetries for the Location of Purchased Property attribute, while in (4), the Property Type Purchased contributes the least to the clusters. The glyph sizes encode the subset sizes, revealing that categories Private Propriety and Central often occur together.

Abstract—Categorical data does not have an intrinsic definition of distance or order, and therefore, established visualization techniques for categorical data only allow for a set-based or frequency-based analysis, e.g., through Euler diagrams or Parallel Sets, and do not support a similarity-based analysis. We present a novel dimensionality reduction-based visualization for categorical data, which is based on defining the distance of two data items as the number of varying attributes. Our technique enables users to pre-attentively detect groups of similar data items and observe the properties of the projection, such as attributes strongly influencing the embedding. Our prototype visually encodes data properties in an enhanced scatterplot-like visualization, visualizing attributes in the background to show the distribution of categories. In addition, we propose two graph-based measures to quantify the plot's visual quality, which rank attributes according to their contribution to cluster cohesion. To demonstrate the capabilities of our similarity-based projection method, we compare it to Euler diagrams and Parallel Sets regarding visual scalability and evaluate it quantitatively on seven real-world datasets using a range of common quality measures. Further, we validate the benefits of our approach through an expert study with five data scientists analyzing the Titanic and Mushroom dataset with up to 23 attributes and 8124 category combinations. Our results indicate that our Categorical Data Map offers an effective analysis method for large datasets with a high number of category combinations.

Index Terms—Categorical data, dimensionality reduction, cluster analysis, similarity-based representation, information visualization

1 Introduction

Categorical data can be encountered in numerous domains, such as representing inventory data describing product properties like color in sales or bioinformatics, encoding the genes formed by nucleotide sequences [2]. In contrast to numeric and ordinal data, categorical data does *not* have an intrinsic order or distance associated with each value pair. The visual analysis of categorical data is challenging since categorical data describes an attribute by name only, with the only supported operators being *equality*, *set membership*, and *mode*.

Currently, there are two widespread methods of visualizing categorical data: (1) *Frequency-based visualizations* [37, 75, 85] map the categorical values to their frequencies, for example, through bar charts, pie charts, or enhanced variants, such as stacked bar charts. In contrast, (2) *set visualizations* solely focus on the set nature of categorical data items, specifically their intersections [4]. Examples include such as Euler diagrams [62] and UpSet plots [53]. Set visualizations like Euler diagrams do not scale well for sets with many intersections because visual clutter is detrimental to their readability. Other, less common solutions treat dimensions independently and map data to a continuous design model [40, 68, 78], leveraging visualization types that initially have been designed for numerical data, such as scatterplots or parallel coordinate plots. However, these approaches deviate from the *discrete* nature of categorical data and suffer from visual clutter and overplotting, limiting their readability [49]. Approaches, such as Parallel Sets [15] and Sankey diagrams [46], follow the frequency and set-based paradigms. These approaches trade effectiveness in visualizing the presence of small subsets for the presentation of frequency information. These approaches require additional design considerations since they tend to emphasize preselected attributes over others [28].

None of the previously described techniques support the similarity-based analysis of categorical data, i.e., deriving the *similarity* of categorical data items as distances such that similar data items are placed close to each other while differing data items are positioned far apart. Analyzing categorical data based on a group or subset similarity is useful, e.g., visually clustering data items only differing in a few attributes can help us better understand important characteristics of the group. Generally, this would allow us to apply methods from cluster analysis to categorical data.

We follow the suggestion by Broeksema et al. [19] to investigate

• *Frederik L. Dennig, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel A. Keim, and Maximilian T. Fischer are with the University of Konstanz (Germany). E-mail: {frederik.dennig, lucas.joos, daniela.blumberg, patrick.paetzold, oliver.deussen, keim, max.fischer}@uni-konstanz.de*

multidimensional scaling to generate *visual mappings* that enable the interpretation of distances and simultaneously convey the properties of data items, i.e., effectively visualizing an item's attributes by using color and position to visually encode attributes. Through this, we address the combinatorial problem of categorical data, i.e., that with the increasing number of attributes and categories, the number of required colors to represent a category with distinguishable colors becomes increasingly difficult. Tackling these challenges, we contribute the following:

(1) A technique applying multidimensional scaling to categorical data while *visually encoding* the category distribution into the background. Through *layout enrichment*, we enable the exploration of the category distribution, enhancing orientation and navigation. Additionally, we contribute *four glyph designs* to represent categorical subsets.
(2) *Quality measures* based on subset distribution to guide the analysis, recommending layout enriched views on attributes contributing strongly to clusters and subset separation.
(3) A *quantitative comparison* to multiple correspondence analysis-based projections and a *qualitative expert study* validating the effectiveness of our approach.
(4) An *online demonstrator* (https://dennig.dbvis.de/categorical-data-map) making the acquired results accessible. To further aid reproducibility, we *openly publish* all our *datasets* and *source code* via OSF (osf.io/jzd46) and the Data Repository of the University of Stuttgart (DaRUS) [29].

With this work, we aim to widen the analytical capabilities for categorical data, particularly for exploratory analysis.

2 Related Work

Our approach is related to visualization and dimensionality reduction methods for categorical data. Furthermore, we propose a layout enrichment for multidimensional projections and contribute visual quality metrics for categorical data projections.

2.1 Visualization Techniques for Categorical Data

Set visualization is one of the core techniques for categorical data. To visualize the members of sets and their intersections, Venn and Euler diagrams are the two most prevalent representations [12]. Multiple adaptations of both techniques mitigate challenges, e.g., to preserve semantics [44], draw area-proportional diagrams [65], or incorporate glyphs to show additional information [58]. Other set visualization techniques use lines to indicate set intersections [66] and matrices to show the cardinality of intersection sets [54], or include the semantic context to visualize sets [57]. Alsallakh et al. presented a comprehensive survey on set visualizations [4]. There are also *frequency-based visualization* methods that focus on attribute frequencies, such as Mosaic plots [37] and Parallel Bargrams [85] by mapping data item occurrences to one or multiple attributes, e.g., a rectangle's area. Other methods map data to a continuous design model, such that they are compatible with visualization for numeric data, e.g., Rosario et al. [68] describe the mapping of categorical data to numeric values for the visualization in Parallel Coordinates [38]. Hybrid methods consider both aspects, e.g., Parallel Sets [15,49] and Sankey diagrams [46]. However, Parallel Sets and Sankey diagrams can suffer from the Müller-Lyer and Sine illusions [27,81] where lines seem to vary in distance or length, affecting the accurate interpretation of frequencies and proportions.

While plenty of approaches visualize categorical data, to the best of our knowledge, none allows identifying groups of similar data items. Thus, we propose a visualization that focuses on similarity.

2.2 Dimensionality Reduction for Categorical Data

Our approach makes use of dimensionality reduction (DR). However, there exist DR methods for categorical data that do not focus on similarity but rather describe the central oppositions in the data [34]. When needing to reduce the dimensionality of categorical data, Correspondence Analysis (CA), similar to Principal Component Analysis (PCA) [42] for numerical data, extracts the standard coordinates, yielding a Biplot [33] of the reduced space. In case of more than two categorical variables, Multiple Correspondence Analysis (MCA) can be used to reduce the number of dimensions showing the central oppositions [34]. Factor analysis of mixed data (FAMD) is a principal component technique for continuous and categorical variables [63]. The continuous variables are scaled to unit variance, and the categorical variables are transformed into a disjunctive data table and then scaled using the specific scaling of MCA to balance the influence of both continuous and categorical variables in the analysis. Multiple Factor Analysis (MFA) combines these methods for mixed data: It uses PCA when variables are quantitative, MCA when variables are qualitative, and FAMD when the active variables belong to both of the two types. The Data Context Map [23] visualizes *mixed-data* using an MDS-based plot and displays categorical attributes on top of the projection while also coloring points and regions according to the predominant category. The approach by Thane et al. [79] uses force-directed graph layouts to visualize categorical datasets representing categories as nodes while edges represent their co-occurrence.

MCA can embed categorical data but, like PCA, is a linear dimensionality reduction technique and thus not able to detect non-linear relationships [19,34]. We propose using MDS to visualize the similarity of categorical data points in a scatterplot-like layout.

2.3 Layout Enrichment for 2-Dimensional Data Projections

The idea to enrich scatterplot layouts by encoding additional information in the background of a projection is not new [61]. The main usage occurs for the visualization of distortions in the topology of the embedding resulting from DR [6]. The following approaches make use of Voronoi diagrams [9] to encode additional information in the background of a projection. Lespinats and Aupetit proposed CheckViz [52], visualizing the presence of *tears* (i.e., missing neighborhood) and *shuffled data* (i.e., wrong neighborhood). Broeksema et al. explored the visualization of categorical data, combining MCA with an enhanced treeview to integrate data record information visualizing user-selected categories. However, they did not address the high redundancy of categorical datasets [19]. Sohns et al. followed a similar approach; however, they used non-linear DR methods to project *mixed data* while using categorical attributes to highlight areas of the embedding space. However, this approach excludes all categorical attributes from the DR process altogether [72]. DICON enables the analysis of multidimensional clusters with an interactive icon-based visualization that encodes additional statistical information visually using space-filling methods, including Voronoi diagrams [21]. Aside from using Voronoi diagrams, other methods for layout enrichment exist [16]. Morariu et al. encode the projection's quality into the plot's background using contours showing the embedding of projections called the metamap [60].

