Info Vis

Article

Information Visualization
2018, Vol. 17(1) 22–40
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1473871616686635
journals.sagepub.com/home/ivi
SAGE

Z-Glyph: Visualizing outliers in multivariate data

Nan Cao[1,*], Yu-Ru Lin[2,*], David Gotz[3,*] and Fan Du[4,*]

Abstract

Outlier analysis techniques are extensively used in many domains such as intrusion detection. Today, even with the most advanced statistical learning techniques, human judgment still plays an important role in outlier analysis tasks due to the difficulty of defining and collecting outlier examples. This work seeks to tackle this problem by introducing a new visualization design, "Z-Glyph," a family of glyphs designed to facilitate human judgment in outlier analysis of multivariate data. By employing a location-scale transformation, a Z-Glyph represents the "normal" data using regular shapes (e.g. straight line and circle), such that the abnormal data can be revealed when deviating from the regular shapes. Extensive controlled experiment and case studies based on real-world datasets indicate the superior performance of the Z-Glyph family, compared with the baselines, suggesting that the proposed design is able to leverage human perceptional features with statistical characterization. This study contributes to a more fundamental understanding about designing visual representations for revealing outliers in multivariate data, which can be applied as a building block in many domain-specific anomaly detection applications.

Keywords

Outlier detection, anomaly detection, information visualization, multidimensional data visualization

Introduction

Outliers, also referred as *anomalies*, are patterns in data that do not conform to expected behavior.[1] Outlier and anomaly detection techniques have been extensively used in a wide range of applications such as fraud detection in financial transactions, or intrusion detection in cyber-security systems. Methods for detecting outliers in data have been proposed since 19th Century,[2] and more analysis techniques have been studied extensively in the literature.[1,3] Particularly, a large category of existing techniques is developed for identifying point outliers in the multivariate data (i.e. data items are shown as points in the multidimensional feature space). However, outlier detection is still considered as a highly challenging problem due to factors such as the availability of labeled data. In this work, we seek to tackle this problem by introducing a new visualization design, called "Z-Glyph," for point outlier analysis of multivariate data.

There are two major challenges in outlier detection. First, defining "normal" (and "anomalous") behavior in data is difficult due to the nature of the data (factors including various data distributions, amount of noise, unknown data-generating process and potential dynamics in data, and so on). Second, labeled data with a high quality for training and validating models used by anomaly detection techniques are often unavailable or difficult to obtain. Hence, in order to better distinguish actual anomalies and collect sufficient

[1]Tongji University, Shanghai, China
[2]University of Pittsburgh, Pittsburgh, PA, USA
[3]The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[4]University of Maryland, College Park, MD, USA
[*]All the authors contributed equally to this article.

Corresponding author:
Nan Cao, Tongji University, Shanghai, China.
Email: nan.cao@gmail.com; yurulin@pitt.edu; gotz@unc.edu; fan@cs.umd.edu

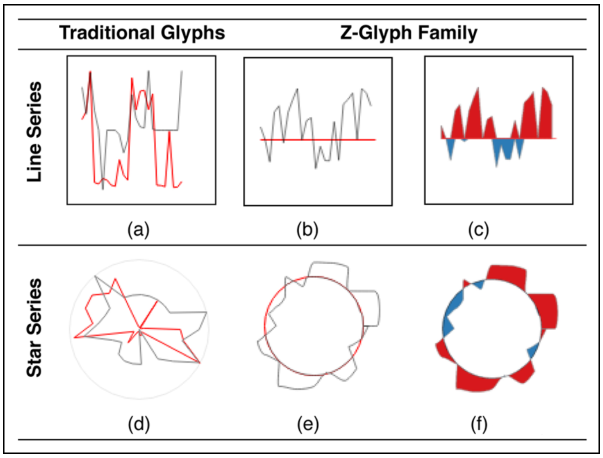


Figure 1. Traditional glyphs and Z-Glyph family for representing the same multivariate data: (a) Line Glyph, (b) Z-Line Glyph, (c) Z-LineD Glyph, (d) Star Glyph, (e) Z-Star Glyph, and (f) Z-StarD Glyph. In traditional glyphs (a, d), baseline values are shown in red. In Z-glyph family (b, c, e, f), data values are transformed and positioned with respect to the "baseline" values shown in regular shapes (such as a straight line and a circle). Dichotomous color encoding is further used to highlight trends deviated from baseline values (c, f).

representatives, human judgement continues to play a critical role in the process of outlier analysis, even with the most advanced statistical learning techniques.[3]

