© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Improving the Robustness of Scagnostics

Yunhai Wang, Zeyu Wang, Tingting Liu, Michael Correll,
Zhanglin Cheng, Oliver Deussen, Michael Sedlmair

Abstract— In this paper, we examine the robustness of scagnostics through a series of theoretical and empirical studies. First, we investigate the sensitivity of scagnostics by employing perturbing operations on more than 60M synthetic and real-world scatterplots. We found that two scagnostic measures, *Outlying* and *Clumpy*, are overly sensitive to data binning. To understand how these measures align with human judgments of visual features, we conducted a study with 24 participants, which reveals that i) humans are not sensitive to small perturbations of the data that cause large changes in both measures, and ii) the perception of clumpiness heavily depends on per-cluster topologies and structures. Motivated by these results, we propose *Robust Scagnostics* (*RScag*) by combining adaptive binning with a hierarchy-based form of scagnostics. An analysis shows that *RScag* improves on the robustness of original scagnostics, aligns better with human judgments, and is equally fast as the traditional scagnostic measures.

Index Terms—Scagnostics, scatterplots, sensitivity analysis, Robust Scagnostics

1 Introduction

Visual quality measures are useful tools for algorithmically assessing visual patterns in data [9, 13]. A prominent example of such measures are the *scagnostics* measures [49, 50] that characterize 2D distributions in scatterplots based on their geometric features. Various visualization tools and techniques have been built upon such measures [4, 19, 20].

While the field has recently begun to evaluate visual quality measures in more detail [9, 37, 39], the *robustness* of these metrics remains underexplored. That is, do scagnostic measures reliably pick out visual patterns of interest in scatterplots, even under noise or the presence of adversarial structures? In order to promote robustness, Wilkinson and Wills [53] proposed criteria that must be met by candidate scagnostics (such as "they should be sensitive to differences in 2D point distributions" and "they should be on a common scale"). While a large-scale evaluation of these factors on synthetic data appears to confirm that scagnostic measures have these properties [53], we identified several issues when working with these measures in practice. Our hypothesis is that synthetic data may not fully capture how scagnostic measures vary, and that more realistic data may reveal robustness concerns [39].

To fill this gap, we present an in-depth study of the sensitivity of scagnostic measures, both theoretically and experimentally. In particular, we examine how much the output values (of scagnostic measures, or human judgments concerning these measures) vary as the result of variations in the input (i.e., changes to individual scatterplots). The results of our studies help us to better characterize potential sensitivity issues of scagnostic measures and to design alternative measures that are robust w.r.t. these issues. To do so, we first augment the data used by Wilkinson and Wills [53] with samples that have a wider variety of cluster characteristics [39]. This extension helps us to test a broader set of visual patterns. In total, we constructed a data set with 60 million synthetic scatterplots and 69K scatterplots obtained from real data.

Based on these scatterplots, we conducted a structured sensitivity analysis of the two different versions of scagnostics published in [50] and [52]. We begin by studying the measures' sensitivity with regard to the two pre-processing steps of data binning and deleting outliers. We then perform *perturbing* operations, such as deleting random points and rotating the scatterplot. We chose these perturbing operations to create relatively little change in the visual structure of the data, so from a human perspective, there ought to be little change in the visual patterns present in the scatterplots. We would then hope that scagnostics, which are meant to capture important visual patterns, would be invariant to these perturbations.

Investigating scagnostics' sensitivity to these changes revealed four main findings. (1) The data binning procedure (a vital pre-processing step to make computation performant on large scatterplots) can result in large changes to all scagnostic measures. (2) The newer *Outlying* measure in [52] (i.e., *Scag-06*), designed to detect both exterior and interior outliers, is less robust than the older *Outlying* measure in [50] (i.e., *Scag-05*). However, the older method lacks the ability to detect interior outliers. (3) Under certain conditions, the *Clumpy* measure does not accurately represent the characteristics of distributions with multiple clusters. (4) All measures except *Outlying* and *Clumpy* are sensitive to deleting outliers but robust to other perturbing operations.

To better understand the large sensitivity of the *Outlying* and *Clumpy*, we furthermore conducted a user study investigating how well human judgments align with these scagnostic measures. From this study we learned that (1) humans are relatively insensitive to small perturbations in scatterplots when assessing outlyingness and clumpiness and (2) cluster-specific densities heavily influence human perception of clumpiness, while the number of clusters has a smaller effect.

Motivated by the study results, we propose *Robust Scagnostics* (i.e. *RScag*) which capture the spirit of the original scagnostic measures, but are designed for additional robustness. *RScag* consists of two major components: an *adaptive binning* approach and *hierarchy-based scagnostics*. Adaptive binning preserves underlying data densities more faithfully than the original scagnostic hexagon binning approach, while hierarchical scagnostics computes measures on local clusters, allowing more flexibility for representing different numbers of clusters and cluster densities. We evaluate *RScag* on our collection of scatterplots and our human response data. The findings indicate that *RScag* outperforms existing scagnostic measures with respect to numerical and perceptual robustness.

In summary our main contributions are:

- we construct a large set of 60M synthetic and 69k real-world scatterplots by expanding existing data distributions, our code and dataset are available at github[1];
- we conduct a sensitivity study for the nine scagnostic measures and two pre-processing steps (data binning and deletion of outliers), discovering patterns of unexpected sensitivity; and

- *Y. Wang, Z. Wang, T. Liu are with Shandong University. E-mail: {cloudseawang, zywangx, sduhammer}@gmail.com.*
- *Z. Wang and Z. Cheng are with Shenzhen VisuCA Key Lab, SIAT, China. E-mail: zl.cheng@siat.ac.cn (corresponding author).*
- *M. Correll is with Tableau Research. E-mail: mcorrell@tableau.com.*
- *O. Deussen is with Konstanz University, Germany and Shenzhen VisuCA Key Lab, SIAT, China. E-mail: oliver.deussen@uni-konstanz.de.*
- *M. Sedlmair is with VISUS, University of Stuttgart, Germany. E-mail: michael.sedlmair@visus.uni-stuttgart.de.*
- *Y. Wang and Z. Wang are joint first authors.*

Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.
Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2019.2934796

[1] https://github.com/ArranZeyuWang/RScag

1077-2626 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

- we propose and evaluate a novel robust scagnostics approach, *RScag*, to preserve more structural information and to be more consistent with human judgments.

2 Related Work

A complete review of visual quality measures can be found in Bertini et al. [13] and Behrisch et al. [9]. Here we restrict our discussion to measures designed for scatterplots and their evaluation.

2.1 Scatterplot Quality Measures

Scagnostics were originally proposed by John and Paul Tukey in the 1980s [49] for identifying interesting scatterplots from large scatterplot matrices by algorithmically characterizing certain visual patterns. Based on that, Wilkinson et al. [50, 52] developed nine computationally efficient scagnostic measures using planar proximity graphs, and demonstrated their utility for selecting interesting scatterplots, sorting scatterplot matrices, and identifying outliers. Seo and Shneiderman [40] present a similar idea called rank by feature, but it is based on classical statistics (means, medians, correlations, etc.) rather than the Tukeys' non-parametric measures. Because of the usefulness of such quantitative measures, scagnostic-type measures have also been developed for other plots [8, 22, 32], and extended to different types of patterns [34].

Scagnostics' main application is to guide the interactive exploration of complex data. For example, Anand et al. [5] use them to explore interesting low-dimensional random projections of high-dimensional data. Dang et al. [19] apply scagnostics to identify interesting subsequences from multivariate time-series data. Hafen et al. [25] use them to sample panels from a trellis display, Anand and Talbot [4] to select good partitioning variables for small multiple displays, and Dang and Wilkinson [21] to choose appropriate data transformations.

Many other visual quality measures exist for scatterplots that follow the spirit of scagnostics. Some of the first measures that have been proposed were Tukey's area of the peeled convex hull [47], Silverman's kernel density isolevel contours [43], as well as related measures by Hastie and Stuetzle [27]. Today, for different purposes a variety of different measure types exits. In terms of scatterplots, visual clutter measures [10, 12], correlation measures [26, 28], and visual cluster separation [1, 6, 39, 44, 45] have gained much attention.

Among them, the visual cluster separation measures are most related to our work, since they aim to characterize the cluster characteristics in distributions similar to the *Clumpy* measure in scagnostics. Aupetit and Sedlmair [6] propose a general framework to construct such visual cluster separation measures, and their quantitative evaluation showed that local density-based measures outperform other measures, a finding that was further confirmed by Shao et al. [41, 42]. In line with these findings, our proposed robust scagnostics are also based on local density and compute each measure in terms of local clusters.

2.2 Evaluation of Scatterplot Quality Measures

Visual quality measures can be evaluated through human subjects studies, sensitivity analysis, or use-case scenarios. Since use-cases are application-dependent, we concentrate on the first two study types.