Layout enrichment methods largely focus on visualizing distortions of the projection. The approach by Broeksema et al. [19] does not address the analysis of a single attribute, so we propose a new enrichment that encodes the category of an attribute using color.

2.4 Metrics for Quality and Patterns in Visualizations

Quality metrics for visualizations describe a set of measurements designed to optimize visualizations in terms of *readability* and *clutter reduction* [14]. Other metrics quantify the presence of *patterns* in a visualization. Instead of measuring quality, pattern metrics can be used to compare and rank different visualizations based on specific properties. Examples are: Magnostics for matrix visualizations [13], Scagnostics for general patterns and trends on scatterplots of numeric data [84], Pargnostics for parallel coordinate plots [25], Visualgnostics for projections of high-dimensional data [51], Pixgnostics for pixel-based visualizations [70], and ParSetgnostics for Parallel Sets [28]. SepMe is a machine-learning-based approach to quantify the presence of clusters in scatterplots [8], while ClustMe quantifies the visual separation of classes in scatterplots [1]. Aupetit and Catz [7] addressed the analysis of high-dimensional labeled data using graphs, including Voronoi diagrams. However, this approach does not address categorical data analysis, i.e., where no numerical attributes are present.

We contribute two novel measures for quantifying visual quality for 2-dimensional projections of categorical data. In this way, we improve

the exploration of categorical data by recommending layout-enriched views according to their visual structure.

3 Constructing the Categorical Data Map

Typically, categorical datasets exhibit inherent *sparsity*, i.e., only a fraction of all possible category combinations is present in a dataset, e.g., for the Mushroom dataset, only 8124 out of 243.799.621.632.000 possible combinations. Thus, we assume that there are relationships among the existing categories restricting their combinations. Additionally, categorical datasets can be *highly redundant*, e.g., the Titanic dataset contains 2201 data items but only 24 unique entries, i.e., all data items can be assigned to one of 24 subsets. Thus, we focus on categorical subsets as subsets of unique attribute values. These subsets are our main representations, enabling us to assign a frequency. We leverage these properties in the design of the *Categorical Data Map* as an analytical approach for the similarity-based analysis of categorical subsets with the following *constraints*:

(C1) Distances of categorical subsets in a scatterplot should indicate similarity, i.e., subsets with a smaller distance should differ in fewer attributes than subsets with a larger distance.
(C2) Allow analysts to find groups of subsets by clustering similar categorical subsets and separating outliers.
(C3) Highlight attributes contributing to the clustering of subsets enabling navigation and orientation in the projection.
(C4) Provide a recommendation for attributes to explore first, linked to the distribution of categories in the plot.

An example of our approach is shown in Fig. 1. (C1) and (C2) are described further in Sec. 3.1. We address (C3) by evaluating different glyph designs and layout enrichments for subsets of categorical data (see Sec. 3.2). We address (C4) in Sec. 3.3, describing measures to rank attributes according to their degree of splitting the embedding into connected areas. In the following, we describe how we derive distance relations of categorical data and how a projection-based approach, i.e., the *Categorical Data Map*, is constructed.

3.1 Projecting Categorical Data

The *Categorical Data Map* enables the visual clustering of similar categorical subsets and separating outliers, addressing (C1) and (C2). At the core, we rely on DR to create a scatterplot-like visualization. In general, we describe encoding E, distance measure M, DR method P, and overlap reduction method O to project a categorical dataset x by applying $O(P(M(E(x))))$.

Encoding (E): We convert all data items into a set representing their categorical data values. We define the set of all attributes as $\mathscr{A} := \{a_1, a_2, \ldots, a_{|\mathscr{A}|}\}$ and the possible categories associated with attribute a_i as the set $\mathscr{C}_i := \{c_i^1, c_i^2, \ldots, c_i^{|\mathscr{C}_i|}\}$ with $i \in \mathbb{N}$. $|\mathscr{A}|$ is the cardinality of a set representing a data item, i.e., the number of attributes since a data item has one category associated with each attribute. We denote a data item as $x_n = (c_1^{n_1}, c_2^{n_2}, ..., c_{|\mathscr{A}|}^{n_{|\mathscr{A}|}})$. From a practical point of view, we make sure that all categories have a unique descriptor across all attributes. We then create a representation compatible with the distance measure. We explored the *set representation* and two variants of *one-hot encoding* [20, 35] (see supplementary material for more details).

Distance Measure (M): With the set representation, we can describe the categories of a data item to define similarity. Based on surveys on distance measures for categorical data [17, 22, 76], we chose and evaluated three set-based distance measures: *Overlap coefficient* [83], *Jaccard Similarity Index* [39], and *Sørenson-Dice coefficient* [73]. By including one-hot encoding, converting each categorical value to a new binary dimension enables us to use classical distance measures, such as *Euclidean* or *Manhattan distance*, to describe a dissimilarity relationship (see supplementary material for more details).

Projection Method (P): DR techniques are a set of non-/linear transformation methods with which a dataset's dimensionality can be reduced. We compared the following two DR methods.

Multiple Correspondence Analysis (MCA): This method is the categorical equivalent of PCA. MCA creates groups of items that are similar according to their categories. Objects sharing the same categories are placed close together, and objects with differing categories are placed far apart [34]. To our knowledge, MCA is the only existing technique that directly uses the set representation of categorical data.

Multidimensional Scaling (MDS): This method describes a set of linear and nonlinear DR techniques that attempt to preserve pairwise distances. Multiple criteria are possible; Kruskal's stress optimization criterion is usually used [50]. We create a dissimilarity matrix to compute the projection given one of the described distance measures.

We chose the two methods based on their popularity and common usage in visual data analysis [19, 30] and compared MDS and MCA as DR methods for categorical data.

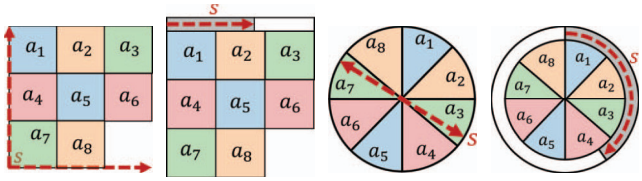
However, a key difference between both methods is that MCA reduces the number of projected points to the number of unique subsets, while MDS, applied naively, would result in a number of projected points equal to the number of categorial data items. Since categorical datasets can contain many duplicates, projecting each data point individually and using a DR method for numeric data (e.g., MDS) could lead to multiple data points being projected to the same position. The main reason is that the distance of identical points is zero. To achieve a comparable result, i.e., the same number of projected points, we remove all duplicates and project one data point for each unique combination of attribute values, i.e., for each categorical subset, describing the prototype of the represented data subset. A second reason is that we want to show the subsets represented by a point irrespective of the method (e.g., MCA or MDS). We visually represent a subset's size (see Sec. 3.2). Reducing the number of data points also improves the runtime of projection algorithms for datasets with duplicate items.

Overlap Reduction (O): Given that some subsets in the categorical data may differ in only one or a few attributes, these subsets will be projected close to each other. This property is desirable in the design of a map by keeping the distances representing similarity coherent. However, it may also introduce overlap if the projected point visually encodes the subset categories through a glyph representation. Additionally, points that are close together will yield small or narrow-shaped Voronoi cells. Thus, we allow users to reduce the overlap after projecting the data using a method based on force-directed graph drawing. This type of layout applies forces to the nodes and edges of a graph [47]. We add a repulsive force to all points with a strength equal to the radius of the glyph while all points are vertices of a fully connected graph, forcing all points into a configuration without overlap but with minimal space in between the glyphs.

3.2 Representing Categorical Data Subsets in Scatterplots

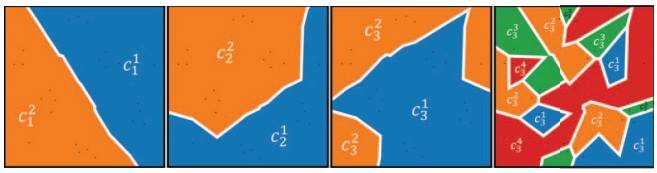
We implemented the visual components of the *Categorical Data Map* using D3 [18]. To represent categorical subsets, we developed four glyph representations and the layout enrichment based on experiences gained during the design phase, addressing (C3). To represent categories, we use the `d3.schemeCategory10` color scale, a well-established color scale for categorical data.