There have been domain-specific visualization techniques designed to facilitate outlier detection in more complex datasets or scenarios, such as visualizing outliers in network traffic data,[4–7] and monitoring anomalies in social media.[8,9] However, there is a very limited understanding about how to generalize these visualization design approaches to reveal outliers in generic multivariate data. In this article, we introduce Z-Glyph, a family of glyphs designed specifically to support outlier detection in multivariate data. Fig. 1 (b,c,e,f) showcases four types of Z-Glyphs proposed and evaluated in this article, extending a preliminary Z-Star design first introduced in Cao et al.[8] This article is motivated by seeing the potential usefulness of this preliminary design as well as the missing of formal evaluations in the original article. The Z-Glyph family is developed based on a common "core idea" that representing "normal" data using regular shapes (e.g., straight lines or circles). This design allows glyphs that depict abnormal data as easily-detected shape deviations. This design follows the one-class assumption that is used in many anomaly detection algorithms.[10,11] It assumes that most data items belong to one large normal category (summarized as the baseline) and only few of them are outliers (revealed by shape deviations). This design not only visually differentiate the abnormal items from the normal ones but also enables a more precise data labeling procedure guided by analyzers through reading and interpreting the intuitive visual representation. Our study results verified the effectiveness of the Z-Glyph design and also revealed that highlighting value differences by colors (Figure 1(c) and (f)) is not very helpful for identifying outliers as expected.

In particular, the main contribution of this article includes:

- *Extending the existing design.* We propose the Z-Glyph family by extending the Z-Star Glyph which is first introduced in Cao et al.[8] based on the same design scheme. Several new glyphs were proposed as the alternative designs and are compared to the Z-Star glyph. These designs leverage human perception features, visual metaphor, and statistical characterization.
- *Extensive controlled experiment.* We propose a new set of experiments to systematically evaluate multiple aspects of different Z-Glyph designs in context of outlier detection. We performed these experiments in a controlled user study to understand the strength and limitations of different Z-Glyphs compared with two baselines designs including Line and Star glyphs. The results not only indicate the proposed design outperforms the baseline glyphs overall, but also reveal design features that are suitable for outlier analysis tasks.
- *Case studies on real datasets.* We developed outlier detection system by applying Z-Glyph design using two real world datasets where ground-truth information is available. We conducted system test and in-depth interview with two expert users using the prototype system. Their feedback showcases the effectiveness of the Z-Glyph design and the feasibility of tackling real-world outlier analysis tasks.

Related work

In this section, we discuss the related work from three aspects: (1) outlier detection with the use of visual analysis techniques, (2) glyph-based visualization, and (3) similar visual designs.

Outlier detection

Outlier analysis techniques, including supervised, unsupervised, and semi-supervised methods, have been studied extensively in the literature.[1,3,12] Typically, the outputs of an outlier or anomaly detection technique are either numeric scores or labels (normal or anomalous).[1] As human judgement is critical in the process of outlier analysis, how to design better representations to enable more effective human judgement and interpretation about outliers in data becomes an important issue.

Visualization techniques have been applied to assist in anomaly detection and evaluation. Statistical diagrams, such as line charts (in particular, time series charts) and histograms, are most commonly used to represent the anomalous changes in variables.[13–15] For spatial data, variogram clouds and pocket plots have been used in finding abrupt changes that violate spatial auto-correlations.[3,16] When dealing with spatial time series data, it is common to find unusual shapes from multiple spatial distributions, such as color distributions in MRI scans.[3]

For multidimensional or multivariate data, various types of dimension reduction techniques, such as multidimensional scaling (MDS)[17] and principal component analysis (PCA),[18] can be applied to create visual mapping in a lower dimensional space. Scatterplot matrices and parallel coordinates[19] are often used to represent data values across multiple dimensions. Although not particularly design for outlier analysis, by depicting the overall pattern of the data, these visualizations can also review outliers to some extent.[20–22] There have been visualization techniques designed for outlier detection in specific domains such as intrusion detection in the field of network security.[4–7] However, these special visualizations are usually not suitable for broader applications.

Several visualization techniques have been proposed to facilitate outlier detection in more complex datasets or scenarios, such as detecting abnormal behaviors in social media. For example, Thom et al.[23] introduced a visual analysis system for monitoring anomalous bursting of keywords at different times and locations based on a tag cloud visualization overlaid on top of a map. Zhao et al.[9] developed the FluxFlow system for detecting and visualizing anomalous information propagation processes in Twitter. Cao et al.[8] introduced TargetVue, a visual analysis system for detecting anomalous user behaviors in online communication systems. These studies showcase comprehensive visual analysis systems that leverage data mining and interaction techniques for outlier detection in a specific application context. Compared with these specific designs, our work focuses on designing a general visual representation for discovering outliers in multidimensional datasets. Our design can be applied to broader application contexts or used in existing visual analysis systems, making the development of domain-specific anomaly detection systems more efficient.

