Evaluating Ordering Strategies of Star Glyph Axes

Matthias Miller*  
University of Konstanz

Xuan Zhang†  
RWTH Aachen University

Johannes Fuchs*  
University of Konstanz

Michael Blumenschein*  
University of Konstanz

Abstract
Star glyphs are a well-researched visualization technique to represent multi-dimensional data. They are often used in small multiple settings for a visual comparison of many data points. However, their overall visual appearance is strongly influenced by the ordering of dimensions. To this end, two orthogonal categories of layout strategies are proposed in the literature: order dimensions by similarity to get homogeneously shaped glyphs vs. order by dissimilarity to emphasize spikes and salient shapes. While there is evidence that salient shapes support clustering tasks, evaluation, and direct comparison of data-driven ordering strategies has not received much research attention. We contribute an empirical user study to evaluate the efficiency, effectiveness, and user confidence in visual clustering tasks using star glyphs. In comparison to similarity-based ordering, our results indicate that dissimilarity-based star glyph layouts support users better in clustering tasks, especially when clutter is present.

Keywords: Star glyph, axes ordering, quantitative evaluation

1 Introduction
Data glyphs are compact visual representations of multi-dimensional data points. Due to their small graphical appearance, they can be used in various settings like within node-link diagrams [12], treemaps [14], tables [23], or geographic maps [16]. For instance, star glyphs are employed in the medical domain [30] and can be used to show the spatial distribution of food production [28].

Due to their use of visual variables, star glyphs [33] are an adequate choice to encode single data points comprising numerical data. The glyph’s axes represent the data dimensions, and their lengths encode numeric values. Since glyphs are versatile, different design variations of star glyphs emerged in literature. Many have already been extensively analyzed by the community (e.g., [16], see [17] for a full enumeration). However, there is not much empirical research about the effect of axes ordering strategy on visual comparison tasks.

The ordering influences the shape of a star glyph and affects its readability and similarity judgment [24, 25]. Therefore, we need (task-based) guidelines to arrange the dimensions in star glyphs [35].

Numerous ordering strategies for star glyphs have been proposed [3, 4, 15, 22, 24, 25, 29, 35, 37] which can be grouped into similarity-based (short: SIM), favoring homogeneous shapes, and dissimilarity-based orderings (short: D1S), emphasizing spikes and salient shapes. Some approaches also discuss symmetry, monotonicity, convexity and concavity, feature saliency, and user-driven relationships among neighboring dimensions. The ordering strategies typically analyze the relationship among all pair-wise dimensions and then adjust the axes of every star glyph simultaneously according to a metric (e.g., SIM or D1S). However, this also means that not all glyphs will result in the desired shape. In particular, outliers may be encoded by shapes which the reordering algorithm is trying to avoid.

We address the research question: “Which ordering strategy is most useful for similarity search and data grouping tasks (clustering) using star glyphs?”

2 Related Work
Finding an optimal star glyph axes ordering has proven to be NP-complete [3] and more research is required [31, 35]. It is related to the ordering of axes in parallel coordinates [3, 11, 21, 34, 38], RadViz [1, 9, 10, 20], ArcViz [27], and other axes-based radial visualizations as summarized by Behrisch et al. [5]. Ordering algorithms typically define an objective function, modeling a good dimension order (according to their interpretation) and apply a heuristic to find an axes order which maximizes the objective function [35].

2.1 Dimension Ordering Strategies
Different visual characteristics can be subject to shape optimization when applying specific ordering strategies of the star glyph axes. Ward [35] summarizes four major strategies which have been extended by others: user- and data-driven, correlation- and similarity-driven, spikes and salient shapes, and symmetry-driven.

User-driven dimension orderings enable experts to adjust the shape of a star glyph based on their domain knowledge [32]. Users can select a data point to sort the data dimensions with ascending or descending order (data-driven) to reveal patterns between records [35].