Glyph Representation: To represent categorical subsets, we developed four glyph representations. All glyphs visualize the attributes and their respective values by dividing a square or circle into segments of equal size, such that each segment represents one attribute. This square-based glyph is inspired by pixel visualizations pioneered by Keim et al. [45]. In Fig. 2, this is represented by the categories a_1 to a_8 for the case of a dataset with eight attributes. For all glyphs, the segments are colored according to the respective category of the attribute. However, we discuss some limitations in Sec. 6. The area-based glyphs represent the relative size of a subset $s \in \mathbb{N}$ by the area (see Fig. 2 (a) and (c)). Thus, we calculate the width and height accordingly. The bar- and arc-based glyphs have a fixed size to minimize space requirements and overlap issues with neighboring glyphs (see Fig. 2 (b) and (d)). To reduce overlap while preserving the relative proximity of the projected points, we decided to map a subset's size $s \in \mathbb{N}$ to a bar at the top or an arc surrounding the glyph as an alternative encoding for the subset



(a) Area square (b) Bar square (c) Area circle (d) Arc circle

Fig. 2: Representation of subsets for a dataset with eight attributes. (a) shows the eight attributes in four segments with the same area while the size encodes the overall subset size. (b) shows a similar glyph, but instead, the size is encoded by a bar at the top, and all glyphs have the same size. (c) Encodes the attributes similar to the area square but is circle-shaped. (d) encodes the size by an arc filled according to the subset size.



$\mathscr{F}_{\text{edge}}(a_1) = 0.17$ $\mathscr{F}_{\text{edge}}(a_2) = 0.17$ $\mathscr{F}_{\text{edge}}(a_3) = 0.42$ $\mathscr{F}_{\text{edge}}(a_4) = 0.79$
$\mathscr{F}_{\text{comp}}(a_1) = 0$ $\mathscr{F}_{\text{comp}}(a_2) = 0$ $\mathscr{F}_{\text{comp}}(a_3) = 0$ $\mathscr{F}_{\text{comp}}(a_4) = 0.69$

Fig. 3: The *fracturedness* of attributes differs a lot and can imply an order, i.e., increasing from left to right. The examples are derived from the Titanic dataset [26]. The edge-based (i.e., $\mathscr{F}_{\text{edge}}$) and component-based fracturedness (i.e., $\mathscr{F}_{\text{comp}}$) values are provided below for each attribute.

size. Hence, each unique subset is represented by a square or circle sized relative to the percentage of data points the subset represents or an indicator filled accordingly. This enables users to perceive similar subsets and assess the size of each group.

Layout Enrichment: To enable the observation of cluster characteristics and explore attributes in the projected space, we show a Voronoi diagram [9] for a selected attribute (see Fig. 1). The Voronoi diagram automatically partitions the map into polygons such that each polygon contains exactly one subset. By selecting one attribute of interest, the partition for the selected attribute gets displayed in the background of the projection. The color of the polygon then encodes the category of the selected attribute. Thereby, it is possible to directly spot cluster regions for the selected attribute and to identify cluster boundaries and outlying data points. The appearance of the background can differ a lot across attributes (see Fig. 3). Attributes form distinct contiguous areas of different sizes, indicating a neighborhood or larger area of subsets of the same category. We added *detail-on-demand* using tooltips, allowing users to see the respective category for each polygon directly.

3.3 Measuring Fracturedness

We quantify *fracturedness*, generally defined as the strength with which the Voronoi partitioning of an attribute appears disjointed and fractured (see Fig. 3). We use *fracturedness* to suggest attributes for analysis, e.g., the lower the fracturedness value, the larger the contiguous areas of categories and thus the more straightforward to orient along, addressing (C4). We use the Delaunay triangulation of the Voronoi diagram [9] as a basis for our measures. In contrast to Aupetit and Catz [7], we describe measures for purely categorical datasets. Before describing the measures, we define the common notations following established notations [7, 24]. Let $G := (V, E)$ be the Delaunay triangulation of the discrete set of points P resulting from the projection (see Sec. 3.1). Thus, G is an undirected graph and the dual graph of the Voronoi diagram of the points P. Therefore, there exists exactly one $v \in V$ for every $p \in P$ defining its x,y-location and categories. Each vertex $v \in V$ has exactly one associated category $\mathscr{C}_n(v) \in \mathscr{C}_n$ for each attribute $a_n \in \mathscr{A}$.

Edge-based Fracturedness: We measure the number of edges in G that connect cells with different associated attributes. This concept is

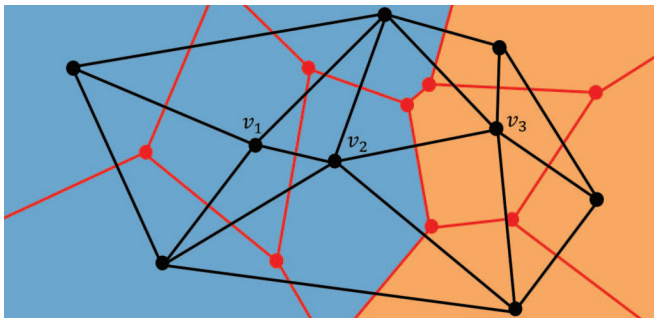


Fig. 4: We illustrate *edge-based fracturedness* with a Delaunay triangulation shown in black, and a Voronoi partitioning with cell borders shown in red. The cells are colored according to the categories of an attribute. v_1, v_2 and v_3 are vertices of the Delaunay triangulation. The edge v_1, v_2 will not contribute to edge-based fracturedness, since it connects cells representing the same category of a given attribute. Edge v_2, v_3 contributes to edge-based fracturedness because it connects cells representing different categories.

shown in Fig. 4. We define an edge $e \in E$ as $\{v_1, v_2\}$ with $v_1, v_2 \in V$ and $v_1 \neq v_2$. An edge contributes to *fracturedness*, if the category for the analyzed attribute a_n and its associated categories in $\mathscr{C}_n$ differ for the connected vertices, i.e., $\mathscr{C}_n(v_1) \neq \mathscr{C}_n(v_2)$ for $\{v_1, v_2\} \in E$. *Edge-based fracturedness* is defined as $\mathscr{F}_{edge} : \mathscr{A} \mapsto [0,1]$ and calculated using Eq. 1.

$$\mathscr{F}_{edge}(a_n) := \frac{|\{v_1, v_2\} \in E : \mathscr{C}_n(v_1) \neq \mathscr{C}_n(v_2)|}{|E|} \text{ with } a_n \in \mathscr{A} \tag{1}$$

Component-based Fracturedness: This measure quantifies the number of continuous areas an attribute produces in the plot through its categories. We show the concept of *component-based fracturedness* in Fig. 5. Each category $c \in \mathscr{C}_n$ defines an induced subgraph $G[S(c)]$ of G, with $S(c) \subset V$ for all $c \in \mathscr{C}_n$ of an attribute $a_n \in \mathscr{A}$. The induced subgraph $G[S(c)]$ is a graph with the vertices $S(c)$ and the edges in E with both of its vertices in $S(c)$. We formally define $S(c)$ for a category $c \in \mathscr{C}_n$ in Eq. 2.

$$S(c) := \{v \,|\, v \in V, \mathscr{C}_n(v) = c\} \text{ for } c \in \mathscr{C}_n \text{ of } a_n \in \mathscr{A} \tag{2}$$

With this definition, a category defines a partition of V, i.e., $\bigcup_{c \in \mathscr{C}_n} S(c) = V$ and a vertex $v \in V$ can only have one category $\mathscr{C}_n(v)$, thus $\bigcap_{c \in \mathscr{C}_n} S(c) = \emptyset$ for a given attribute a_n. Therefore, there exits $|\mathscr{C}_n|$ subgraphs of G for attribute $a_n \in \mathscr{A}$. Let $\omega(G)$ be the number of connected components of any graph G. The *component-based fracturedness* is dependent on the number of connected components of all subgraphs $\omega(G[S(c)])$ for each $c \in \mathscr{C}_n$ (see s_1 to s_6 in Fig. 5). We define the sum of the number of components of all induced subgraphs as $\Omega(a_n)$ for an attribute $a_n \in \mathscr{A}$. $\Omega(a_n)$ is formally defined in Eq. 3:

$$\Omega(a_n) := \sum_{c \in \mathscr{C}_n} \omega(G[S(c)]) \text{ with } a_n \in \mathscr{A} \tag{3}$$

We can also quantify the fracturedness a single category contributes to the overall measure. This allows us to differentiate categories forming contiguous areas and highly fractured ones. The fracturedness $f_{\text{comp}}(c)$ of a single category $c \in \mathscr{C}_n$ is defined in Eq. 4:

$$f_{\text{comp}}(c) := \frac{\omega(G[S(c)]) - 1}{\Omega(a_n)} \text{ with } c \in \mathscr{C}_n \text{ of } a_n \in \mathscr{A} \tag{4}$$

Component-based fracturedness is defined as $\mathscr{F}_{comp} : \mathscr{A} \mapsto [0,1]$ and calculated using Eq. 5. It allows us to compare different attributes and is an alternative measure to $\mathscr{F}_{edge}(a_n)$.