Human Subjects Studies. Various studies rely on human judgments to assess the nature and strength of visual patterns of interest in charts. The human judgments are compared to the corresponding quality measures of these patterns. Ideally, human judgments and statistical quality measures would be tightly correlated. Using this approach, Sips et al. [44] evaluated measures for class separability in scatterplots. Their results indicated a good correlation between the proposed measures and human judgments. Tatu et al. [46], and Lewis et al. [33] also studied class separation in controlled user studies; both studies found that some measures contrast with human judgments while others align relatively well.

Instead of asking a few people to observe many datasets, Sedlmair et al. [38, 39] set out to conduct a *data study* in which class separation is judged by a small number of trained *experts*. The study revealed that the tested measures failed in almost 50% of the cases under these more realistic conditions. In a follow-up work, they used this carefully collected human data as an input to a machine learning framework [37]. This framework was then used to automatically evaluate and compare how well measures predict human judgments in both existing measures [37] and the new ones they proposed [6]. Recently, Behrisch et al. [8] conducted a similar data study in order to systematically evaluate and rank measures for adjacency matrices.

The closest methodologies to our work are in studies done by Lehmann et al. [31] and Pandey et al. [35]. The former study compared the consistency between filtering relevant scatterplots based on human perception versus selecting them by a subset of scagnostic measures and shows that the selected scagnostic measures outperform the other measures [2, 30, 45]. Pandey et al. [35] conducted a study, in which users had to group sets of scatterplots according to their subjective judgment of similarity. Comparing the results with the nine scagnostic measures, they concluded that the measures do not align well with factors that humans would take into account for their similarity judgments. This result prompted us to investigate if there are additional data factors not captured in existing scagnostic measures.

Sensitivity Studies. A sensitivity analysis [36] refers to quantifying the change in outputs due to small perturbations of the inputs. A good quality measure should be insensitive to small input changes but sensitive to large ones. A few methods [7, 15–18] have been proposed to compute the sensitivity information of specific visualization processes and augment visualizations with such information. Here, we mainly focus on the ones developed for quality measures.

Wilkinson and Wills [53] created a large set of synthetic datasets and selected a few real datasets to understand the distribution of their scagnostic measures. The study shows that their measures are sensitive to distributional changes, but it is unclear how sensitive the measures are to small changes of different data factors. Furthermore, they did not test how human judgments are sensitive to such data changes. In contrast, Behrisch et al. [8] evaluated the sensitivity of Magnostics by observing how such measures change as different levels of noise are added to the data. Similar to that approach, we add different levels of perturbation to different data factors and observed how much they influence scagnostic measures. Additionally, we examine human judgments under perturbation and compare them to our algorithmic results.

3 Background: Graph-Theoretic Scagnostics

In this section, we briefly review scagnostics [50, 53] including its pipeline, basic geometric graphs, preprocessing steps, and the nine measures. In particular, we examine the relationship between scagnostic measures and the basic geometric graphs and highlight the connections between the involved preprocessing steps and the different measures. For a full description of the measures we refer the reader to the original papers [50, 53] or our supplementary materials.

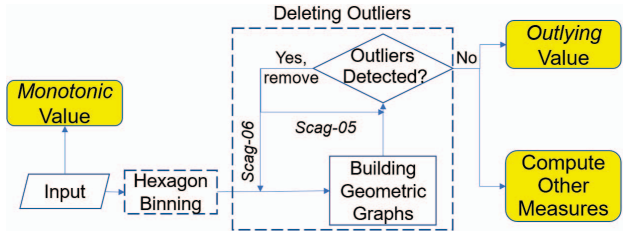

Fig. 1. Pipeline for computing scagnostics, where the *Outlying* value is obtained after all outliers are removed, and then the other measures are computed.

3.1 Algorithm Pipeline

Scagnostics provide nine measures for characterizing different patterns: *Outlying*, *Skewed*, *Clumpy*, *Convex*, *Skinny*, *Striated*, *Stringy*, *Sparse*, and *Monotonic*. As shown in Fig. 1, computing most measures except *Monotonic* involves two preprocessing steps: data binning and deleting

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

outliers. Since geometric graphs need to be re-built once outliers are detected and removed, building geometric graphs is also a core step.

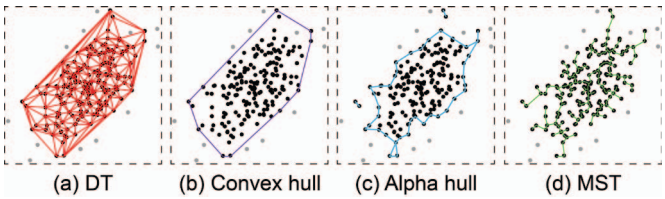

Fig. 2. Examples of four geometric graphs built after deleting outliers highlighted in gray.

Data Binning: To improve the performance of scagnostics, hexagon binning [14] is used to reduce data while preserving data characteristics. Starting with a 40×40 hexagonal grid, the points of a scatterplot are binned and checked to see if the number of non-empty cells is more than 250. If so, the points will be re-binned with a twice as coarse bin size until there are no more than 250 nonempty cells.

To attenuate the influence of data binning in scagnostic measures, Wilkinson et al. [52] suggest using a weighting function w to adjust the measures *Skewed, Sparse* and *Convex*:

$$w = 0.7 + \frac{0.3}{(1+t^2)}\,, \tag{1}$$

where $t = n/500$ and n is number of points.

Building Geometric Graphs: Scagnostics are based on geometric graphs. The Delaunay Triangulation (DT) is constructed first, and then a Minimum Spanning Tree (MST), convex hull, and alpha hull are built based on the DT. MST, convex hull and alpha hull all are subgraphs of the DT, although they are defined using different criteria. By setting the value of α to the 90th percentile of the MST edge lengths [53], the formed alpha hull does not include sparse or striated point sets, see black points outside of alpha hull in Fig. 2 (c).

Detecting Outliers: To improve the robustness of scagnostic, outliers are deleted before computing the measures. Following Tukey [48], a potential outlier is a point whose adjacent edges in the current MST have edges larger than ω:

$$\omega = q_{75} + 1.5(q_{75} - q_{25}), \tag{2}$$

in which q_i refers to the i-th percentile of the sorted edge lengths of the MST. After deleting outliers, the output is the *Outlying* value and the updated geometric graphs (examples are given in Fig. 2). The convex hull and alpha hull are used for computing *Convex* and *Skinny* values, while the MST is used to compute the other measures.

3.2 Scagnostic Measures

The nine scagnostic measures reveal many hidden features such as density, shape, or association level in the input scatterplot [50]. In the following, we mainly review the *Outlying* and *Clumpy* measures, which are most relevant to our findings. Note that a measure called *Straight* appeared in *Scag-05* [50], but was removed in *Scag-06* [52]; hence, we did not test it.

Outlying Measure: The *Outlying* measure indicates the impact of outliers on the data. Based on the edge lengths of the MST, it is defined:

$$c_{outlying} = length(T_{outliers})/length(T) \tag{3}$$

where $length(T)$ is the total length of edges in the initial MST and $length(T_{outliers})$ measures the total length of edges adjacent to outliers.

In the definition of the *Outlying* measure from *Scag-05* [50], a point is classified as an outlier if it satisfies the condition in Eq. 2, but also has a degree of one. This additional condition prevents the measure from detecting interior outliers (see v_2 in Fig. 3 (a)). To address this issue, the newer definition of the *Outlying* measure in *Scag-06* [52] ignores this condition, but might therefore remove additional points.

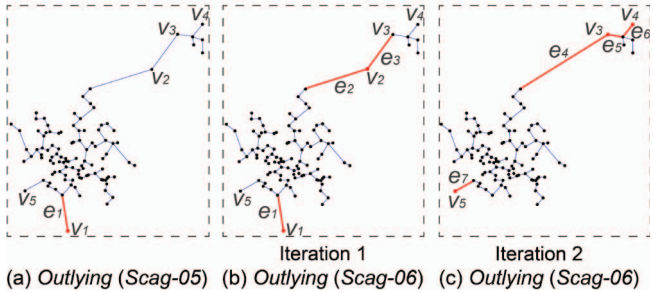

Fig. 3. Differences between the definition of the *Outlying* measure in *Scag-05* (a) and *Scag-06* (b,c). (a) Only the point v_1 is removed by the older version of outlying. In *Scag06*, v_1 and v_2 are removed in the first iteration (b) and v_3, v_4 and v_5 are further removed in the second iteration (c) , which results in a new long edge of the graph.

Fig. 3 (b,c) shows an example, where v_1 and v_2 are deleted in the first iteration and then three non-outlier points v_3, v_4, and v_5 are further removed at the second iteration. Deleting these points results in a newly formed long edge e_4. As per the pipeline shown in Fig. 1, different versions of the *Outlying* measure can result in different values of the other scagnostic measures.