Glyph-based visualization

In information visualization, a glyph refers to a small and compact graphic representation that represents a data point with multidimensional features.[24] Compared with other multidimensional visualization techniques, such as multidimensional scaling (MDS)[17], parallel coordinates[19], scatterplot matrices, and various advanced designs for reducing clutter in multidimensional data[25] or for representing data from heterogeneous dimensions[26–30], glyphs transform multidimensional data features to composite visual properties (such as shape, color, and size), producing various "visual signatures" of data points that reveal more complex data patterns and offer a richer description about data points. The composite visual form of a glyph also allows it to be used in small-multiple settings, or to be flexibly combined with other types of data representation or graphics such as tables and maps.[31]

Glyph based designs have been proven to be effective for representing rich data in a wide range of domains. Examples include visualizing poetry,[32] sport event,[33] medical data,[34–36] time series data,[37] workflow data,[38] vector fields,[39,40] or representing data uncertainty[41] or sensitivity[42] and comparing subject survey data.[43] A glyph's composite visual form makes it suitable to be used in distinguishing some sort of "activities" in a dynamic environment. For example, Erbacher et al.[44] introduced a radial glyph that shows a web server's activities for connecting to other servers over time. Fry[45] introduced a glyph that summarizes and represents users' visits to web pages at a time and allows comparing changes across time. Xiong and Donath[46] developed "PeopleGarden," a flower-shaped glyph that summarizes a user's aggregated interaction histories in a discussion group. These existing glyph designs can be useful in revealing outlier activities in a particular setting; however, there is still a lack of understandings about how to design generic glyphs for supporting outlier analysis.

Similar visual designs

Comparing different items in a dataset is a key step for detecting outliers. Therefore, an effective representation of multivariate data for outlier detection should facilitate a fast visual comparison of data features. Gleicher et al.[47] comprehensively summarized various different types of visual comparison techniques in their survey paper. Following their taxonomy, the design of Z-Glyph falls into the category of "signal subtraction." While the proposed Z-Line design may appear at first glance similar to the one shown in Fig.1(c) in Gleicher et al.,[47], which illustrates the comparison of the value differences of two variables X, and Y by showing "X-Y", our design targets on identifying outliers from a set of multivariate data items, thus making a distinct contribution. In particular, we show differences between the feature values of an item using the baseline values across multiple variables. Here, the goal is not to compare two variables but to compare multiple data items. In addition, to the best of our knowledge, little visual comparison technique has been developed to detect outliers in multivariate data.

Another similar design is the horizon graph,[48,49] a variant of the line chart, which is originally designed to help illustrate multiple time-series within a compact display area. In this design, the line chart is divided into layered bands by multiple baselines, each of which indicates a data value. Different from horizon graph in which each baseline indicates a single value, the baseline in Z-glyph indicates the mean values of multiple different data features. It distorts and visualizes different mean values onto the same line segment, thus facilitating a fast comparison between normal and abnormal values across multiple data dimensions, which cannot be achieved by a horizon graph. Therefore, Z-glyphs are essentially generalizations of horizon graphs where the baseline value (regardless of the actual value) are aligned on the same horizontal line.

Design of Z-Glyphs

In this section, we introduce the visualization design of the proposed Z-Glyph visualization.

Visual design and rationales

The proposed glyphs aim to facilitate human judgment in the process of outlier analysis. A critical question to be answered here is *how to represent outlier information that can be easily perceived and recognized by human*. Our design is motivated by the following design guidelines and data analysis strategies:

Choosing optimal visual channels. A variety of visual attributes, such as shape, color, size, orientation, closure, can be incorporated into designing a glyph for outlier detection purpose. The proposed glyphs should be designed based on visual channels that are mostly effective for encoding outlier information. In this study, we investigate several visual channels that have been shown effective in glyph-based visualizations and further test their effectiveness in the context of outlier analysis.

Utilizing visual metaphor. Metaphoric visual representation is a powerful way to establish association between a visual channel and the concept(s) to be encoded.[38] If possible, visual metaphor should be employed to facilitate establishing an intuitive mental model for perceiving outliers. The proposed glyphs are designed based on the metaphor of "compliance versus non-compliance" where normal data patterns are represented as a regular shape (a straight line or a circle) and outlier patterns are displayed as shapes departed from regular shapes.

Incorporating statistical distribution concept. Outlier detection methods commonly rely on determining the statistical estimation of the underlying distribution to characterize the normal behavior of the data. This common analysis strategy should be incorporated when designing the visual encoding of outlier information. The proposed glyphs leverage the concept of distributions widely studied in the statistics literature. However, unlike traditional outlier detection methods that simply output scores or labels to represent the "outlierness," our design visually encodes the statistical information to better support human recognition and interpretation.