$$\mathscr{F}_{\text{comp}}(a_n) := 1 - \frac{|\mathscr{C}_n|}{\Omega(a_n)} \text{ with } a_n \in \mathscr{A} \tag{5}$$

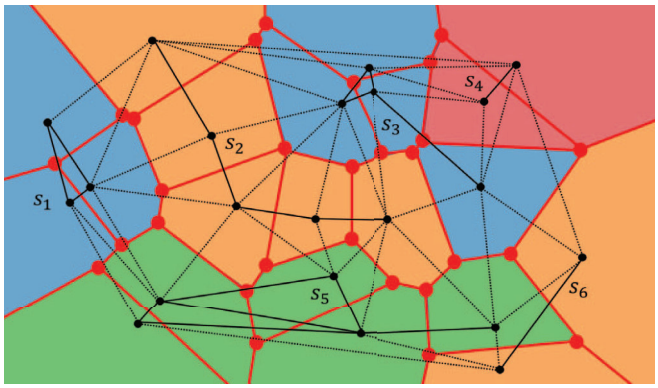


Fig. 5: We describe *component-based fracturedness* with a Voronoi partitioning with cell borders shown in red. The associated Delaunay triangulation is shown in black. The cells are colored according to the categories of an attribute. s_1 to s_6 are six components induced by an attribute through the subgraphs associated with a category. Solid lines connect each subgraph, while dashed lines are not part of any subgraph. With six components $\mathcal{F}_{comp} = 0.33$ for the attribute (see Eq. 5).

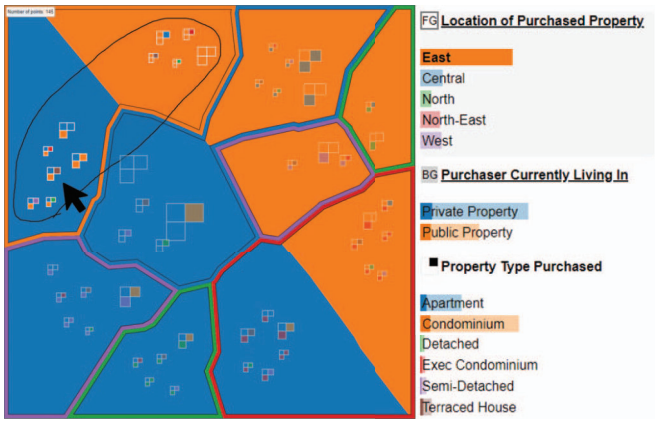


Fig. 6: Through user selection, the borders of a second attribute can be added to the foreground of the plot, e.g., Purchaser Currently Living In is shown in the background as the primary attribute, and Location of Purchased Property is shown in the foreground.

The sum of all component-based fracturedness values of individual categories $c \in \mathscr{C}_n$ is equal to the fracturedness of the attribute $a_n \in \mathscr{A}$. We express this relationship in Eq. 6:

$$\mathcal{F}_{\text{comp}}(a_n) = \sum_{c \in \mathscr{C}_n} f_{\text{comp}}(c) \text{ with } a_n \in \mathscr{A} \tag{6}$$

A mathematical proof of the equivalence described in Eq. 6 can be found in the supplementary material.

3.4 Interacting with Attributes and Subsets

Our prototype allows interactions on the attributes of the dataset shown in the side panel and projected subsets.

Attribute Selection: Users can change the attribute visualized through layout enrichment. We also show the outline for categories of a second selected attribute (see Fig. 6). We add the borders of categories to the foreground if another attribute is already selected and visualized in the background. This visual cue does allow for the observation of one main attribute and a second attribute, similar to the outline of MosaicSets [69]. This introduces less clutter and thus requires less effort to perceive. We initially used textures with different colors to represent different categories. However, using textures of different colors to fill each cell in the Voronoi portioning introduced excessive clutter, and the interpretation of common regions was difficult.

Subset Selection: We allow for the selection and highlighting of groups of subsets. Once the user has selected data items, we show the common categories of the selection using Lasso selection and highlight all data items outside of the selection with the same combination of categories in the side panel on the left, similar to the proximity visualization for continuous data proposed by Aupetit and Catz [6]. This interaction enables cluster analysis since all common categories among the selected items are highlighted (see side panel in Fig. 6). Thus, visual groupings can be compared with respect to the categories and attributes contributing to cluster cohesion. Additionally, all subsets matching the common categories of the selection are also highlighted (see plot in Fig. 6). Together, this allows analysts to observe and judge group cohesion along with the contributing attributes.

Attribute and Category Ordering: A user can select attributes of the dataset listed on the side panel to change the attribute encoded in the foreground and background of the plot. By default, attributes are sorted by their edge-based fracturedness in ascending order, and categories are ordered by their individual contributions to component-based fracturedness in ascending order, allowing for a focus on attributes forming clear splits in the projection space. When selecting subsets (see previous paragraph), the lists of common attributes and distinct attributes are also ordered similarly.

4 Interpreting the Categorical Data Map

In the following, we perform a case study on cluster and attribute analysis, using the Property Sales dataset [48] (see Fig. 1) to show how to interpret emerging patterns for cluster, outlier, and similarity analysis. We chose this dataset because of its relative simplicity. However, it lacks the complexity of large categorical datasets, which we will address in an expert study (see Sec. 5).

Cluster Analysis: There exist a total of $\Pi_{n \in \{1,\ldots,|\mathscr{A}|\}} |\mathscr{C}_n|$ possible data items, given that all combinations of attributes are allowed, resulting in an exponential growth in the number of possible and unique data items. Hence, we can assume that there are dependencies and relationships among the categories contained in a dataset impacting their distribution. This means that groups of subsets that share a set of attributes should form perceivable structures (i.e., clusters) when projected using DR methods. Thus, our approach benefits from and leverages the sparsity of categorical data.

For the Property Sales dataset, there are ten clusters (see Fig. 1 (1)). There is a symmetric split along the center of the projection. Given the size of this dataset, we can observe that the two attributes Purchaser Currently Living In and Location of Purchased Property dominate the appearance of the projection. The glyph sizes indicate that the categories Private Propriety and Central often occur together while {Private Propriety, Central, Condominium} is the largest unique subset. Thus, we can see that most private property is purchased in the central areas, and in this general group, the large majority are condominiums.

Attribute Analysis: By encoding the attribute values in the background, we enable users to analyze the distribution of subsets in the projection with respect to one or two attributes. For the Property Sales dataset, we found that the attribute Purchaser Currently Living In creates a clear and straight division between subsets (see Fig. 1 (2)). We can also see a second level of grouping by the Location of Purchased Property attribute forming a close to orthogonal split in the projection, which can be spotted with our visualizations (see Fig. 1 (3) and Fig. 6). Thus, Purchaser Currently Living In and Location of Purchased Property are the primary attributes. This finding is substantiated when checking the side panel entry of the attribute Property Type Purchased, which has three categories with low frequency. The appearance of the partitioning depends a lot on the selected attributes. When observing the layout enrichment, attributes present themselves on a spectrum from a few clearly separated groups to intermingled and highly fractured appearances. Property Type Purchased does not contribute to elements' clustering (or cluster cohesion) since most

Dataset	TW (↑)			CT (↑)			SC (↑)			NS (↓)			Avg. NH (↑)			Med. NH (↑)		
	MDS+O	MDS+J	MCA	MDS+O	MDS+J	MCA	MDS+O	MDS+J	MCA	MDS+O	MDS+J	MCA	MDS+O	MDS+J	MCA	MDS+O	MDS+J	MCA
Audiology [11]	.58	**.89**	.83	.64	**.90**	.89	.29	**.77**	.69	.17	**.09**	.81	.89	**.92**	**.92**	.98	.98	.98
Mushroom [55]	.96	**.97**	.91	.92	.93	**.97**	.78	.77	**.79**	**.08**	.09	.64	.89	**.90**	.84	.90	**.92**	.87
Titanic [26]	**.86**	**.86**	.76	**.84**	**.84**	.81	**.76**	.75	.59	**.07**	.07	.28	**.68**	**.68**	.63	.74	**.75**	.60
Cyber-Security [86]	.84	**.87**	.79	.83	**.86**	.81	**.87**	.82	.68	**.04**	.06	.30	**.56**	.55	.54	**.66**	.62	.58
Property Sales [48]	**.91**	.89	.73	**.86**	.85	.81	**.70**	.66	.46	**.09**	.10	.24	**.65**	**.65**	.51	**.76**	**.76**	.39
HCI Study (1) [67]	**.84**	.77	.71	**.81**	.79	.73	**.72**	.69	.61	**.07**	**.07**	.18	**.59**	.58	.58	**.53**	**.53**	.52
HCI Study (2) [67]	1.0	1.0	1.0	1.0	1.0	1.0	.86	.85	**.89**	**.03**	**.03**	.11	.56	.56	.56	.57	.57	.57

Table 1: We compare projections of MDS using the Overlap coefficient (*MDS+O*) and the Jaccard distance (*MDS+J*) to *MCA* by applying them to seven real-world datasets. The *MDS* outperforms *MCA* for most datasets and quality metrics. In the case of the *Audiology* dataset with high category overlap, usually present in datasets with many attributes, we found that *MDS* combined with the Jaccard distance outperforms both alternatives.