Clumpy Measure: This measure depicts the clustering of data points based on the edge lengths of the MST. It is obtained by testing each edge e_j with the following procedure:

- remove edges which are longer than e_j;
- select two point subsets linking to the vertices of e_j;
- find the longest edge e_k from the smaller subset; and
- compute the *Clumpy* value by:

$$c_{clumpy} = 1 - length(e_k)/length(e_j). \tag{4}$$

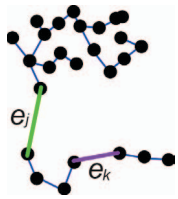

Fig. 4. Two edges used for computing the *Clumpy* measure.

After iterating over all edges, the maximal value is taken as the output of *Clumpy* measure. This measure is therefore built on the assumption that the data consists of two clusters and only takes into account the intra-cluster distances and the largest inter-cluster distance within the small cluster (see Fig. 4). Such a definition is not able to accurately characterize certain cluster structures in some scatterplots.

3.3 Existing Limitations

From the above brief review, we identified three computational aspects of scagnostics that might be further improved:

- It is unclear how binning impacts the robustness of the final scagnostic measures, even when considering the included weight function– we examine the effect of binning in Sec. 4.2;
- Both versions of the *Outlying* measure have drawbacks and it is unclear which one is more robust; and
- the *Clumpy* measure is determined by two edges, which might not accurately characterize patterns within complex distributions [41].

To address these limitations, we perform a sensitivity analysis of scagnostic measures to assess their robustness (Section 4) and conduct a user study to assess the consistency between human judgments and scagnostics (Section 5). Based on the results of these two studies, we propose a new robust set of scagnostic measures for better capturing a wide variety of data patterns (Section 6).

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

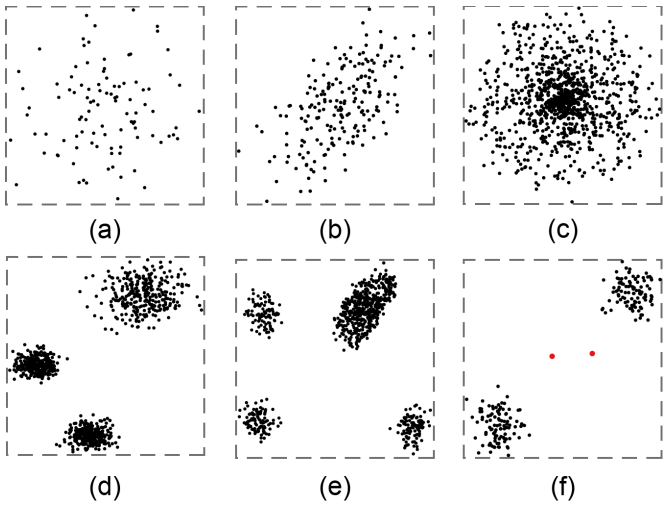

Fig. 5. Examples of scatterplots with different cluster-specific characteristics generated by *Binormal* distributions: (a) a circular distribution; (b) a rotated elliptical distribution; (c) a circular distribution with varying densities; (d) a three-cluster distribution with varying cluster sizes; (e) a four-cluster distribution with varying cluster densities and shapes; (f) the two-cluster distribution with interior outliers (shown in red).

4 Scagnostics Robustness

In this section, we perform a sensitivity analysis to test the robustness of scagnostic measures and identify the factors influencing this robustness. We use the R implementation of Scagnostics provided by Wilkinson and Anand [51]. This package contains all *Scag-05* and *Scag-06* measures except the *Scag-05 Outlying* measure, which we re-implemented in R.

First, we test how strongly the data binning algorithm influences the resulting scagnostic measures, with the goal of removing the binning step from our future sensitivity analyses if it has an outsized influence on the resulting measure. We then conduct our main sensitivity analyses across both versions of scagnostic measures on various datasets.

4.1 Data Augmentation

To simulate a variety of 2D point distributions, we first generated a large number of scatterplots using ten 2D point distributions as in Wilkinson and Graham [53]. However, they only considered some simple distributions, while we were interested in more complex data features; in particular, those known to impact visual cluster separation [39].

To address this gap, we sampled a large parameter space of the *Binormal* distribution to generate both single and multi-cluster scatterplots with varying cluster-specific characteristics. Since this distribution enables us to control the size, density and shape of clusters with different parameters, and to adjust the number and comparative distance of distributions, many within-cluster and between-cluster data factors [39] are incorporated into the data. In addition, we randomly placed a few interior and exterior outliers into our multi-cluster scatterplots to simulate contaminated data. In doing so, we generated around 800K scatterplots, which are non-perturbed scatterplots used for sensitivity analysis (see Section 4.3). More details about our parameter space sampling can be found in the supplementary materials. Fig. 5 shows six typical examples with variations in cluster size, shape, density and outliers.

Besides our synthetic plots, we intended to gather more realistic data. We first collected 1703 real-world datasets from various sources [24, 35, 39]. Since most of these are high dimensional data, we created scatterplots for each combination of two dimensions as well as creating additional scatterplots through standard dimensionality reduction techniques [37]. In total, we created 69K real-world scatterplots with a wide variety of shapes, number of points, and sizes.

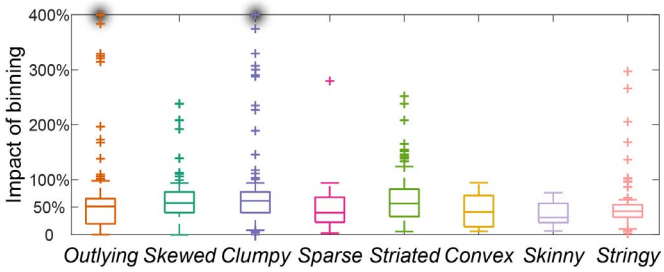

Fig. 6. Relative changes of various measures in *Scag-06* as a result of including or omitting the binning step. In case a result is out of the plot range (see *Outlying* and *Clumpy*), we draw a dark transparent shadow to indicate the amount.

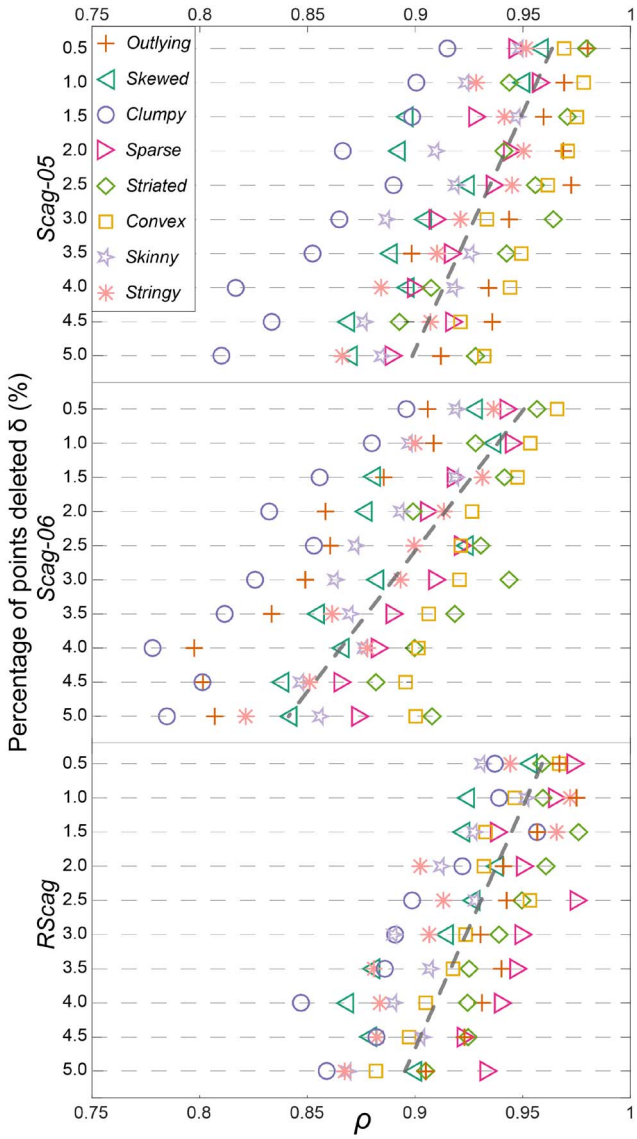

Fig. 7. Dot plot showing Spearman's rank correlation coefficient ρ of eight scagnostic measures in *Scag-05*, *Scag-06*, and our proposed *RScag* (see Section 6) obtained by applying random deletion operations to our scatterplot dataset. The regression trend lines of ρ values except *Outlying* and *Clumpy* measures are shown in gray dashes. Values close to 1 indicate that the measure was relatively consistent across plots even after points were deleted.