Correlation- and similarity-driven strategies improve star glyphs by adjacent placement of similar axes to support understanding of clusters, outliers, and relationships [7]. Ankerst et al. propose heuristic algorithms based on similarity for star glyphs to improve the overall perception [3]. Similarly, Artiero et al. use similarity heuristics of attributes to apply dimension-ordering and take perceptual aspects as Gestalt Laws into account.
by applying dimension reduction [4]. Yang et al. combine similarity-based ordering with a hierarchical structure of the dimension to enable interactive exploration of high-dimensional subspaces [37]. Friendly and Kwan argue that using correlation-based ordering in star glyphs supports the identification of shape irregularities [15]. The authors did not conduct a survey to underpin their statement. **Spikes and salient shapes** such as “has-one-spike” are helpful in visual grouping tasks of data points according to Klippel et al. [24]. They argue, that concavity is more suitable for comparability than convexity, which is especially true for the “star” glyph, due to the large variations between adjacent dimensions. The salience of dissimilar neighboring axes shall enhance the comparison speed and help to detect changes. Klippel et al. showed that the star glyph shape with eight dimensions influences classification tasks [25]. Especially, in contrast to earlier work that state that similarity-driven orderings improve high-dimensional visualizations, dissimilarity between neighboring axes contribute salient properties that are perceptually more noticeable. **Symmetry-driven** reordering methods help to reduce the visual complexity of star glyphs and, therefore, support comparison tasks by improving memorability [22]. By providing some examples, Peng et al. argue that orderings with simple and symmetric as well as monotonic shapes of the star glyphs facilitate the identification of value differences between multiple dimensions [29]. They emphasize that symmetry and similarity are primary factors to identify patterns. For this, Gestalt Laws are a solid foundation for perception design [36]. Peng et al. state that star glyphs can be optimized by aligning the symmetry on the vertical or horizontal axis [29]. An additional rotation optimization step can be included in the pipeline to find the best global rotation for all star glyphs of a dataset. Rotation can be applied, e.g., on top of similarity or spike-based ordering.

### 2.2 Empirical Studies and Research Gap

While many ordering strategies, algorithms, and heuristics have been proposed star glyph dimension ordering, empirical evaluation is missing. Previous approaches mainly argue by showing examples or providing arguments w.r.t. to e.g., Gestalt laws. While this is useful to find differences between strategies, we also need empirical evidence to directly compare strategies respecting scalability, performance, analysis tasks, data characteristics, and user perception [35].

We are only aware of two studies conducted by Klippel et al. [24, 25]. Their results indicate that spikes and salient shapes have a positive effect on visual grouping tasks and colored axes positively affect the processing speed and reduce the negative influence of shape saliency on rotated data glyphs. However, Klippel et al. did not directly compare different ordering strategies or evaluated them against a benchmark. Instead, they designed different star glyph shapes and analyzed how participants grouped them by their understanding of similarity during an exploratory analysis task. In our study, we aim to close this research gap by comparing two proposed reordering strategies using a controlled, empirical user experiment.

### 3 Empirical User Study

We evaluate whether a similarity- (homogeneous shape, short: SIM) or dissimilarity-based layout (spike and salient shape, short: DIS) is more efficient and effective for a visual clustering task. We designed our study based on Klippel et al.’s work [24, 25]. We adopted the task, user interface, glyph design (including colored axes), and datasets’ dimensionality (eight dimensions). However, in contrast, we applied two different reordering algorithms (SIM and DIS) and different clutter levels as independent factors, and evaluate the results using a benchmark dataset.

#### 3.1 Experimental Design and Hypotheses

The participants had to manually assign star glyphs into reasonable clusters and identify noise, i.e., items not belonging to any cluster. In the study, we used the term group instead of cluster. To assess the performance, we use four dependent variables as quality measures: (i) task completion time, (ii) quality of groups, (iii) quality of identified noise, and (iv) the confidence of the participants.