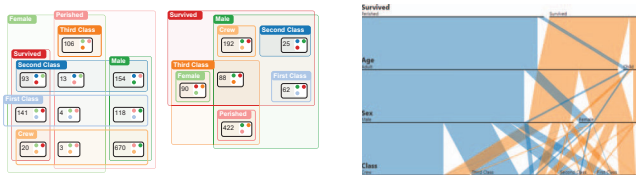


Fig. 7: Two visualizations of the Titanic dataset [26]. A *split* Euler diagram without the Age attribute (left) and an overlap reduced Parallel Sets visualization (right) with *very thin ribbons*. Both have drawbacks with a small dataset and do not scale with an increasing number of attributes.

groups contain subsets of the majority of its categories. Thus, the areas of the categories are disjointed, which reflects the fact that most property types are sold as both private and public property, as well as most of the geographic locations.

5 Evaluation

We qualitatively compare our *Categorical Data Map* to existing visualizations for categorical data and quantitatively compare our approach to MCA used by Broeksema et al. [19]. Additionally, we performed an expert study on two representative datasets with five data scientists.

5.1 Comparison to Euler Diagrams and Parallel Sets

For categorical data, each data point has exactly one category for each attribute, while in Euler diagrams, the number of sets an element is included in is not restricted, i.e., it could be in less. Thus, to truthfully represent categorical data in Euler diagrams, there need to be $\Sigma_{a_i \in \mathscr{A}} |\mathscr{C}_i|$ sets, i.e., one set for each category of all attributes. Euler diagrams may require the selection of specific subsets of attributes and, therefore, are less suitable for exploratory data analysis. For highly intersecting sets, automatic layout methods might not create a single diagram [62]. We show an example of an automatically generated split Euler diagram for the Titanic dataset in Fig. 7 (left). The attribute Age was removed to reduce the diagram's complexity. The Titanic dataset requires ten sets. However, even with eight sets, the visualization is disjointed. Parallel Sets are alternative categorical sets visualization, combining principles from stacked bars and parallel coordinate plots [15, 49]. Fig. 7 (right) shows the Titanic dataset in a Parallel Sets visualization, where the readability is improved through overlap reduction. Small subsets are represented as very thin ribbons on the lowest level, which can be hard to perceive. Visualizing the Mushroom dataset with classical Parallel Sets is not visually feasible since it will have 22 ribbon layers and 8123 subsets on the lowest level (see supplementary material). Alsakran et al. [3] addressed this issue by only visualizing 2-dimensional subsets in a modified Parallel Sets visualization. However, the relation between 2-dimensional subsets is lost. Thus, we argue that Euler diagrams and Parallel Sets, as examples of established visualizations for categorical data, do not scale with an increasing number of attributes.

5.2 Quantitative Evaluation of Projection Quality

We use five quality metrics commonly used in related work for DR to evaluate and compare the quality of our categorical data projections [32]. The result of comparing MDS with Overlap coefficient (*MDS+O*) and Jaccard distance (*MDS+J*) to *MCA* are shown in Tab. 1. We briefly describe each metric below and use them to compare our MDS-based method to MCA using seven real-world categorical datasets.

Trustworthiness (TW) [82] quantifies the proportion of points that remain close in the lower-dimensional representation to assess how accurately local patterns in the projection represent the data patterns. This is linked to the occurrence of "false neighbors" in the protection.

Continuity (CT) [82] measures the ratio of points in the projection that remain close in the original space. This is related to the "missing neighbors" of a projected point.

Normalized Stress (NS) [41] quantifies how well the distances between pairs of points are preserved when mapping from the original space to the projected space. This measure should be as low as possible.

Shepard Diagram Correlation (SC) [31] measures the rank correlation of all distances of the original and the projected space, assessing the quality of distance preservation globally using Spearman's ρ [74].

Neighborhood Hit (NH) [64] measures the proportion of a point's neighbors in the projection space that share the same label as the point itself, averaged across all points in its neighborhood. This metric is related to the separation of labeled data in the projection. In our case, we evaluate every attribute as a set of labels. Thus, we calculate the mean and median values of **NH** across all attributes of a dataset.

TW, **CT**, **NH** require a parameter k defining a neighborhood size. We set $k = 7$, a commonly used value [31]. We found that our approach generally outperforms *MCA* quantitatively. Additional measurements of MDS with other distance measures, detailed descriptions of the quality metrics, and datasets are provided in the supplementary material.

5.3 Qualitative Expert User Study

To evaluate the *Categorical Data Map* we performed a paired analytics study [43]. We conducted an expert study with five data scientists, **E1–E5**, with varying backgrounds. All participants were Ph.D. candidates and students. All were male, and the age range was 25 to 30 years. All experts had experience in the area of information visualization and visual analytics. During the study, we asked the experts to verbalize their thought process to capture it. The following studies are set up using MDS projections of the Mushroom and Titanic dataset using the Overlap coefficient (*MDS+0*). Tab. 1 shows that these projections are higher quality than MCA-based ones regarding most quality metrics.

All trials followed a predefined structure and took between 43 and 57 minutes. The study was conducted in German. The study started with an introduction to the *Categorical Data Map* using the Property Sales dataset by Hassan et al. [36] shown in Fig. 1 and included a description of the square area glyph, layout enrichment, and interactions

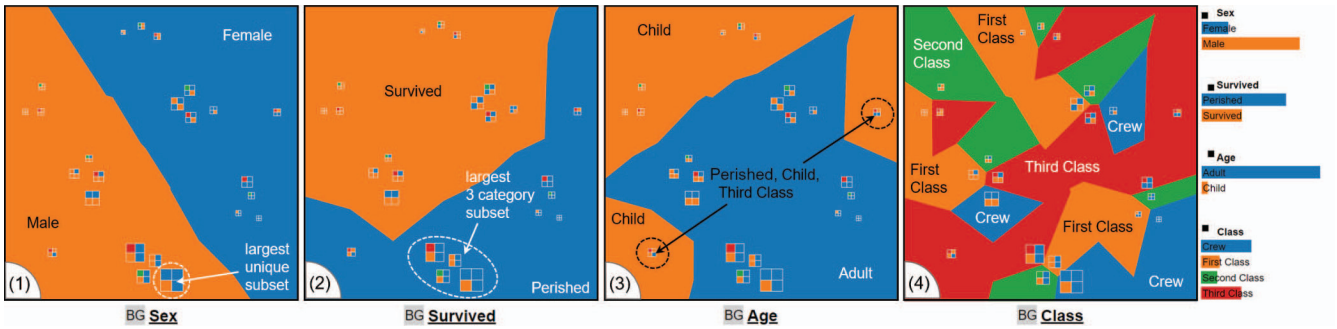


Fig. 8: *Categorical Data Map* visualizations of the *Titanic dataset* [26] using *MDS* and *Overlap coefficient* [77]. (1) The visualization shows six clusters and two outliers. The largest cluster is the subset of Adult, Male, Perished (at the bottom). The background encoding shows that the Survived and Sex attributes are relevant for this dataset, clearly separating the data items. For Sex, the separation is left and right. (2) For Survived, the separation is bottom-right/top-left. (3) The Age dimension also yields a separation, while (4) Class shows no clear structure.

to introduce the expert to the prototype. After the introduction, the experts had the opportunity to ask questions regarding our approach.

Titanic Dataset: The experts had to analyze the Titanic dataset [26] using the *Categorical Data Map* shown in Fig. 8. **E1–E5** were able to locate the largest subset {Male, Perished, Adult, Crew} by looking at the visualization without any additional interaction (Fig. 8 (1)). **E1–E5** used Lasso selection to find and validate that the largest subset regarding three attributes is {Male, Perished, Adult} (Fig. 8 (2)). Additionally, **E1–E5** were able to find six clusters and two outliers. **E1**, **E3**, and **E4** found that the outliers represent the subsets defined by the categories {Perished, Child, Crew} (Fig. 8 (3)). **E1**, **E3**, and **E5** commented on the high number of perished males and the large number of casualties among the {Male, Crew}. **E1–E5** used the layout enrichment to navigate and reason about the location of subsets, including the Class attribute (Fig. 8 (4)). **E2** commented on the close to orthogonal split in the projection between Sex and Survived shown in Fig. 8 (1) and (2).