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

4.2 Binning Effect

To study the influence of data binning on the scagnostic measures, we compute the relative changes, of the *Scag-06* [52] scagnostic measures on our scatterplot dataset both with and without binning:

$$\text{Impact of binning} = \frac{|sm_{w/o} - sm_{w/}|}{sm_{w/o}} \quad (5)$$

where $sm_{w/}$ and $sm_{w/o}$ are the values of the scagnostic measures with and without binning, respectively. Since the *Monotonic* measure is computed from the input data, we do not consider it in this study.

Fig. 6 summarizes the changes of the eight remaining measures. We can see that all of them have large changes, where the ratios of *Outlying* and *Clumpy* measures both exceed 400% for some data, while the median of the change in the *Clumpy* measure is 63%. Binning has a similar influence on the earlier scagnostic measures, *Scag-05* [50]; this result is shown in the supplementary materials. This observation is inconsistent with the one shown by Wilkinson et al. [52]. Because of this large sensitivity, we exclude the binning step in our remaining sensitivity study and compare measures based on the full scatterplots.

4.3 Sensitivity Analysis

To assess the sensitivity of scagnostic measures, we study the impact of small perturbations in the scatterplots. Ideally, minor perturbations should result in only minor changes to the resulting measures.

Perturbing Scatterplots: We employ two operations to perturb scatterplots: i) randomly deleting a percentage of δ data points 10 times which results in 10 plots for each δ or ii) rotating the entire plot by θ degrees which outputs one plot for each θ. By setting $\delta = \{0.5, 1, 1.5, ..., 5\}$ and $\theta = \{1, 5, 15, 45, 90, 180\}$, we generated around 60 million scatterplots using the real and synthetic data introduced in Sec. 4.1. We do not employ other perturbations such as scaling and translation here, because the MST is translation and scale invariant [29].

Quantifying Sensitivity: Since values of some of the scagnostic measures (e.g. *Outlying* and *Clumpy*) are typically quite small [53], value changes of these measures might not be able to clearly indicate their sensitivity. Therefore, we use the *rank* instead of the numerical *value* [3] to analyze their sensitivity.

Given a set of scatterplots $\{s_1, \cdots, s_n\}$, we compute the measures for each scatterplot and then rank the results in terms of one specific scagnostic measure. For each measure and all associated scatterplots, we compute the Spearman's rank correlation coefficient $\rho \in [-1, 1]$. If ρ is 1 (i.e., the ranking-based scatterplot aligns with curve $y{=}x$), the ranks are perfectly correlated and the measure is insensitive to the perturbation; if ρ is far from 1, the measure exhibits more sensitivity. In our experiment, ρ is always larger than 0, see an example in our supplementary materials.

Results: Fig. 7 shows the ρ values of eight scagnostic measures of *Scag-05* and *Scag-06* generated by randomly deleting different amounts of points. In general, we can see that when δ (the percentage of deleted points) is not larger than 3%, most measures of *Scag-05* are larger than 0.95, while the threshold is 1.5% for *Scag-06*. Through a closer inspection of the results, we make the following additional three observations:

(1) *The ρ values of the Clumpy measure are smaller than the others in both versions of scagnostics.* — This observation indicates that, unlike in the prior sensitivity analyses [53], the *Clumpy* measure is less robust than the other measures.

(2) *The ρ values of the Outlying measure are most of the time smaller than the other measures in Scag-06.* — The *Outlying* measure in *Scag-06* is thus more sensitive to perturbations than the corresponding version in *Scag-05*. Since many of the other measures are contingent on the initial computation of the *Outlying* measure, their ρ values seem also to be higher in general in *Scag-06*.

(3) *Except the Clumpy and Outlying measures, the ρ values of all other measures decrease quite smoothly as δ increases in* Scag-05*, and even stronger in Scag-06.* — This observation indicates that large perturbations result in large changes to scagnostic measures for most scatterplots. We also computed trends of the ρ values from all the obtained measures (excluding *Outlying* and *Clumpy* for the reasons mentioned above). The resulting regression lines are shown on top of the values in Fig. 7 (dashed gray lines). They also show a clear decrease in robustness for the measures of *Scag-06*.

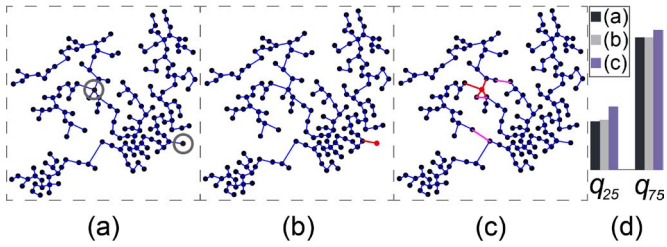

Fig. 8. Illustrating the stability of MST. (a) MST of an input scatterplot with the two points to be deleted indicated by circles; (b) deleting one boundary point and its adjacent edge (in red) does not change the MST structure; (c) deleting an interior point and its adjacent edges introduces new edges in purple, resulting in a structural change of the MST. 25th and 75th percentiles of the sorted lengths of the MST edges in (a,b,c)

The ρ values of measures obtained by rotating the scatterplots are shown in the supplementary materials, they describe similar patterns and observations. These results motivated us to further explore the underlying reasons for the larger sensitivity of the *Clumpy* and *Outlying* measures and to discuss their common limitations, aiming at providing us insights for improving the robustness of scagnostics.

4.4 Rationale for High Sensitivity of Outlying and Clumpy

Perturbing a scatterplot may change the underlying MST. This, in turn, impacts the scagnostic measures that rely on the MST. Taking the deletion operation as an example, deleting a boundary point might not change the MST structure, since any subtree of an MST is still an MST that spans all the nodes of that subtree [29]. However, deleting interior points will change the MST structure because of newly created edges. Fig. 8 illustrates such changes, where the 25th and 75th percentiles of the MST edge lengths increase by 22% and 6% after deleting the point shown in Fig. 8 (c). Because of this, the *Outlying* and *Clumpy* measures both exhibit high sensitivity to these sorts of perturbations. In the following section, we show how MST changes influence the sensitivity of *Outlying* in both versions of scagnostics.

4.4.1 Outlying Measure

To understand why the *Outlying* measure in *Scag-06* is sensitive to data perturbations, we investigated scatterplots with high sensitivity. Fig. 9 shows an example, where the *Outlying* value changes drastically from 1.31 to 0.45 after deleting the circled point in Fig. 9 (a) while the *Outlying* value in *Scag-05* remains 0 (see Fig. 9 (c)). Note that this example reveals a case in which the value of the *Outlying* measure in *Scag-06* is larger than 1. This is inconsistent with the originally stated requirement that scagnostic measures "should be on a common scale of [0,1]" [53].

Carefully looking at Fig. 9 (a,b) shows that a few MST edges (labeled in pink) are constructed during the deletion of outliers. These new edges might be even longer than most of the edges in the initial MST, resulting in a corresponding *Outlying* measure larger than 1. In contrast, only deleting nodes with degree 1 as in *Scag-05* does not introduce new edges to the MST of the remaining points. This explains why the *Outlying* measure in *Scag-05* is less sensitive than the one in *Scag-06* and why its value always fits into a common scale. Ideally, an *Outlying* measure should combine aspects of both versions: being less sensitive to perturbations while also enabling the identification of interior outliers.

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

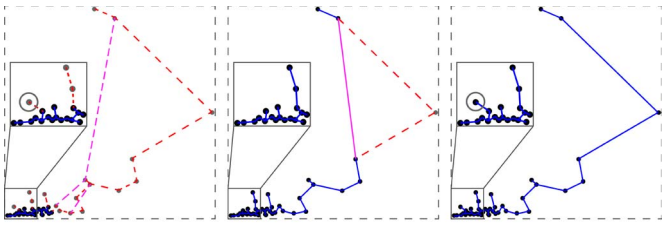

Fig. 9. An example where the *Outlying* measure defined in *Scag-06* is sensitive to perturbation, while the version in *Scag-05* does not detect any outliers. The scatterplot and its final MST are defined by solid edges, while the dotted and pink lines depict deleted and the newly inserted edges during outlier deletion. (a,c) Original scatterplot and MST; (b) Scatterplot and MST generated by deleting the circled point in (a).

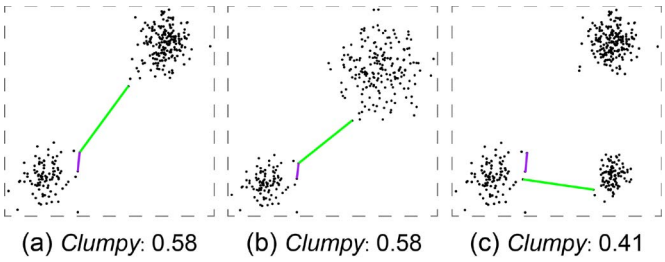

Fig. 10. Scatterplots with varying densities and numbers of clusters and the two edges in each scatterplot used for computing *Clumpy* value (shown in green and purple).