Visual encoding

Typically, data with multidimensional feature values can be represented using Line glyphs or Star glyphs (Figure 2(a) and (c)). In a Line glyph, feature axes are parallel arranged. A data item is shown as a polyline connecting with the points indicating the data item's feature values (e.g., the black polyline shown in Fig. 2(a)). In a Star glyph, a data item is shown with feature axes arranged radially (Figure 2(c)). A Naïve way to introduce the outlier-related information would be to overlay the "normal" feature values on the same glyph,

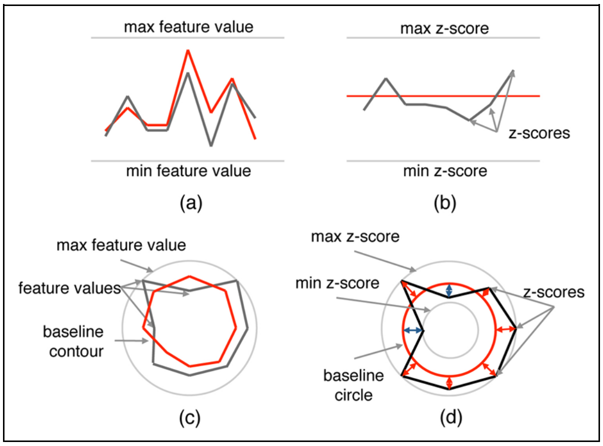


Figure 2. Visual design: (a) traditional Line glyph plots data and baseline values in a re-scaled space limited by the min. and max. feature values, (b) Z-Line glyph plots data with location-scale transformation (z-scores), where the location parameter values are viewed as the baseline, (c) traditional Star glyph plots data and baseline values in a re-scaled circle limited by the max. feature values, and (d) Z-Star glyph plots data with location-scale transformation in a scaled circular band.

such as the red polylines shown in Figure 2(a) and (c). Such representation, however, does not directly guide users/viewers to judge or recognize outliers.

We propose a new glyph design for encoding outlier information. First, we represent the "normal" data using regular shapes including straight line and circle, such that abnormal data can be revealed if their feature representation deviates from the regular shapes (Figure 2(b) and (d)). Second, to enable the visual comparison between shapes, a data item's feature values should have common scales across dimensions, such that certain types of shapes (e.g. smoothing or fluctuated lines) can be interpreted in a similar way regardless of the original feature units. To create such a feature representation, we employ a location-scale transformation for each feature dimension as follows.

Let X be a feature variable, the transformed feature variable is defined as $Z = (X - a)/b$, where a is the *location* parameter and b is the *scale* parameter. The location parameter can be chosen to measure the central tendency of the distribution, such as mean, median, and mode. The scale parameter should measure the dispersion or variation of the variable X. When a is the mean of X, and b is the standard deviation of X, the transformation corresponds to standardization. Z is called standard score or z-score. The standard score measures the distance from the mean to the random variable in terms of standard deviations, and hence, it is dimensionless (i.e. it has no physical units). This standard transformation can be applied to arbitrary distributions.

To simplify the interpretation of visual mapping resulted from the transformation, we assume the underlying feature values follow or can be transformed to follow a certain location-scale distribution such as normal distribution and exponential distribution. In this way, the standard scores remain unchanged in the location-scale transformation, making the visual perception of similar visual mappings consistent. If the feature values follow a normal distribution, the outliers can be easily recognized with high absolute z-scores. Even for arbitrary distributions, the transformation still provides heuristics to compare the relative "outlierness" of data and hence has been commonly used in outlier analysis.[3] In our glyph design, encoding this outlier information as shapes in a glyph allows users to visually compare and recognize potential outliers in the data, which leverage human judgment in better distinguishing actual anomalies.

We propose Z-Glyph family following the idea of visually encoding the feature z-scores. Based on different visual encoding strategies, the Z-Glyph family has four variants: Z-Line, Z-Star, Z-LineD, and Z-StarD (as shown in Figure 1(b), (c), (e), and (f)). In Z-Line and Z-Star glyphs, the features' z-scores are plotted as polylines or stars against the mean, shown as the red straight line in Figure 1(b), or the red circle in Figure 1(e), respectively. The mean line/circle forms a stable visual *baseline* in the entire dataset which simplifies the visual detection (sometimes, the mean value can be replaced by the baseline values of the features which are already known. For example, the standard lab test results in an electronic health records). The two design variants utilize different combinations of visual channels for comparison. In Z-LineD and Z-StarD glyphs, the areas between the feature polylines/stars and the mean line/circle are filled with two colors to enhance the dichotomous region—values above the means are colored in red and values below the mean are colored in blue. The dichotomous coloring incorporates an additional visual channel to assist visual comparison across shapes. Figure 3 illustrates the normal and abnormal patterns shown using Z-StarD, where colored area emphasizes the deviance of feature values. In this study, we will examine these different design choices and their effectiveness in supporting outlier analysis.