Mushroom Dataset: **E1–E5** had the opportunity to perform an open exploration task and were only given the information that the dataset is about mushrooms and that the class attribute indicates their poisonousness. The glyph was replaced with a simple black dot to reduce the visual complexity. **E1–E5** perceived five clusters right at the outset. **E2** and **E3** used the Lasso selection together with layout enrichment to determine differentiating categories for cluster separation, e.g., evanescent, large, and pendent for the ring-type attribute (Fig. 9 (3)). **E1–E5** found the poisonous outliers nested in the group representing edible mushrooms (Fig. 9 (1)) being poisonous mushrooms very similar to edible ones. **E1–E5** found the general rule that mushrooms with an fishy foul, musty, spicy or other unpleasant smells indicate a poisonous mushroom (Fig. 9 (2)). During the open exploration task, **E1**, **E2**, and **E3** found the rule without additional information. **E4** and **E5** needed help to find the class and odor combination. However, **E4** and **E5** could deduce the rule by only interpreting the plot. **E3**, quickest in exploring the dataset, found that stalk-surface-below-ring is silky for the majority of poisonous mushrooms and the stalk-surface-below-ring is mostly smooth for edible ones (Fig. 9 (4)).

General Comments: Before concluding the study, the participants were asked to comment on their preferences for the available glyph designs. **E1**, **E2**, and **E4** preferred a circular glyph design (Fig. 2 (c) and (d)) over a square design. **E3** and **E5** preferred square glyph designs (Fig. 2 (a) and (b)). **E1** and **E3** found that the area-based glyphs are inferior to the alternative designs for reading off precise subset sizes. **E1** mentioned as a drawback that the glyphs are not rotation invariant. **E1** commented that the layout enrichment is very useful for navigation and orientation and helps to perceive the impact on category groups. However, **E1** also noted that the layout enrichment does not reflect the ratio of data items with a given category. **E3** mentioned a general preference for the map metaphor by being helpful for orientation among different subsets. **E2** mentioned potential scalability issues with the glyph for large datasets, e.g., for a high number of attributes, and proposed semantic zoom as a potential option. **E1–E5** commented that ordering attributes according to their fracturedness was understandable and useful. During the general questions at the end, **E2–E4** freely explored plots created with other distance measures and DR methods. **E3** commented that the result of MCA-based plots was hard to interpret, noticing the disjointed layout enrichment and thus having larger fracturedness. **E1** mentioned issues with the encoding of categories, such as the category North not being located north of the plot or the category brown not having the color brown, and suggested being able to select the color of a category manually.

6 Discussion and Future Work

In this section, we discuss the lessons-learned, reflect on the design decisions, and discuss computational complexity and future work.

Visualizing Attributes and Categories: We initially used circular glyphs as shown in Fig. 2 (c) and (d), which had the benefit of using the available space effectively since overlap minimization relative to the radius is straightforward to implement. The subset size encoding by the arc around the circle enables finer-grained distinction of sizes since it offers more space. However, during the design phase, users misinterpreted the circle segments as pie charts, a common method for displaying categorical data. Thus, we decided to circumvent this common misconception by using square-based representation for the categorical subsets. However, three out of five experts preferred a circular glyph design.

There are visual limitations to the number of dimensions and categories that our approach is able to support. The number of visually distinguishable categories is limited by the number of square segments that fit into the glyph, which is limited by the screen space. The number of attributes is limited by the number of colors, which have to be distinguishable and memorizable. Thus, we suggest following Miller's Law [59] for the number of dimensions and attributes, which proposes a maximum of seven plus or minus two. Alternatively, we suggest interactions such as semantic zoom, e.g., removing attributes for which all subsets have the same category after zooming in on a specific area.

Encoding of Subset Sizes: We evaluated four different visual encodings for the size of a categorical data subset (see Fig. 2). The area-based glyph makes it easier to perceive subset sizes at a glance, and thus, a user can spot the distribution of the dataset directly. Still, it suffers from overlap, especially for tight clusters. Thus, there is a benefit to applying methods to reduce overlap. We are able to mitigate the overlap problem with the force-directed overlap reduction largely. Simplifying the representation of a dot requires less space, but the assessment of subset sizes requires interaction. It is possible to remove the subset size information altogether. However, this may limit analysis tasks where the subset sizes it not important, e.g., the Mushroom dataset. All glyph designs benefit from a mouse-over mechanism that moves the currently selected glyph to the top so that all attributes can be observed.

Encoding of Attributes Into the Background: Fig. 8 shows that encoding an attribute into the visualization gives insight into the topology of the projection. We could also show the benefit of encoding multiple

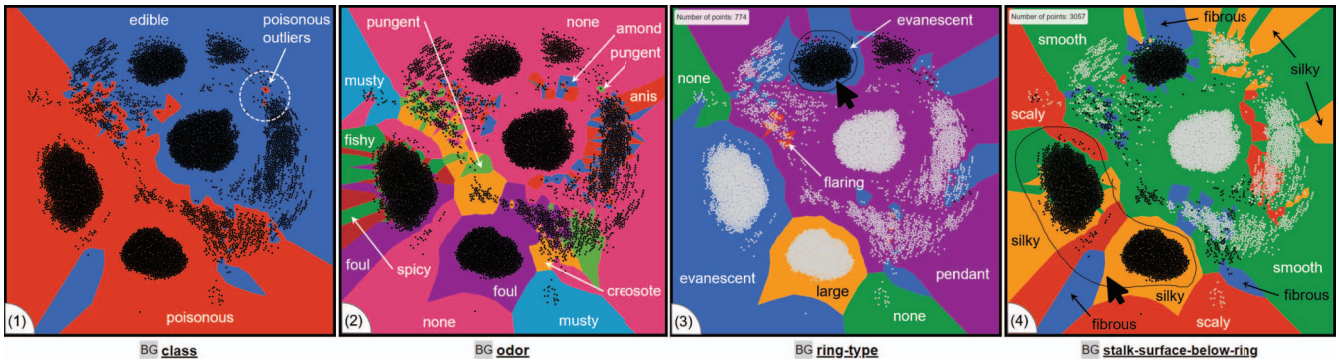


Fig. 9: *Categorical Data Map* visualizations of the *Mushroom dataset* [55] using the *MDS* and *Overlap coefficient*. (1) Two poisonous mushrooms very similar to edible mushrooms. (2) Comparing class and odor reveals that the poisonous outlier has a pungent odor. Continued analysis reveals that mushrooms with an unpleasant smell are poisonous. (3) After the selection of a cluster, the ring-type is identified as a defining characteristic for the cohesion of visible clusters and is used as a property for the classification of mushrooms. (4) Selecting two poisonous clusters, reveals that the vast amount of poisonous mushrooms are silky at the stalk-surface-below-ring, while there exist very few silky mushrooms that are edible.

attributes into the background to allow for a more complex representation of the topology. We found that the number of categories of an attribute weakly influences the fracturedness of an attribute. However, the main factor is the number of subsets containing the attribute, i.e., an attribute with two categories and an occurrence roughly equal among all subsets will yield a low fracturedness for that attribute. With increased imbalance between the categories, the fracturedness may increase if other more balanced attributes are present.

We discussed the use of *weighted Voronoi diagrams* [5] to better reflect the subset size in the background encoding. The use of a weighted Voronoi diagram will conflict with local cluster patterns; more specifically, for imbalanced datasets, the area of one Voronoi cell extends below the point of its neighbors, requiring restrictions on the range weights. This behavior makes the layout enrichment hard to interpret since points are placed inside or close to an area representing a category they do not belong to. For datasets with only unique entries, the weight Voronoi diagram will be identical to the regular Voronoi diagram. To organize subsets, we also considered *Voronoi Treemaps* [10]. However, Voronoi Treemaps require a hierarchical structure, just like regular Tree Maps [71] and, thus, cannot be applied to categorical data without additional information to derive a hierarchy of attributes.

Computational Complexity: The number of data records n poses potential limitations. The time complexity of projecting data is determined by the DR methods. However, since categorical data sets are sparse, as discussed in Sec. 4, the number of projected subsets is significantly lower than that of data records. The Voronoi diagram calculation and the corresponding Delaunay triangulation are both in $O(n\,log(n))$ [9]. The time complexity of calculating the fracturedness measures depends on the number of vertices and edges of the Delaunay triangulation, which will have n vertices and $3n-3-h$ edges, where h is the number of vertices on the convex hull. The time complexity of calculating *edge-based fracturedness* is based on enumerating all edges of the Delaunay triangulation and has a time complexity of $O(|E|)$. The time complexity of calculating *component-based fracturedness* is dependent on the algorithm for determining the number of components. We use a depth-first search-based approach with time complexity of $O(|V|+|E|)$. Thus, the dimensionality reduction method employed poses the highest contribution to the time complexity, $O(n^3)$ for MDS.

Future Work: We found that Voronoi cells can overrepresent the amount of data associated with a specific category. Thus, there is a need for a new layout enrichment method following these constraints: (1) The global area associated with one category should be relative to the occurrence in the dataset (data-ink ratio), (2) the extent of individual category areas should remain close to their projected data point positions, (3) where meaningful (e.g., among clusters), the layout enrichment should visually enclose the data points with the same category if the data-ink ratio allows. The expert study showed that the color assignment for foreground and background colors could be improved. We suggest assigning attributes to a few sets of colors based on an exploration phase. Later in the analysis, we require one color set for the attribute used in the background, one for the foreground using a distinctive palette, and one for the attribute under focus by the user. All the other attributes would be assigned a neutral color (e.g., grey). In this paper, we studied the use of MDS for categorical data analysis. However, following the approach of encoding categorical data into distances, other DR methods could be used (e.g., t-SNE [80] or UMAP [56]). These can be evaluated and compared quantitatively, following the evaluation presented in this paper. We think that the concept of *fracturedness* can be transferred to high-dimensional space when analyzing categorical data. Such a measure can be used to compare the low- and high-dimensional representations and provide a quality measure for projections of categorical data.