**Participants.** We recruited 15 participants from the local student population (seven female, two bachelor, twelve master, one PhD student). The age ranged from 20 – 27 years with a median of 23. The participants had a different background in data analysis and visualization: ten had general knowledge in data analysis, four had data visualization experience, and one has used star glyphs before. All participants received a compensation of 10 EUR.

**Glyph Design and Implementation.** The glyphs are designed analog to Klippel et al.’s work [24, 25] using a contour, gray background, and colored axes. We used ColorBrewer [19] to select diverging colors and applied the ordering algorithm by Ankerst et al. [3]. The Euclidean distance is used to measure the (dis)similarity between dimensions. We ran an exhaustive search to find the permutation with the highest (SIM) and lowest (DIS) similarity. An example of star glyphs with the two orderings is depicted in Fig. 1. Orientation (rotation) of the star glyphs is not considered and chosen randomly. All orderings are pre-computed not to influence the run time during the study. We provide the study and the ordering strategy implementation on our websites.

**Hypotheses.** We address the following two hypotheses:

H1. Clutter negatively influences visual comparison. With increasing clutter, the performance of grouping tasks drops, independent of the axes ordering. In particular, we expect users to be (a) slower, (b) less accurate in grouping accuracy, (c) less accurate in noise identification, and (d) less confident of their grouping.

H2. Klippel et al. [24, 25] argue that spikes and salient shapes support users in similarity estimation and grouping tasks. Therefore, the performance of users should increase with a dissimilarity-based ordering. Furthermore, we hypothesize that the salient shapes should support users even more if the data contains clutter since sharp edges are more perceptually apparent. In particular, we expect users to be (a) faster, (b) more accurate in grouping accuracy, (c) more accurate in noise identification, and (d) more confident of their grouping when dimensions are ordered by dissimilarity.

#### 3.2 Benchmark Datasets

We manually created 18 different datasets using the PCDC tool [8]. Every dataset contains 50 records of which 2–7 data points are selected as noise (randomly distributed across all dimensions). The remaining data points are grouped into 2, 3, or 4 clusters with similar cluster sizes. Besides, we introduced clutter dimensions which do not discriminate any cluster, since we uniformly distributed all data points across the clutter dimensions. 6 datasets contain no clutter (0C), 6 one- (1C), and 6 two clutter dimensions (2C). For instance, in condition 2C a dataset consists of six dimensions discriminating.

3 E

http://subspace.dbvis.de/sg-study and */sg-ordering.
which do not belong to any group (noise).

Tasks and Procedure. Each study took an hour on average. Participants filled out a consent form, demographics, and report on previous knowledge in data analysis, information visualization, and star glyphs. Afterward, we described how to read the visual encoding of a star glyph using an artificial car dataset as an example. Specifically, we clarified that star glyphs with similar shapes on different axes are not similar (rotation invariance). Finally, the participants performed three training trials before the study was recorded.

To conduct the study, we used a 27-inch screen with 2560x1440 resolution and a mouse to execute given tasks. Every participant had to perform 18 trials, leading to 15 participants $\times$ 18 trials $= 270$ trials for the entire study. In between two trials, we showed a blank screen with the term ‘break’ to motivate the participants to have regular breaks. Each trial consisted of manually grouping all 50 star glyphs of one dataset into distinct groups and noise. Then, the participants stated the confidence level about their selection on a 7-point Likert-scale. We did not provide the number of clusters per dataset and explicitly told the participants that there might be glyphs which do not belong to any group (noise).

Fig. 2 shows our interface. Participants were able to add new or delete groups. Glyphs can be interactively assigned to groups by drag&drop. If a glyph was considered to be noise, then it remained in the left panel. Participants were able to undo or change a grouping also using drag&drop. In the study, we did not constrain the task completion time. We ended the study with an interview about the participants’ strategy and preferences regarding the SIM and DIS ordering by showing examples. Questions and answers were recorded.