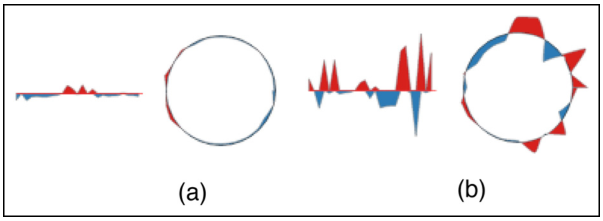


Figure 3. Visualizing (a) normal and (b) abnormal data values.

Discussion

Assumption on data distribution. It is worth mentioning that the aforementioned Z-Glyph design based on the assumption of the underlying data following an unimodel-based distribution. The underlying rationales for making this assumption are from multiple aspects: (1) many nonparametric methods in outlier detection, for example, those that are designed to search for low-density objects in Euclidean space, are using the same assumption and are verified to be effective in practice;[1,3] (2) even for arbitrary distributions, this assumption still provides good heuristics that allows for comparing the relative "outlierness" of data and hence has been commonly used in outlier analysis;[3] and (3) the proposed visualization follows three design rationales with the goal to better support human recognition and interpretation. Note that we do not assume the data should follow a normal distribution, but we do assume they should follow the location-scale distribution, which is a broader family containing normal distribution. Our framework allows users to choose measures for the central tendency of a distribution, for example, mean, median, and mode (page 3). We believe that this design contributes to provide a novel linkage to bridge external representation (visualization) and the statistical distribution concept (users' conceptual model related to outlier recognition).

Readability of the design. Another potential constraint of Z-Glyph design is that scaling data around a baseline transforms the data into a relative instead of an absolute scale, which makes it difficult to read actual values from the visualization. We believe in most of the cases that Z-Glyphs will be used for providing visual cues of outliers in a multidimensional dataset. Therefore, supporting a precise reading of the feature values is not the major goal of the Z-Glyph design as other visualization views that facilitate data reading can always be used at the same time as shown in Cao et al.[8]

Experiment design

We examine the effectiveness of different glyph design choices in a controlled user study. In this section, we describe the design of the experiment and provide rationales for some of the particular experiment design decisions, which were made based on prior studies and our pilot studies.

User task: outlier detection

This study focuses on evaluating the glyphs' capability of revealing outliers in a multivariate dataset. To this end, we design a task that simulates a typical outlier detection task in the process of outlier analysis, in which a large collection of data items are considered normal but a small portion of items are potentially abnormal and requires additional human inspection. Human evaluators need to be able to find actual outliers from this small set of potentially abnormal items. Hence, in our experiment, the user task is as follows:

- Determine outlier items (i.e. the items have significant different feature values compared with that of other items) from a given small set of multidimensional data based on their glyph representation.

In this task, the primary factor to be tested is the six design choices, as shown in Figure 1(a)–(d). Additionally, when these glyphs are used for representing data in the outlier detection task, the results are affected by two major factors: (a) the numbers of data items shown to the users and (b) the numbers of features represented by the glyphs. We have conducted a pilot study with six users to determine the proper conditions for examining how these two factors affect the study results.

In real-world applications, identifying actual outliers is not a trivial task and usually requires evaluators to inspect data with dozens or even hundreds of feature dimensions.[8] In order to simulate the real-world scenario, we decided to show data with few dozens of feature dimensions through glyphs. We tested a wide range of possible number of feature dimensions in our pilot study and selected 25 as the low-dimensional case and 50 as the high-dimensional case as the two conditions best differentiated users' detection ability. We believe 50 dimensions are also high enough to verify the Z-Glyph family's scalability in terms of representing high dimension data. In comparison, most existing techniques, as shown in a recent survey[50], are able to concurrently visualize only a relatively small number of dimensions (most often less than 20). We also tested a range of possible numbers of data items to be shown to the users during the formal experiment. We determined to use 5×5 = 25 items as the small-size dataset case and use 10×10 = 100 items as the large-size dataset case.

Study hypotheses

The goal of this experimental study is to understand the strengths and limitations of different glyph designs in terms of their effectiveness of facilitating human judgment in outlier analysis. Based on the design rationale provided in the last section, we hypothesize the core design of the Z-Glyph family (i.e., showing the data means as a stable visual baseline) will better facilitate the outlier recognition than the two Naïve baseline designs, i.e., the Line glyph and the Star glyph.

H1: The Z-Glyph family is more effective than the baseline glyphs (Line and Star) in assisting outlier detection task.

These design variants utilize different visual channels. Considering line-based glyphs simply require visual comparison of positions along vertical direction, and human visual system is most efficient in position comparison, we hypothesize that line-based glyphs better facilitate the outlier recognition than star-based glyphs (which also requires visual comparison in orientation).