7 Conclusion

We presented a novel projection-based visualization method to address the need for similarity-based analysis techniques for categorical data. We leverage distance relations based on set intersections to create enhanced and interactive glyph-based scatterplot-like visualizations called the *Categorical Data Map*. We visualized attributes and categories by calculating a Voronoi partitioning and coloring the cells according to the category of the associated attribute. Our method allows for exploring the categorical data space through segmentation, enabling the orientation along an automatic or user-selected attribute. For automatic selection, we rank-order attributes along a visual property we defined as *fracturedness* measures. We quantitively evaluated different distance measures for the projection of categorical data with MDS, suggesting that the Overlap coefficient and Jaccard distance yield results outperforming MCA. Through a case study, we showed that our *Categorical Data Map* can support the identification of similar subsets and clusters, as well as the detection of attributes with a strong influence on the topology of the embedding. In an expert study, we were able to confirm that our approach facilitates the analysis of categorical data, especially for large datasets, by grouping similar subsets while, through layout enrichment, visualizing the distribution of categories of an attribute. We published a demonstrator and our results online so that users can interactively experiment with our approach and build upon our results. We conclude that the *Categorical Data Map* effectively analyzes large categorical datasets, 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 (Projects A01 and A03) and the Federal Ministry for Economic Affairs and Climate Action (BMWK, grant No. 03EI1048D).

REFERENCES

[1] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. ClustMe: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. *Comput. Graph. Forum*, 38(3):225–236, 2019. doi: 10.1111/CGF.13684 2

[2] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, 3rd ed., 2018. 1

[3] J. Alsakran, X. Huang, Y. Zhao, J. Yang, and K. Fast. Using entropy-related measures in categorical data visualization. In *IEEE Pacific Vis. Symp.*, pp. 81–88, 2014. doi: 10.1109/PacificVis.2014.43 6

[4] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. J. Rodgers. The state-of-the-art of set visualization. *Comput. Graph. Forum*, 35(1):234–260, 2016. doi: 10.1111/cgf.12722 1, 2

[5] P. F. Ash and E. D. Bolker. Generalized Dirichlet tessellations. *Geometriae Dedicata*, 20(2):209–243, 4 1986. doi: 10.1007/BF00164401 8

[6] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7-9):1304–1330, 2007. doi: 10.1016/J.NEUCOM.2006.11.018 2, 5

[7] M. Aupetit and T. Catz. High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing*, 63:139–169, 2005. doi: 10.1016/j.neucom.2004.04.009 2, 4

[8] M. Aupetit and M. Sedlmair. SepMe: 2002 new visual separation measures. In C. Hansen, I. Viola, and X. Yuan, eds., *IEEE Pacific Vis. Symp.*, pp. 1–8. IEEE Computer Society, 2016. doi: 10.1109/PACIFICVIS.2016.7465244 2

[9] F. Aurenhammer. Voronoi diagrams - A survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405, 1991. doi: 10.1145/116873.116880 2, 4, 8

[10] M. Balzer and O. Deussen. Voronoi treemaps. In J. T. Stasko and M. O. Ward, eds., *IEEE Symp. Inf. Vis.*, pp. 49–56. IEEE Computer Society, 2005. doi: 10.1109/INFVIS.2005.1532128 8

[11] R. Bareiss, B. W. Porter, and C. C. Wier. Protos: An exemplar-based learning apprentice. *Int. J. Man Mach. Stud.*, 29(5):549–561, 1988. doi: 10.1016/S0020-7373(88)80012-9 6

[12] M. E. Baron. A note on the historical development of logic diagrams: Leibniz, Euler and Venn. *Math. Gaz.*, 53(384):113–125, 1969. 2

[13] M. Behrisch, B. Bach, M. Hund, M. Delz, L. von Rüden, J. Fekete, and T. Schreck. Magnostics: Image-based search of interesting matrix views for guided network exploration. *IEEE Trans. Vis. Comput. Graph.*, 23(1):31–40, 2017. doi: 10.1109/TVCG.2016.2598467 2

[14] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, T. Schreck, D. Weiskopf, and D. A. Keim. Quality metrics for information visualization. *Comput. Graph. Forum*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446 2

[15] F. Bendix, R. Kosara, and H. Hauser. Parallel sets: Visual analysis of categorical data. In *IEEE Symp. Inf. Vis.*, pp. 133–140, 2005. doi: 10.1109/INFVIS.2005.1532139 1, 2, 6

[16] D. Blumberg, Y. Wang, A. Telea, D. A. Keim, and F. L. Dennig. Inverting Multidimensional Scaling Projections Using Data Point Multilateration. In *16th Int. EuroVis Workshop Visual Analytics*. The Eurographics Association, 2024. doi: 10.2312/eurova.20241112 2

[17] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *SIAM Int. Conf. Data Min.*, pp. 243–254. SIAM, 2008. doi: 10.1137/1.9781611972788.22 3

[18] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185 3

[19] B. Broeksema, A. C. Telea, and T. Baudel. Visual analysis of multi-dimensional categorical data sets. *Comput. Graph. Forum*, 32(8):158–169, 2013. doi: 10.1111/cgf.12194 1, 2, 3, 6

[20] J. Brownlee. Why one-hot encode data in machine learning? Online, 2017. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/, last accessed 2023-03-01. 3

[21] N. Cao, D. Gotz, J. Sun, and H. Qu. DICON: interactive visual analysis of multidimensional clusters. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2581–2590, 2011. doi: 10.1109/TVCG.2011.188 2

[22] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Models Methods Appl. Sci.*, 1(4):300–307, 2007. 3

[23] S. Cheng and K. Mueller. The Data Context Map: Fusing data and attributes into a unified display. *IEEE Trans. Vis. Comput. Graph.*, 22(1):121–130, 2016. doi: 10.1109/TVCG.2015.2467552 2

[24] J. Clark and D. A. Holton. *A First Look at Graph Theory*. World Scientific, 1991. doi: 10.1142/1280 4

[25] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1017–1026, 2010. doi: 10.1109/TVCG.2010.184 2

[26] R. J. M. Dawson. The "unusual episode" data revisited, 1995. `http://jse.amstat.org/v3n3/datasets.dawson.html`, last accessed 2023-03-01. 4, 6, 7

[27] R. H. Day and E. J. Stecher. Sine of an illusion. *Perception*, 20:49–55, 1991. doi: 10.1068/p200049 2

[28] F. L. Dennig, M. T. Fischer, M. Blumenschein, J. Fuchs, D. A. Keim, and E. Dimara. Parsetgnostics: Quality metrics for parallel sets. *Comput. Graph. Forum*, 40(3):375–386, 2021. doi: 10.1111/cgf.14314 1, 2

[29] F. L. Dennig, L. Joos, D. Blumberg, D. A. Keim, and M. T. Fischer. The Categorical Data Map - Replication Data, 2024. doi: 10.18419/darus-3372 2

[30] F. L. Dennig, M. Miller, D. A. Keim, and M. El-Assady. FS/DS: A theoretical framework for the dual analysis of feature space and data space. *IEEE Trans. Vis. Comput. Graph.*, 2023. Early Access. doi: 10.1109/TVCG.2023.3288356 3

[31] M. Espadoto, N. S. T. Hirata, and A. C. Telea. Deep learning multidimensional projections. *Inf. Vis.*, 19(3):247–269, 2020. doi: 10.1177/1473871620909485 6

[32] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Trans. Vis. Comput. Graph.*, 27(3):2153–2173, 2021. 6

[33] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 12 1971. doi: 10.1093/biomet/58.3.453 2

[34] M. Greenacre and Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, 1 ed., 2006. doi: 10.1201/9781420011319 2, 3

[35] J. T. Hancock and T. M. Khoshgoftaar. Survey on categorical data for neural networks. *J. Big Data*, 7(1):28, 2020. doi: 10.1186/S40537-020-00305-W 3

[36] S. Hassan and G. Pernul. Efficiently managing the security and costs of big data storage using visual analytics. In *16th Int. Conf. Inf. Integr. Web-based Appl. Serv.*, pp. 180–184, 2014. doi: 10.1145/2684200.2684333 6

[37] H. Hofmann, A. Siebes, and A. F. X. Wilhelm. Visualizing association rules with interactive mosaic plots. In *6th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 227–235. ACM, 2000. doi: 10.1145/347090.347133 1, 2

[38] A. Inselberg. The plane with parallel coordinates. *Vis. Comput.*, 1(2):69–91, August 1985. doi: 10.1007/bf01898350 2