Randomization. Each participant performed 18 trials, i.e., the grouping task on all benchmark datasets was equally distributed between SIM and DIS. We randomized the order of the trials as follows: First, we grouped the datasets into their level of difficulty based on the amount of clutter ($0C, 1C, 2C$). Then, participants performed the trials with increasing difficulty, i.e., $6 \times 0C$, then $6 \times 1C$, and finally $6 \times 2C$. For every clutter condition, we randomized the dataset order and randomly assigned $3 \times$ SIM and $3 \times$ DIS. We attached the randomization algorithm and our configuration in the supplementary material. A summary of our trials:

<table>
<thead>
<tr>
<th>Levels of Difficulty</th>
<th>Ordering Strategies (SIM, DIS)</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3$</td>
<td>$\times$</td>
<td>$15$</td>
</tr>
<tr>
<td>$2$</td>
<td>$\times$</td>
<td>$3$</td>
</tr>
<tr>
<td>$1$</td>
<td>$\times$</td>
<td>$270$ trials in total</td>
</tr>
</tbody>
</table>

Data Collection, Post-Processing, and Analysis. In each trial, we recorded the grouping task completion time, the selected groups and noise, and the participants’ confidence. Some participants created groups with only one or two glyphs. Thus, in a post-processing step, we converted such small groups into noise to execute a more coherent analysis. We measured the quality of the identified noise by computing the Jaccard index between noise and ground truth noise.

The grouping quality is also based on the Jaccard index between the grouping and ground truth. However, since participants could have also selected too few or too many groups, we structured our quality computation as a two-step process: First, we computed the average Jaccard index of each group to its best match in the ground truth. Second, we computed the average Jaccard index of every ground truth cluster to its best match in the selection. Using this method, we considered too few, too many groups, as well as too few and too many records per group. The final clustering quality is the average score of both steps.

3.4 Results and Statistical Analysis

We executed a statistical analysis to summarize the study results. We report all statistically significant findings ($p < .05$) and some interesting trends visible in the data. We check for normal distribution using a one-sample Kolmogorov-Smirnov test. For a better comparison, we always report the median ($\bar{x}$) and, additionally, the mean ($\mu$) for normally distributed samples. Analysis scripts and detailed results can be found in the supplementary material.

Statistical tests. Confidence is measured as Likert-scale. Therefore, a Pearson’s Chi-squared test is used for the analysis. Given the non-normal nature of the measures time, cluster quality, and noise identification quality w.r.t. $0C, 1C, 2C$, we used a non-parametric Friedman’s test and a Wilcoxon signed rank test with Bonferroni correction (Post-hoc). The same measures do not follow a normal distribution w.r.t. the strategies SIM and DIS. Hence we used a Wilcoxon signed rank test with continuity correction. Considering SIM and DIS within $1C$ and $2C$ reveal normal distributed samples for the measures time, quality of clustering, and quality of noise identification. Hence, we use a paired t-test for the statistical analysis.

Task Completion Time.

H1a. Task completion time increased with clutter levels, but not significantly: $0C (\bar{x} = 162.5s), 1C (179.5s), 2C (184.0s)$.

H2a. Using the ordering DIS $(176.0s)$ users completed the grouping task slightly faster than SIM $(178.0s)$, but only for datasets with clutter dimensions. $0C$: DIS $(168.0s)$ vs. SIM $(158.0s), 1C$: $180.0s / 179.0s$, and $2C$: $180.0s / 190.0s$. Differences are not significant.

Cluster Quality.

An overview of the cluster quality is depicted in Fig. 3.

H1b. There were significant effects of clutter level on cluster quality ($\chi^2 (2, N = 270) = 109.92, p < .001$). Post-hoc tests showed a higher participants’ accuracy with $0C (\bar{x} = 0.85)$ compared to $1C (\bar{x} = 0.67, p < .001)$ and $2C (\bar{x} = 0.55, p < .001)$, and between $1C$ and $2C (p < .001)$.