H2: The line-based glyphs (Line, Z-Line, and Z-LineD) are more effective than the star-based glyphs

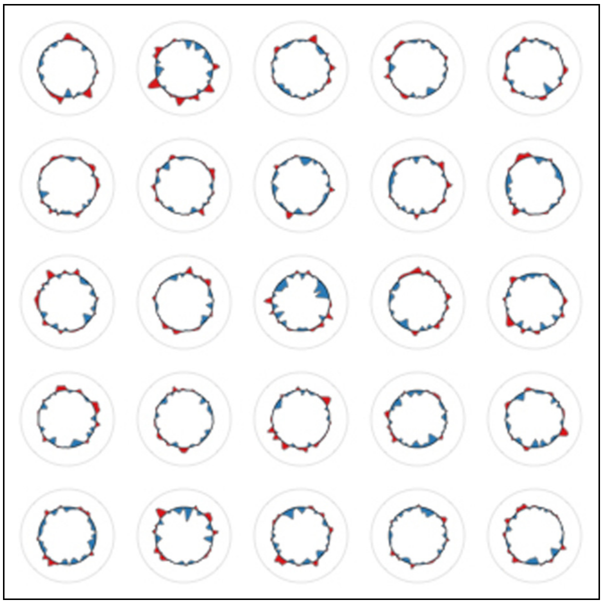


Figure 4. In the study, glyphs are randomly laid out in an *N* by *N* grid.

(Star, Z-Star, and Z-StarD) in assisting outlier detection task.

Furthermore, we hypothesize that adding dichotomous color encoding helps with outlier recognition as the dichotomous colored region highlight the deviation of feature values.

H3: The dichotomous color–encoded glyphs, Z-LineD and Z-StarD, are more effective than Z-Line and Z-Star in assisting outlier detection task.

Glyph display

We would like to minimize the influence of other visual properties irrelevant to the glyph design, such as the positioning of the glyphs, in the study. To this end, we randomly position data glyphs in an N by N grid, where the glyphs' positions do not encode any information (Figure 4).

Task performance measures and test data

To evaluate users' performance of detecting outliers via different glyph designs, we quantify the accuracy and the completion time of performing the task. There are two alternative ways to measure the task accuracy:[52] "probe one" in which users need to identify a single item with the highest "outlierness" and "select all" in which users need to identify all outlier items in a given dataset. In our pilot study, we have tested the two experiment designs. We found that the "probe one" is not proper in this study as there was no clear way of judging what "the most" abnormal pattern might be. Thus, instead of "probe one," we asked users to select three outlier items without explicitly ranking the most abnormal one. The number of outliers was chosen because even with many state-of-the-art anomaly detection techniques (e.g. One-Class SVM[10] and OCCRF[11]), the accuracy may be less than 10% in real-world applications,[9] that is, about 3 out of 25 data items. In addition, we have chosen to fix this number regardless of dataset sizes. Fixing target numbers enable a comparison of users' task completion time in all cases, as selecting more targets require more operations (e.g. mouse clicks) that could confound the study results.

The task completion time was automatically recorded in our experimental system. It measures the duration starting at the time when each testing dataset is loaded and presented to users as glyphs, and ending at the time when users click the "next" button to continue the next trial. The duration includes both the data inspection time and answering time.

Simulated data. In the experiment, we assumed that the underlying multivariate data were normal deviate, and users were asked to find three actual outliers from each of the given datasets. We generated each of the testing dataset that contained N data items with D-dimensional features as follows. We first produced sufficient amount of samples following the D-dimensional multivariate normal distribution and computed the sample mean μ and sample standard deviation σ. We randomly selected three sample points whose distances to the mean were greater than 3σ, and randomly selected $N - 3$ points with the distances to the mean less than 3σ.

Consideration of study baselines

We consider line glyph and star glyph as two design baselines (Figure 1(a) and (d)) as they are the most popular glyph design choices.[37] In terms of star glyph, there exist several design variants that could influence the study results. It has been shown in previous study[31] that a star glyph with data lines outperforms those star glyphs attached with contours in terms of revealing data similarities. However, the prior study results cannot be directly applied in our study for two key reasons. First, previous study only considered data with relative small dimensions (not more than 10), and our study considers much larger feature dimensions. Second, previous study focused on evaluating the design choices for a task of revealing similar patterns

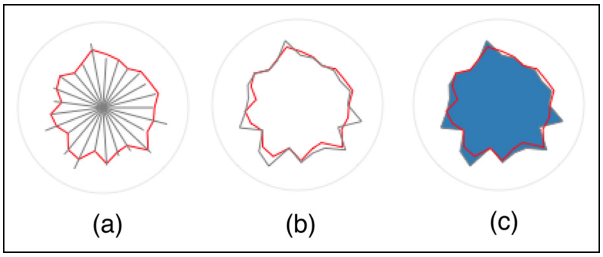


Figure 5. Different variations of star glyphs: (a) data line only, (b) data contour only, and (c) data contour with a filling color. In all these variations, the baseline is visualized as a red contour.

with respect to one target item, and our goal is to evaluate the designs in terms of how they help reveal a small portion of abnormal patterns. Thus, we conducted an additional pilot study to determine a specific star glyph design as the baseline in our experiment.