[39] P. Jaccard. The distribution of the flora in the alpine zone.1. *New Phytol.*, 11(2):37–50, 1912. doi: 10.1111/j.1469-8137.1912.tb05611.x 1, 3

[40] D. F. Jerding and J. T. Stasko. The information mural: A technique for displaying and navigating large information spaces. *IEEE Trans. Vis. Comput. Graph.*, 4(3):257–271, 1998. doi: 10.1109/2945.722299 1

[41] P. Joia, D. B. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2563–2571, 2011. doi: 10.1109/TVCG.2011.220 6

[42] I. T. Jolliffe. *Principal Component Analysis*. Springer, 1986. doi: 10.1007/978-1-4757-1904-8 2

[43] L. T. Kaastra and B. D. Fisher. Field experiment methodology for pair analytics. In *5th Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 152–159. ACM, 2014. doi: 10.1145/2669557.2669572 6

[44] R. Kehlbeck, J. Görtler, Y. Wang, and O. Deussen. SPEULER: Semantics-preserving euler diagrams. *IEEE Trans. Vis. Comput. Graph.*, 28(1):433–442, 2022. doi: 10.1109/TVCG.2021.3114834 2

[45] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Trans. Vis. Comput. Graph.*, 6(1):59–78, 2000. doi: 10.1109/2945.841121 3

[46] A. B. W. Kennedy and H. R. Sankey. The thermal efficiency of steam engines. *Minutes Proc. Inst. Civ. Eng.*, 134:278–312, 1898. doi: 10.1680/imotp.1898.19100 1, 2

[47] S. G. Kobourov. Spring embedders and force directed graph drawing algorithms. *CoRR*, abs/1201.3011, 2012. 3

[48] L. C. Koh, A. Slingsby, J. Dykes, and T. S. Kam. Developing and applying a user-centered model for the design and implementation of information visualization tools. In *15th Int. Conf. Inf. Vis.*, pp. 90–95, 2011. doi: 10.

1109/IV.2011.32 1, 5, 6
[49] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Trans. Vis. Comput. Graph.*, 12(4):558–568, 2006. doi: 10.1109/TVCG.2006.76 1, 2, 6
[50] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. doi: 10.1007/BF02289565 3
[51] D. J. Lehmann, F. Kemmler, T. Zhyhalava, M. Kirschke, and H. Theisel. Visualnostics: Visual guidance pictograms for analyzing projections of high-dimensional data. *Comput. Graph. Forum*, 34(3):291–300, 2015. doi: 10.1111/cgf.12641 2
[52] S. Lespinats and M. Aupetit. CheckViz: Sanity check and topological clues for linear and non-linear mappings. *Comput. Graph. Forum*, 30(1):113–125, 2011. doi: 10.1111/J.1467-8659.2010.01835.X 2
[53] A. Lex and N. Gehlenborg. Points of view: Sets and intersections. *Nature Methods*, 11(8):779, 2014. doi: 10.1038/nmeth.3033 1
[54] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248 2
[55] G. Lincoff and N. A. Society. *National Audubon Society field guide to North American mushrooms*. Audubon Society field guide series. Knopf: Distributed by Random House New York, 1981. 6, 8
[56] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. 8
[57] W. Meulemans, N. H. Riche, B. Speckmann, B. Alper, and T. Dwyer. KelpFusion: A hybrid set visualization technique. *IEEE Trans. Vis. Comput. Graph.*, 19(11):1846–1858, 2013. doi: 10.1109/TVCG.2013.76 2
[58] L. Micallef, P. Dragicevic, and J. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2536–2545, 2012. doi: 10.1109/TVCG.2012.199 2
[59] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, pp. 81–97, 1956. 7
[60] C. Morariu, A. Bibal, R. Cutura, B. Frenay, and M. Sedlmair. Predicting user preferences of dimensionality reduction embedding quality. *IEEE Trans. Vis. Comput. Graph.*, 29(1):745–755, 1 2023. doi: 10.1109/TVCG.2022.3209449 2
[61] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. Vis. Comput. Graph.*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735 2
[62] P. Paetzold, R. Kehlbeck, H. Strobelt, Y. Xue, S. Storandt, and O. Deussen. RectEuler: Visualizing intersecting sets using rectangles. *Comput. Graph. Forum*, 42(3):87–98, 2023. doi: 10.1111/cgf.14814 1, 6
[63] J. Pagès. *Multiple Factor Analysis by Example Using R*. R Series. Chapman & Hall/CRC, 1 ed., 2014. doi: 10.1201/b17700 2
[64] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Trans. Vis. Comput. Graph.*, 14(3):564–575, 2008. doi: 10.1109/TVCG.2007.70443 6
[65] J. G. Pérez-Silva, M. Araujo-Voces, and V. Quesada. nVenn: generalized, quasi-proportional Venn and Euler diagrams. *Bioinformatics*, 34(13):2322–2324, 2018. doi: 10.1093/bioinformatics/bty109 2
[66] P. J. Rodgers, G. Stapleton, and P. Chapman. Visualizing sets with linear diagrams. *ACM Trans. Comput. Hum. Interact.*, 22(6):27:1–27:39, 2015. doi: 10.1145/2810012 2
[67] K. Rogers, J. Wiles, S. Heath, K. Hensby, and J. Taufatofua. Discovering patterns of touch: A case study for visualization-driven analysis in human-robot interaction. In *11th ACM/IEEE International Conference on Human Robot Interaction*, pp. 499–500, 2016. doi: 10.1109/HRI.2016.7451825 6
[68] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Inf. Vis.*, 3(2):80–95, 2004. doi: 10.1057/palgrave.ivs.9500072 1, 2
[69] P. Rottmann, M. Wallinger, A. Bonerath, S. Gedicke, M. Nöllenburg, and J. Haunert. MosaicSets: Embedding set systems into grid graphs. *IEEE Trans. Vis. Comput. Graph.*, 29(1):875–885, 2023. doi: 10.1109/TVCG.2022.3209485 5
[70] J. Schneidewind, M. Sips, and D. A. Keim. Pixnostics: Towards measuring the value of visualization. In *IEEE Symp. Vis. Anal. Sci. Technol.*, pp. 199–206, 2006. doi: 10.1109/VAST.2006.261423 2
[71] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, 1992. doi: 10.1145/102377.115768 8
[72] J. Sohns, M. Schmitt, F. Jirasek, H. Hasse, and H. Leitte. Attribute-based explanation of non-linear embeddings of high-dimensional data. *IEEE Trans. Vis. Comput. Graph.*, 28(1):540–550, 2022. doi: 10.1109/TVCG.2021.3114870 2
[73] T. Sørenson. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34, 1948. 3
[74] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159 6
[75] M. Spenke and C. Beilken. Visualization of Trees as Highly Compressed Tables with InfoZoom. In *IEEE Symp. Inf. Vis.*, pp. 122–123, 2003. 1
[76] Z. Sulc and H. Rezanková. Comparison of similarity measures for categorical data in hierarchical clustering. *J. Classif.*, 36(1):58–72, 2019. doi: 10.1007/S00357-019-09317-5 3
[77] D. Szymkiewicz. Une conlribution statistique à la géographie floristique. *Acta Societatis Botanicorum Poloniae*, 11(3):249–265, 1934. doi: 10.5586/asbp.1934.012 7
[78] S. T. Teoh and K. Ma. Paintingclass: Interactive construction, visualization and exploration of decision trees. In *9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 667–672. ACM, 2003. doi: 10.1145/956750.956837 1
[79] M. Thane, K. M. Blum, and D. J. Lehmann. CatNetVis: Semantic visual exploration of categorical high-dimensional data with force-directed graph layouts. In *25th Eurographics Conf. on Vis.*, 2023. doi: 10.2312/evs.20231049 2
[80] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008. 1, 8
[81] S. VanderPlas and H. Hofmann. Signs of the sine illusion—why we need to care. *J. Comput. Graph. Stat.*, 24(4):1170–1190, 2015. doi: 10.1080/10618600.2014.951547 2
[82] J. Venna and S. Kaski. Visualizing gene interaction graphs with local multidimensional scaling. In *14th European Symposium on Artificial Neural Networks*, pp. 557–562, 2006. 6
[83] W. M. Waggener. *Pulse Code Modulation Techniques*. Springer, 1994. 3
[84] L. Wilkinson, A. Anand, and R. L. Grossman. Graph-theoretic scagnostics. In *IEEE Symp. Inf. Vis.*, pp. 157–164, 2005. doi: 10.1109/INFVIS.2005.1532142 2
[85] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton. Parallel bargrams for consumer-based information exploration and choice. In *14th Annual ACM Symp. User Interface Softw. Technol.*, pp. 51–60. ACM, 2001. doi: 10.1145/502348.502357 1, 2
[86] Y. Yano, R. G. Kula, T. Ishio, and K. Inoue. Verxcombo: an interactive data visualization of popular library version combinations. In *23rd International Conference on Program Comprehension*, pp. 291–294, 2015. doi: 10.1109/ICPC.2015.43 6