4.4.2 Clumpy Measure

Although the *Clumpy* measure has been used in several applications for identifying highly clustered areas [4, 5], only using two edges for its definition might not accurately characterize complex patterns in scatterplots, or distributions with multiple clusters [41]. In the following, we show how the *Clumpy* measure behaves for scatterplots with different cluster-like patterns as illustrated in Fig. 10.

Density Variance: Fig. 10 (a,b) shows two scatterplots that both consist of two clusters. The large clusters on the top right of these scatterplots have varying densities. However, they all have the same *Clumpy* value, indicating that the *Clumpy* measure is not able to characterize the density of the large cluster. For distributions with a single cluster, the *Clumpy* measure not only cannot represent the density information, but also has a very small value no matter how compact the cluster is (see an example in the supplementary material).

Number of Clusters: In comparison to Fig. 10 (a), the scatterplot in Fig. 10 (c) has an additional cluster on the bottom right, which is more compact than the other two clusters. Although the green edge that connects the other cluster to the additional one is quite long, the *Clumpy* value in Fig. 10 (c) is smaller than the one in Fig. 10 (a). This demonstrates that the *Clumpy* measure is not able to characterize distributions with multiple clusters well. A good *Clumpy* measure should be able to capture the characteristic of all clusters of a scatterplot. The current version is based on two edges only and is therefore limited.

4.5 Summary

To characterize the robustness of scagnostic measures, we summarize the behaviors of these measures below:

- The *Outlying* measure as defined in *Scag-06* is not robust with respect to data perturbations and special cases, while the version defined in *Scag-05* is more robust, but cannot identify interior outliers;
- The *Clumpy* measure is also not robust, and is also not able to characterize distributions with multiple clusters, or with complex variation in cluster densities;
- The other scagnostic measures except *Outlying* and *Clumpy* are less sensitive to data perturbations.

These findings motivated us to improve the definitions of the *Outlying* and *Clumpy* measures in order to increase their robustness (see Sec. 6). We discuss an additional limitation caused by collinear points in the DT in the supplementary materials.

5 User Study

Pandey et al. [35] investigated perceived similarity in scatterplots by comparing the results of user-driven groupings with the Euclidean distances of all 9 scagnostic measures. However, it is unclear how *individual* scagnostic measures align with human judgments. As we saw in the previous section, small perturbations in scatterplots can radically alter the *Outlying* measure in *Scag-06* and the *Clumpy* measure, while certain cluster-like patterns are not captured by *Clumpy* measure. Hence, to find out if small perturbations in scatterplots also alter the *perceived* features of the plot and understanding how users perceive cluster-like patterns, we designed a study with three parts. The first two parts focus on how human judgments of perceived outlyingness and clumpiness align with computed *Outlying* and *Clumpy* measures, while the last one investigates the judgments of clumpiness compared to the *Clumpy* measure specifically for scatterplots with complex, multimodal cluster patterns. Due to the space limits, we only show the comparison between *Scag-06* and human judgments; our results using *Scag-05* exhibit similar patterns and can be found in the supplementary materials.

Hypotheses: For the first two parts, our hypothesis is that human judgments of outlyingness and clumpiness would not align with the existing scagnostic measures for small perturbations (see Section 4.4), but might be consistent when perturbations are large (H1). More specifically, participants would be *insensitive* to minor perturbations of the chart when comparing measures, but would be *sensitive* to these perturbations once they were sufficiently large.

For the last part, we similarly hypothesized that human perception of clumpiness would be contingent on cluster-related features like the number and size of clusters, which would not be consistent with the existing *Clumpy* measure that cannot characterize distributions with multiple clusters (see Section 4.4.2). We therefore expected high misalignments in judgments for plots with multiple clusters (H2).

Participants: We recruited 24 participants (15 male, 9 female) from the computer science department of our local university for our study. Their ages ranged from 19 to 27 years ($M = 23$, $SD = 1.87$). All participants reported normal or corrected-to-normal vision, and had no color vision deficiencies. Subjects completed the study in one and half hours on average and were compensated with $20.00 USD. We selected this group, rather than recruiting users via a crowd-working platform, as all participants had more than 3 years experience in designing and reading scatterplots.

Apparatus: The study was conducted on a desktop machine with a 3.4GHz Intel i7-6700 CPU, 8 GB of RAM and Windows 10 operating system using a 23.8-inch LCD display with a resolution of 1920 x 1080 pixels. Participants only used the mouse to complete their tasks.

Tasks: Since our goal is to understand human sensitivity to data perturbations in terms of *Outlying* and *Clumpy*, the main experimental task was choosing which of two plots (one with, and one without perturbations) has higher *Outlying* (task I) or *Clumpy* (task II and III) values. Fig. 11 shows two instances of task I and task II; task III can be found in the supplementary material. Each participant was given 30 minutes in total to complete the entire task (with the remaining time displayed on screen), but we did not impose any per-trial time constraints. We recorded the response times, the specific plots that were

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

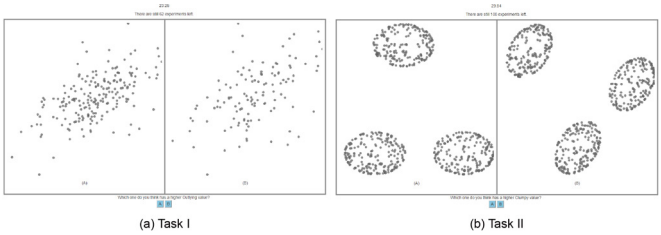

Fig. 11. Example comparisons from tasks I and II.

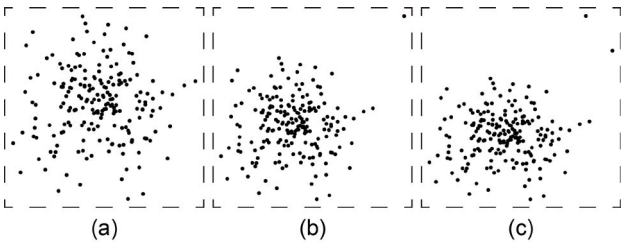

Fig. 12. Visual explanation of the concept of *Outlying* measure by three similar scatterplots in which the *Outlying* value gradually increases from (a) to (c).

chosen, and the error rate (the proportion of times that the participant did not choose the plot with the higher scagnostics value).

Procedure: After a short explanation of the task and a training session, the participants completed each of the three study tasks in order, with a short interview and a five-minute break after each task.

In order to explain the concepts of *Outlying* and *Clumpy* measures, we first gave them their word-wise definitions and then provided them with some visual explanations. Specifically, we showed them three examples of scatterplots exhibiting a gradual increase of the scagnostic measure in question. For space reasons here we show only the *Outlying* example in Fig. 12; examples for the *Clumpy* measure can be found in the supplementary material.

To further understand how humans define outliers and clumpiness, we asked participants a number of questions during the post-task interviews: "which data points are considered to be outliers?" for task I, "how do you compare the clumpiness of a pair of plots with different amounts of perturbations?" for task II, and "which data factors influence your perception of clumpiness: size, density, number of cluster, or others?" for task III. All questions were derived from a small pilot study, in which we interviewed 6 visualization experts after showing them 25 pairs perturbed/non-perturbed scatterplots and asking them to rank them in terms of outlyingness and clumpiness. The black curves in Fig. 14 show the results of the measures in *Scag-06*; red curves show our robust scagnostics measures described in Section 6.

5.1 Task I: *Outlying*

Data: We randomly selected ten plots from each of the ten distributions used by Wilkinson and Wills [53], resulting in 100 reference scatterplots in total. Based on these references we generated two perturbed plots for each of them with two different kinds of perturbation. We randomly selected half (50) of the reference scatterplots for the deletion operation. We in turn selected 5 scatterplots a piece in which we deleted $\delta \in \{1,2,3,5,7.5,10,15,20,30,45\}\%$ of the data points. Similarly, of the remaining 50 scatterplots, we selected 5 scatterplots a piece which we rotated by $\theta \in \{1,2,3,5,7.5,10,15,20,30,45\}$ degrees. Users were required to choose the plot with the higher *Outlying* value from each pair of reference plot and its corresponding perturbed plot.