In the pilot study, we compared three types of star glyph designs shown in Figure 5. Eighteen users were asked to select three outliers out of 100 50-dimensional data items. The results, as summarized in Figure 6, suggested that the design (b) performs the best, both in terms of low completion time and high accuracy. In particular, accuracy of design (b) is significantly better ($p < .05$ when compared to design (a) and $p < .05$ when compared to design (c)). Therefore, we have chosen design (b) as the baseline in our main experiment.

The order of glyph axes is another relevant design factor that is also investigated in the pilot study. However, we decide to omit this factor from our final study and paper for the following two reasons: (1) the pilot study results suggested that reordering the axes in glyphs did not have a significant effect on the performance measures when using Z-Glyphs; (2) Z-glyphs can be extended to visualize time-series data in which the order of axis represents timestamps which cannot be reordered; and (3) reordering is a visual clutter reduction technique which can be applied in Z-Glyph, but reordering itself is not related to the design of Z-Glyph.

User study

In this section, we first describe the study procedures that were followed to realize the above experimental design. We then present the study's results and discuss the findings.

Participants and apparatus

We recruited 18 users (8 females) to participate in our study with the goal of comparing six distinct glyph designs: Line Glyph, Z-Line Glyph, Z-LineD Glyph, Star Glyph, Z-Star Glyph, and Z-StarD Glyph, as shown in Figure 1. The users were researchers or graduate students in computer science, art, and psychology. Their ages ranged from 23 to 34 (mean: 28, SD = 3.16) and all had normal vision.

Testing environment. The study was performed on a 15.4-in laptop computer with a display resolution of 1440×900 pixels and a 60 Hz refresh rate. Users sat approximately 50–60 cm from the display. The experiment was conducted within a 960×650 pixel window with a white background. Glyphs were randomly positioned in the experiment window across a two-dimensional grid with a cell size of 52×52 pixels. The glyphs are re-sized such that users do not need to scroll the window in any of the varying conditions.

Procedure

Before the formal study, we organized an 1-h orientation seminar. During the seminar, we first introduced the concept of outlier detection and its wide application in many real-world scenarios. Next, we reviewed

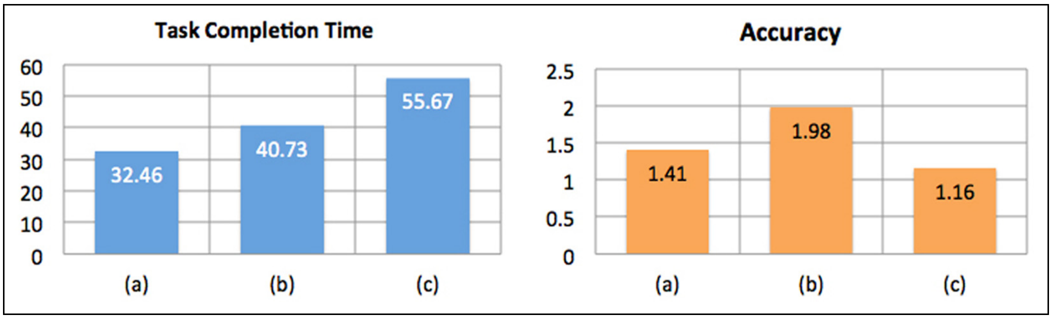


Figure 6. Comparing three different star glyph designs in terms of task completion time (in second) and number of correct answers (3 is the max corresponding to 100% accurate). The labels (a–c) indicate three different types of star glyph designs illustrated in Figure 5.

in detail the six different glyph designs and their interpretation in the context of outlier detection. Finally, we provided a brief lesson with instructions regarding the use of the prototype system.

During the instructional lesson portion of the seminar, users were shown how the study system would display a set of glyphs (all of the same type) from which the users would be asked to identify three outliers. Users were told to identify and select the outliers by clicking on the corresponding glyphs. The selection, which displays a blue highlight on the glyph, could be unselected by a second click on the glyph. Users were also shown the "next" button which was to be clicked when they considered themselves finished with the task. Clicking next would record the results and surface the visualization for the next task.

Following the group lesson, users were asked to practice using the study system using a sample dataset (24 tasks addressing all six glyph designs, two data scales, and two dimensionality scales). Finally, a question-and-answer session was held to address any remaining questions.