H2b. When comparing ordering strategies, participants were more accurate with DIS ($\bar{x} = 0.69$) compared to SIM ($\bar{x} = 0.67, p < .05$), which is also true within clutter levels $1C$ and $2C$, but not $0C$: DIS ($\bar{x} = 0.85, \mu = 0.81$) vs. SIM ($\bar{x} = 0.85, \mu = 0.82$), 1C: DIS ($\bar{x} = 0.68, \mu = 0.66$) vs. SIM ($\bar{x} = 0.66, \mu = 0.63$). 2C: DIS ($\bar{x} = 0.58, \mu = 0.57$) vs. SIM ($\bar{x} = 0.52, \mu = 0.50, p < .001$).
Noise Identification Quality.  

H1c. There was a significant effect of clutter level on noise identification ($\chi^2(2, N = 270) = 80.02, p < .001$). Post-hoc tests revealed that participants were more accurate with $\text{DIS}$ ($\bar{x} = .8$) compared to $\text{1C}$ ($\bar{x} = .5, p < .001$) and $\text{2C}$ ($\bar{x} = .33, p < .001$). In addition, there was a significant effect between clutter conditions $\text{1C}$ and $\text{2C}$ ($p < .001$). However, we could also see that, without clutter, participants were faster and more confident using a similarity-based ordering. The remaining question is whether it would be possible to combine $\text{SIM}$ and $\text{DIS}$ into a combined ordering strategy. Our study did not reveal whether participants need as many spikes as possible or whether a few important spikes are enough to improve the cluster quality. Further research needs to be done in this area. Another relevant question is also how the rotation of entire glyphs influences grouping quality in clustering tasks and further investigation in this direction is advisable as, for example, already started by Fuchs et al. [16].

Design Considerations. With the results gained from our study, we derive the following design considerations:

1) As the performance of users drop considerably when clutter dimensions are present, glyph designers should try to avoid clutter by applying a feature selection method first, if possible.

2) Since, for datasets with clutter, salient shapes and spikes support grouping tasks, we recommend using DIS strategies.

3) For datasets without clutter, we did not find a clear difference between $\text{SIM}$ and $\text{DIS}$. As $\text{SIM}$ seem to be slightly faster and less error prone to rotation [24, 25]. We recommend to use this strategy.

Limitations. We identified two main threats to our results’ validity.

1) The number of trials (270) is rather small, in particular, for the effectiveness and efficiency analysis of a specific clutter level. This affects not only the statistical analysis, but outliers may also distort the results. The number of trials per participant cannot be increased with the current study design; otherwise, the study would take much longer than one hour. Therefore, we suggest repeating the study with more participants to increase the number of trials.

2) While we designed our datasets with different cluster structures and distributions, we limited them by eight dimensions as Klippel et al. [25]. There might be differences for datasets with less, more, or an odd number of dimensions.

5 CONCLUSION AND FUTURE WORK

We conducted an empirical user study to evaluate the impact of clutter and axes ordering to clustering performance with star glyphs. Our results show that users perform better when the glyphs represent salient shapes and spikes, which is achieved by a dissimilarity-based ordering of the dimensions. Furthermore, we elicited that there is a significant impact of clutter on the clustering performance in general.

As future work, we plan to extend and re-run the study based on our discussed limitations and include other reordering strategies, as well. Extending to that, we want to investigate whether there is an influence of the data characteristics and rotation (e.g., favor symmetrical glyph shapes) to the ordering strategy. If so, we are interested in developing techniques to select the most useful ordering strategy based on the given data and task. Finally, automatic ordering strategies should be compared to user-driven axes arrangements, which are determined by experts based on their domain knowledge.

ACKNOWLEDGMENTS

We thank David Pomerene and the anonymous reviewers for their valuable feedback and support. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 251654672 – TRR 161 (Project A03).