Results: We analyzed our results using bootstrapped 95% confidence intervals of the sample means of the error rate. Fig. 14 (a) summarizes the results. The error rate is larger than 50% when δ is less than 5%. On the other hand, the error rate is less than 40% when δ is larger than 10%. This is consistent with Hypothesis H1, indicating that the

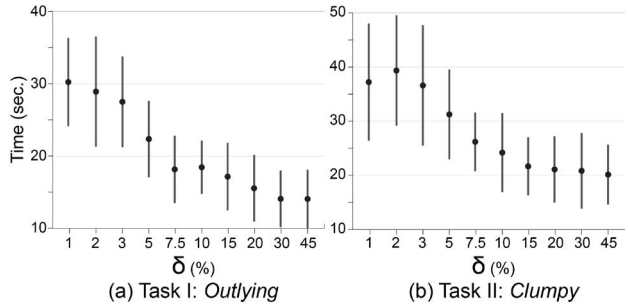

Fig. 13. Response times of the lab study with perturbation by deletion. We show mean values and deviation as 95% CIs of response times in terms of *Outlying* (a) and *Clumpy* (b).

Outlying measure, as defined in *Scag-06*, does not align well with human perception for small perturbations. In other words, human judgments are largely insensitive to small perturbations.

Fig. 13 (a) shows the response times in this study. Participants spent more time in comparing plots with small perturbations, and less time on the ones with large perturbations. This observation is partially consistent with Hypothesis H1. The results of rotation show similar implications and are given in the supplementary material.

The answers to the interview questions indicate that most participants interpreted any point far away from its nearest cluster as an outlier, no matter whether it is an interior or exterior point. This indicates the goal of the *Outlying* measure defined in *Scag-06* is justified: both interior and exterior outlying points are relevant to the visual perception of outliers.

5.2 Task II: *Clumpy*

Data: We constructed reference scatterplots with a variety of cluster-like patterns. Specifically, we chose spherical and clustered distributions [53] and eight binormal distributions with varying cluster number, size, and density (see Sec. 4.1). We created 10 scatterplots for each of these 10 distributions, resulting in 100 reference scatterplots. We followed the same procedure as in Section 5.1 to perturb each scatterplot.

Results: We present our results in Fig. 14 (b). Initially, there appears to be a similar pattern as in Fig. 14 (a), in that error rates are large when the perturbations are small (indicating that humans are insensitive to small perturbations whereas computer scagnostic measures are not). A closer look reveals that the error rates are higher and remain high even as δ increases. Our hypothesis was that the *Clumpy* measure does not fully correspond to our participants' intuitions about clumpiness. Our results therefore only partially support Hypothesis H1. Fig. 13 (b) shows the response times, which partially support Hypothesis H1.

During the interview we found out that most participants randomly made a choice when the perturbations were small. Similarly, mostly arbitrary decisions were made when perturbations (large or small) did not impact the variance in density between the two plots. The participants were most comfortable making decisions when large variations in cluster density and size were present. This suggests that defining the *Clumpy* measure across only two edges does not match visual judgments of clumpiness by humans.

5.3 Task III: *Clumpy* for Complex Clusters

Data: To explore human judgments of cluster-like patterns, we constructed scatterplots using the clusters based on the *Binormal* distribution. Based on that we introduced three kinds of scatterplot pairs for the comparisons: *same-cluster-number*, *one-more-cluster* and *random-cluster-number* aiming to understand how humans judge clumpiness in terms of cluster-specific characteristics. For *same-cluster-number* pairs, the two plots had the same number of clusters. For *one-more-cluster* pairs, one plot had one more cluster than the other. For *random-cluster-number* pairs, the number of clusters in both plots was randomly deter-

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

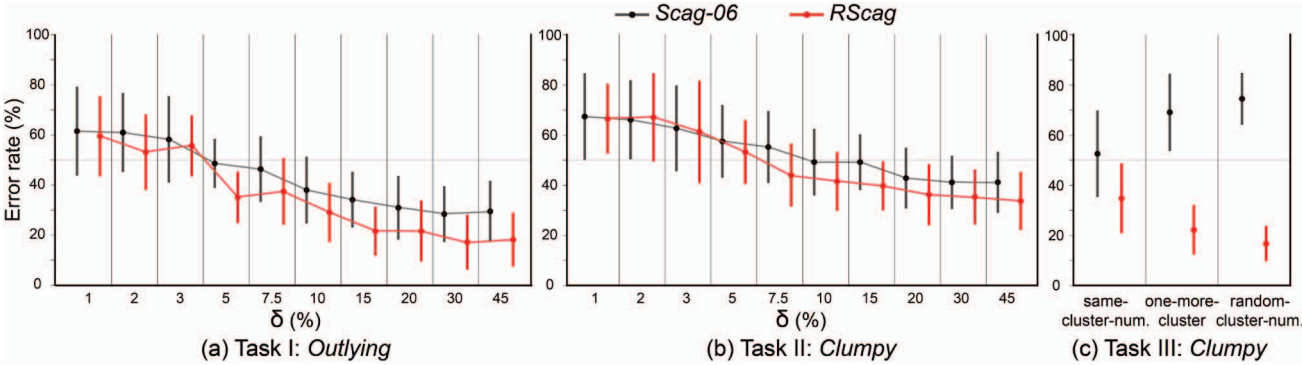

Fig. 14. Results of the lab study: Mean values and deviation for 95% CIs of error rates in terms of *Outlying* (a) and *Clumpy* (b,c) measures, which are defined in *Scag-06* are shown in black (see Section 5), while our proposed measures in red of each study (see Section 6).

mined. Since the cluster number varies from one to five, we created five different cluster configurations for pairs with the *same-cluster-number*, four different point distributions for the *one-more-cluster*, and one random configuration. For each configuration, we randomly created 10 pairs of plots with varying cluster sizes, shapes, and densities.

Results: Fig. 14 (c) shows the overall results. Error rates were very high across all three types of cluster pairs (larger than the chance of 50%), supporting our hypothesis H2: human judgments of clumpiness seem to systemically differ from the *Clumpy* measure. These rates were especially high for pairs with dissimilar cluster numbers *one-more-cluster* and *random-cluster-number*, exceeding an error rate of 70%. This suggests that humans made differing judgments based on the number and internal density of the clusters, a phenomena that is not captured in the existing *Clumpy* measure.

We specifically asked the participants about the different cluster pairs in the post-task interview. When two plots had the same number of clusters or multiple different clusters, participants mainly used the variability in density to judge clumpiness; the number of clusters had only a weak influence on their judgments. When a plot with one cluster was compared to a plot with multiple clusters, participants always assumed the plot with multiple clusters to have a higher clumpiness. Such observations motivated us to re-define the *Clumpy* measure with more cluster-specific factors rather than only using two edges.

6 Robust Scagnostics

Our results suggest that scagnostic measures might not be able to accurately and robustly characterize visual features in scatterplots, especially the *Outlying* and *Clumpy* measures. To alleviate this issue, we propose *Robust Scagnostics* (*RScag*) that compute each measure based on a cluster hierarchy rather than one or two (*Clumpy*) global cluster(s). Fig. 15 illustrates the pipeline, which also contains two preprocessing steps: adaptive binning and cluster hierarchy construction.

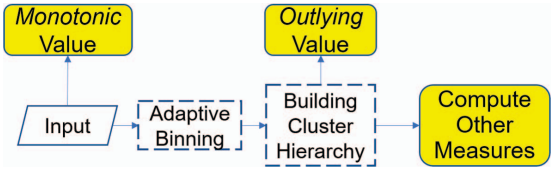

Fig. 15. The pipeline for computing *RScag*. After adaptive binning and building the cluster hierachy, all measures are computed except the *Monotonic* measure.

6.1 Adaptive Binning

Binning can significantly improve computational performance. However, it can dramatically change the characteristics of the input data and resulting measures, as shown in Fig. 6. To address this issue, we introduce *adaptive binning*, which combines hexagonal binning and uniform sampling to preserve the relative densities of an input scatterplot [11].

Given a scatterplot with n points, we first map all points to a 20×20 hexagonal grid and compute the average number of points in each cell m. Next, we define the sampling ratio as $\gamma = n/m$ and uniformly sub-sample the points in each cell with the ratio γ, while requiring that each cell contains at least one point. Fig. 16 (b,c) shows the results generated by applying the existing binning and our binning strategies. The density variability of the input scatterplot in Fig. 16 (a) is lost by hexagonal binning (subfigure (b)), but is kept by our strategy in (c).

By using our adaptive strategy, the computed scagnostic measures more closely resemble the values from the original plots. Fig. 16 (d) compares the relative change ratios (see Eq. 5) of *Outlying* and *Clumpy* values generated by our binning strategy and traditional hexagonal binning. Our strategy preserves the *Clumpy* value while reducing the change ratio of *Outlying* values from 56% to 24%.

6.2 Hierarchy-based Scagnostics

After the adaptive binning, we construct a cluster hierarchy for computing scagnostics while deleting outliers. Before describing the construction procedure for the cluster hierarchy, we first need to define robust versions of *Clumpy* and *Outlying* measures.

Robust Outlying Measure: To preserve the robustness of the *Outlying* measure in *Scag-05*, we also take all nodes with degree 1 and associated edge weight greater than ω (see Eq. 2) as outliers, but we compute ω in terms of local clusters rather than the whole MST. Suppose a cluster hierarchy is obtained from the MST, interior outliers would become exterior outliers for each local cluster. In this way, both exterior and interior outliers can be identified while preserving the robustness. Fig. 17 (a) shows an example of this process.

For a cluster hierarchy with leaf clusters $\{l_1, \cdots, l_c\}$, our *Outlying* measure is defined as:

$$c_{outlying} = \sum_i \frac{n_i}{n} \frac{length(T_o(l_i))}{length(T(l_i))}, \tag{6}$$

where n refers to the overall number of points, n_i is the number of points in the i-th sub-cluster, $T_o(l_i)$ refers to the set of outlier edges in l_i, and $T(l_i)$ is the set of MST edges in l_i.

Robust Clumpy Measure: Our robust *Clumpy* measure is based on splitting large clusters C into sub-clusters separated by the edge e_j under review, whereas the existing *Clumpy* measure ignores the density of large sub-clusters (see Figs. 10 (a,b)). To address this issue, we re-define the *Clumpy* measure by incorporating the longest edge e_m of the larger sub-cluster and the number of points in each of them:

$$c_{clumpy}(C) = 1 - \frac{length(e_k)n_k + length(e_m)n_m}{length(e_j)(n_k + n_m)}, \tag{7}$$

where n_k and n_m refer to the number of points in the small and large sub-cluster linked to e_j, respectively.

The subtrahend in Eq. 7 can be considered as the weighted Davies-Bouldin (DB) Index [23], which is defined as the ratio of the sum

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

of the within-cluster scatter to the between-cluster separation. The only difference is that the DB Index uses averaged point distances to compute the within-cluster scatter, while we use the largest point distances, see an instance in Fig. 17 (b).

For a cluster hierarchy with leaf nodes $\{c_1, \cdots, c_l\}$, we compute the overall *Clumpy* value by computing $c_{clumpy}(c_i)$ for each leaf cluster and sum them up:

$$c_{clumpy} = \sum_i \frac{n_i}{n} c_{clumpy}(c_i). \quad (8)$$

Building A Hierarchy: Taking the whole point set as a single cluster C, we perform the following procedure to build the cluster hierarchy:

1. delete outliers by computing ω (Eq. 2) for all edges of C.
2. find the edge e_j that maximizes $c_{clumpy}(C)$.
3. if e_j is smaller than ω then terminate.
4. split C into two clusters C_l and C_r.
5. repeat the whole procedure for C_l and C_r.

Due to space limits, the illustrations for this algorithm can be found in the supplementary material.

Other Measures: After finishing this procedure, we have a cluster hierarchy with leaf nodes $\{c_1, \cdots, c_l\}$ and the resulting *Outlying* value. For each of the other measures, we compute the value within each leaf cluster and then determine the weighted average of all values, as for the *Clumpy* measure defined in Eq. 8.

6.3 Comparison with Scagnostics

We validated *RScag* through conducting several comparisons with the original scagnostics. We generate the results of *RScag* by using the same perturbations as in Section 4.3, compute the error rate in the results of the user studies, and analyze the runtime of both methods for comparison. The results show that our *R-Scag* better aligns with human perception while their computation time is less than for *Scag-06*.

Perturbations: Fig. 7 shows our results at the bottom. The ρ values of our measures are larger than 0.9 when δ is not larger than 3%, and they are always larger than 0.85. And the slope of its trendline is relatively small and is similar with *Scag-05* which reflects that they change smoothly as δ increases. Our measure shows a similar sensitivity for the other perturbing operations that are shown in the supplementary material. These results confirm that our *R-Scag* is generally robust to perturbations.

User Study: Fig. 14 compares human error rate in our user study using both *Scag-06* and *RScag* as ground truths by following the same setup (participants, tasks, procedure) introduced in Sec. 5. Fig. 14 (a) shows that when α is less than 3%, our *Outlying* measure has similar error rates as that in *Scag-06*, but our error rate is more than 10% lower as δ grows larger than 3%. This indicates that our *Outlying* measure aligns better with human judgment of outliers. Fig. 14 (b) shows a similar pattern for the *Clumpy* measure, although the improvement is slightly smaller (around 7% when δ is greater than 5%). For the cluster pattern study (Fig. 14 (c)), our robust version performs much better, with improvements of 18%, 46%, and 59% across the three types of cluster pairs.

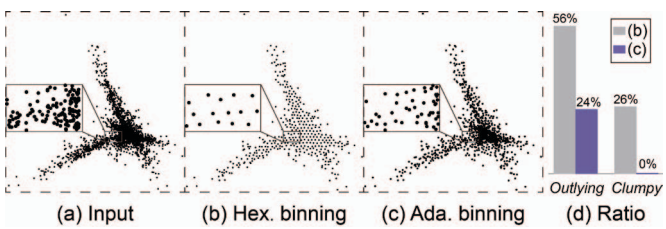

Fig. 16. Comparison of two binning methods and resulting changes for the scagnostic measures. (a) input; (b) result of hexagonal binning; (c) result of our adaptive binning; (d) relative change ratios (in percentage) of *Outlying* and *Clumpy* measures defined in *Scag-06*, obtained by comparing the scagnostic measure from (a) to the ones from (b,c).

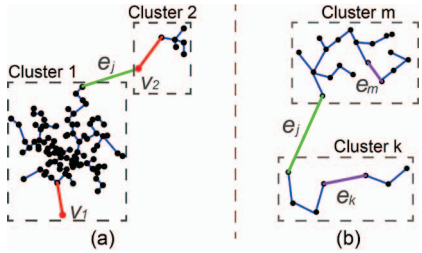

Fig. 17. Illustration of robust hierarchy-based measures: (a) *Outlying* measure; and (b) *Clumpy* measure. Both are defined on local clusters.

From the above observations, we conclude that our *Outlying* and *Clumpy* measures are more consistent with human judgments, especially when computing *Clumpy* values for complex cluster distributions.

Performance: The computational complexity of constructing the DT and MST can be reduced to $O(nlogn)$ in both cases, where n is the number points after binning. Since the complexity of the cluster hierarchy construction is $O(kn)$ ($k << \log n$), our algorithm has the same time complexity as *Scag-05* and *Scag-06*. In practice, *Scag-06* may be the slowest because of the iterative rebuilding of the DT and MST after deleting outliers. *RScag* performs the second best because of the additional hierarchy construction compared to *Scag-05*.

Moreover, we compare their runtime by applying them to scatterplots with various *Binormal* distributions on the same machine as mentioned in Sec. 5. Table 1 shows the results. While *Scag-05* is the fastest and *Scag-06* the slowest, all of them reveal similar costs. These observations are consistent with the above analysis.

Table 1. Average runtime (in ms) for three versions of scagnostics for *Binormal* scatterplots with different numbers of points.

Number of points	100	200	500	1000	1500	2000
Scag-05	23	28	35	49	57	67
Scag-06	27	33	46	69	77	89
R-Scag	23	29	39	53	62	77

7 Conclusion and Future Work

In this paper, we presented a comprehensive robustness analysis for scagnostics. We find that *Outlying* and *Clumpy* measures are negatively effected by data binning. We further conducted a user study for assessing how human judgments of outlyingness and clumpiness correlate with *Outlying* and *Clumpy* measures, which revealed that: i) human perceptions do not align with these measures for small perturbations, and ii) the perception of clumpiness mainly depends on a few cluster relevant factors such as the per-cluster density. To address these issues, we propose *Robust-Scagnostics* (*RScag*), which is robust to perturbations and more in line with human judgments.

For future work, we plan to conduct a large-scale user study to further examine the effect of different cluster relevant factors on perceived clumpiness. Second, we intend on applying our robust scagnostics to different applications [4, 19]. Lastly, we would like to extend robust scagnostics to quantify visual features in multi-class scatterplots and parallel coordinates.

8 Acknowledgements

This work is supported by the grants of the National Key Research & Development Plan of China (2016YFB1001404), NSFC (61772315, 61861136012), Science Challenge Project (TZ2016002), the Leading Talents of Guangdong Program (00201509), the CAS grant (GJHZ1862), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 251654672 – TRR 161, and the DFG Center of Excellence 2117 "Centre for the advanced Study of Collective Behaviour" (ID: 422037984).

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

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. *Computer Graphics Forum*, 38(3):225–236, 2019. doi: 10.1111/cgf.13684

[2] G. Albuquerque, M. Eisemann, D. Lehmann, H. Theisel, and M. Magnor. Improving the visual analysis of high-dimensional datasets using quality measures. In *Proc. IEEE Symp. Visual Analytics Science & Technology*, pp. 19–26, Oct 2010. doi: 10.1109/VAST.2010.5652433

[3] M. Alvo and L. Philip. *Statistical methods for ranking data*. Springer, 2014. doi: 10.1111/insr.12095_11

[4] A. Anand and J. Talbot. Automatic selection of partitioning variables for small multiple displays. *IEEE Trans. Visualization & Computer Graphics*, 22(1):669–677, 2016. doi: 10.1109/TVCG.2015.2467323

[5] A. Anand, L. Wilkinson, and T. N. Dang. Visual Pattern Discovery using Random Projections. In *Proc. IEEE Conf. Visual Analytics Science & Technology*, pp. 43–52, 2012. doi: 10.1109/VAST.2012.6400490

[6] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *Proc. IEEE Pacific Visualization Symp.*, pp. 1–8, 2016. doi: 10.1109/PACIFICVIS.2016.7465244

[7] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. In *Proc. IEEE Symp. Visual Analytics Science & Technology*, pp. 147–154. IEEE, 2008. doi: 10.1109/VAST.2008.4677368

[8] 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. Visualization & Computer Graphics*, 23(1):31–40, Jan 2017. doi: 10.1109/TVCG.2016.2598467

[9] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, et al. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446

[10] E. Bertini and G. Santucci. By chance is not enough: preserving relative density through nonuniform sampling. In *Proc. Int. Conf. Information Visualization*, pp. 622–629. IEEE, 2004. doi: 10.1109/IV.2004.1320207

[11] E. Bertini and G. Santucci. Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter. In *Proc. Int. Symp. Smart Graphics*, pp. 77–89. Springer, 2004. doi: 10.1007/978-3-540-24678-7_8

[12] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006. doi: 10.1057/palgrave.ivs.9500122

[13] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Trans. Visualization & Computer Graphics*, 17(12):2203–2212, 2011. doi: 10.1109/TVCG.2011.229

[14] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82(398):424–436, 1987. doi: 10.2307/2289444

[15] Y.-H. Chan, C. D. Correa, and K.-L. Ma. Flow-based scatterplots for sensitivity analysis. In *Proc. IEEE Symp. Visual Analytics Science & Technology*, pp. 43–50. IEEE, 2010. doi: 10.1109/VAST.2010.5652460

[16] Y.-H. Chan, C. D. Correa, and K.-L. Ma. The generalized sensitivity scatterplot. *IEEE Trans. Visualization & Computer Graphics*, 19(10):1768–1781, 2013. doi: 10.1109/TVCG.2013.20

[17] C. D. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *Proc. IEEE Symp. Visual Analytics Science & Technology*, pp. 51–58. IEEE, 2009. doi: 10.1109/VAST.2009.5332611

[18] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE Trans. Visualization & Computer Graphics*, 25(1):830–839, 2019. doi: 10.1109/TVCG.2018.2864907

[19] T. N. Dang, A. Anand, and L. Wilkinson. Timeseer: Scagnostics for high-dimensional time series. *IEEE Trans. Visualization & Computer Graphics*, 19(3):470–483, 2013. doi: 10.1109/TVCG.2012.128

[20] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *Proc. IEEE Pacific Visualization Symp.*, pp. 73–80. IEEE, 2014. doi: 10.1109/PacificVis.2014.42

[21] T. N. Dang and L. Wilkinson. Transforming scagnostics to reveal hidden features. *IEEE Trans. Visualization & Computer Graphics*, 20(12):1624–1632, 2014. doi: 10.1109/TVCG.2014.2346572

[22] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Trans. Visualization & Computer Graphics*, 16(6):1017–1026, 2010. doi: 10.1109/TVCG.2010.184

[23] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Analysis & Machine Intelligence*, (2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909

[24] A. Frank and A. Asuncion. University of California Irvine (UCI) Machine Learning Repository, 2010.

[25] R. Hafen, L. Gosink, J. McDermott, K. Rodland, K. Kleese-Van Dam, and W. S. Cleveland. Trelliscope: A system for detailed visualization in the deep analysis of large complex data. In *Proc. IEEE Symp. Large-Scale Data Analysis & Visualization*, pp. 105–112. IEEE, 2013. doi: 10.1109/LDAV.2013.6675164

[26] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using Weber's law. *IEEE Trans. Visualization & Computer Graphics*, 20(12):1943–1952, 2014. doi: 10.1109/TVCG.2014.2346979

[27] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989. doi: 10.1080/01621459.1989.10478797

[28] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Trans. Visualization & Computer Graphics*, 22(1):469–478, 2016. doi: 10.1109/TVCG.2015.2467671

[29] D. C. Kozen. *The design and analysis of algorithms*. Springer Science & Business Media, 2012. doi: 10.1007/978-1-4612-4400-4

[30] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, and H. Theisel. Selecting coherent and relevant plots in large scatterplot matrices. *Computer Graphics Forum*, 31(6):1895–1908, Sept. 2012. doi: 10.1111/j.1467-8659.2012.03069.x

[31] D. J. Lehmann, S. Hundt, and H. Theisel. A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *it - Information Technology*, 57(1):11–21, 2015. doi: 10.1515/itit-2014-1070

[32] D. J. Lehmann, F. Kemmler, T. Zhyhalava, M. Kirschke, and H. Theisel. Visualnostics: Visual guidance pictograms for analyzing projections of high-dimensional data. *Computer Graphics Forum*, 34(3):291–300, 2015. doi: 10.1111/cgf.12641

[33] J. M. Lewis, M. Ackerman, and V. de Sa. Human cluster evaluation and formal quality measures: A comparative study. In *Proc. Annual Meeting of the Cognitive Science Society (CogSci)*, pp. 1870–1875, 2012.

[34] J. Matute, A. C. Telea, and L. Linsen. Skeleton-based scagnostics. *IEEE Trans. Visualization & Computer Graphics*, 24(1):542–552, 2018. doi: 10.1109/TVCG.2017.2744339

[35] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)*, pp. 3659–3669. ACM, 2016. doi: 10.1145/2858036.2858155

[36] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models*. Wiley Online Library, 2004. doi: 10.1002/0470870958

[37] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015. doi: 10.1111/cgf.12632

[38] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Visualization & Computer Graphics*, 19(12):2634–2643, 2013. doi: 10.1109/TVCG.2013.153

[39] M. Sedlmair, A. Tatu, T. M., and T. Munzner. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3):1335–1344, 2012. doi: 10.1111/j.1467-8659.2012.03125.x

[40] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005. doi: 10.1057/palgrave.ivs.9500091

[41] L. Shao, A. Mahajan, T. Schreck, and D. J. Lehmann. Interactive regression lens for exploring scatter plots. *Computer Graphics Forum*, 36(3):157–166, 2017. doi: 10.1111/cgf.13176

[42] L. Shao, T. Schleicher, M. Behrisch, T. Schreck, I. Sipiran, and D. A. Keim. Guiding the exploration of scatter plot data using motif-based interest measures. In *2015 Big Data Visual Analytics (BDVA)*, pp. 1–8, Sep. 2015. doi: 10.1109/BDVA.2015.7314294

[43] B. W. Silverman. *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986. doi: 10.1201/9781315140919

[44] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009. doi: 10.1111/j.1467-8659.2009.01467.x

[45] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnork, and D. Keim. Combining automated analysis and visualiza-

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.

tion techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science & Technology*, pp. 59–66, 2009. doi: 10.1109/vast.2009.5332628

[46] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: An initial study on 2D projections of large multidimensional data. In *Proc. Int. Conf. Advanced Visual Interfaces (AVI)*, pp. 49–56, 2010. doi: 10.1145/1842993.1843002

[47] J. W. Tukey. Mathematics and the picturing of data. In *Proc. Int. Congress of Mathematicians*, vol. 2, pp. 523–531, 1975.

[48] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[49] J. W. Tukey and P. A. Tukey. Computer graphics and exploratory data analysis: An introduction. In *Proc. Sixth Annual Conf. and Expo.: Computer Graphics*. National Computer Graphics Association, 1985.

[50] L. Wilkinson and A. Anand. Graph-theoretic scagnostics. *Proc. IEEE Information Visualization Symp.*, pp. 157–164, 2005. doi: 10.1109/INFVIS.2005.1532142

[51] L. Wilkinson and A. Anand. *scagnostics: Compute scagnostics - scatterplot diagnostics*, 2018. R package version 0.2-4.1.

[52] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: interactive exploration guided by pairwise views of point distributions. *IEEE Trans. Visualization & Computer Graphics*, 12(6):1363–72, 2006. doi: 10.1109/TVCG.2006.94

[53] L. Wilkinson and G. Wills. Scagnostics Distribution. *Journal of Computational & Graphical Statistics*, 17(2):473–491, 2008. doi: 10.1198/106186008X320465

Authorized licensed use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on October 26,2020 at 12:06:32 UTC from IEEE Xplore. Restrictions apply.