Once all users had received their orientation, we scheduled individual study sessions with each user. For each individual session, the order of the experiment was randomized, including both the order of the tasks and the order of glyphs. For each user study task, we used the same dataset with each type of glyph. The choice to reuse datasets across glyphs was made to allow a fair comparison of the observed results.

To avoid learning effects, glyph locations were shuffled when switching glyph designs, resulting in new locations for the outliers that users were asked to identify. In addition, the dimension ordering was shifted each time the location was changed. A shift in order, rather than a randomized order, was used because sequential relationships between dimensions can significantly affect the resulting visualized pattern (e.g. reordering is an important visual clutter reduction method[53]). Together, these two techniques ensured that for each of the six glyph types in a task, the users were looking at the same set of targets using the same dataset, but were unable to memorize the correct answer.

The users' task completion time and answer accuracies were recorded automatically by the study system and captured in a quantitative performance report. After performing the study tasks, the users completed a post-study questionnaire to gather subjective feedback. From start to finish, each session lasted approximately 30–45 min.

Task conditions

We performed a within-subjects study in which each user was required to complete 12 tasks using each of the six glyph designs, resulting in 72 trials per user. As mentioned above, we considered both large- and small-scale datasets, with both high and low dimensionality. We generated three distinct datasets for each of these trails, thus resulting 72×3 = 216 datasets, one of each testing trail. Considering the 18 users, the design produced 1296 unique trials.

Table 1. The design of study tasks.

	18	Users
×	6	Designs
×	2	Scales of the data (*small (25), large (100)*)
×	2	Scales of dimensions (*low (25), high (50)*)
×	3	Repetitions
	1296	Trials

Results

In this section, we report the results of our analysis of both the quantitative and qualitative results gathered during the study. First, we describe the effect of our two study variables (data size and dimensionality) on the overall task performance. Then we focus on a direct comparison of the glyph designs. Finally, we present the results from the post-study questionnaire.

Effects of data dimensionality and size. We investigate how the two study variables (dimensionality and data size) affect the task performances in a series of analysis. To this end, we separate the study results into four datasets based on different testing conditions (i.e. low/high dimension, small/large size). In each dataset, one variable was fixed and the other was tested based on repeated measures analysis of variance (RM-ANOVA) to take the glyph type into consideration, while making the comparison. Before the RM-ANOVA analysis, the data's normality and homogeneity were tested and the unsatisfied data were transformed (The Shapiro–Wilk test showed that some of the datasets were non-normally distributed. The inverse degree of freedom was used to transform the data into a normal distribution.). During the test, the assumption of sphericity was verified based on Mauchly's test. The degree of freedom was corrected using Greenhouse–Geisser estimate of sphericity when the assumption is violated. The following figures and descriptions summarize the testing results in detail.

According to Figure 7(a), when the number of dimensions was low, the task-completion time of the Z-Glyph family was less sensitive to the change of data size (i.e. time differences were relatively small) when compared to the baseline glyphs. However, RM-ANOVA analysis showed that size was a key

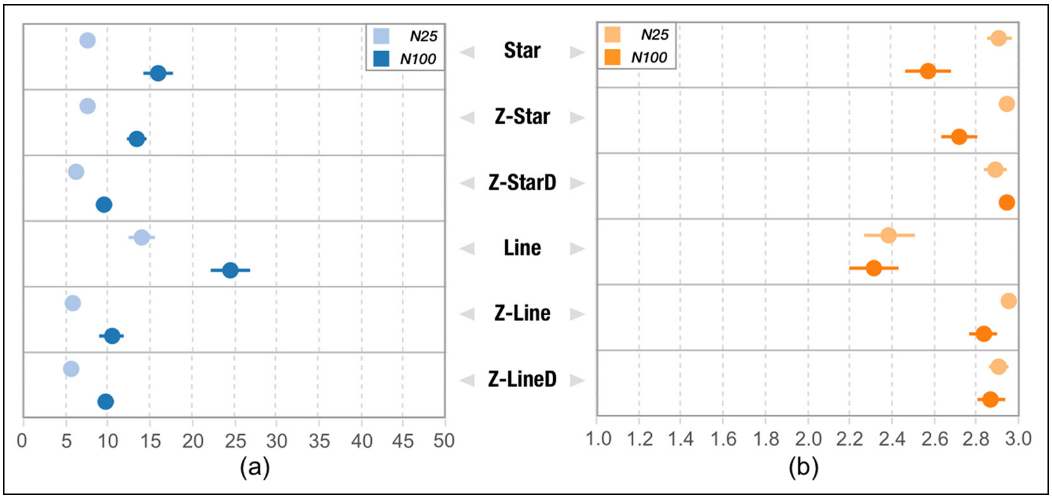


Figure 7. The effect of data size when dimensionality is 25 (low-dimensional): (a) mean time (D25) and (b) mean accuracy (D25).

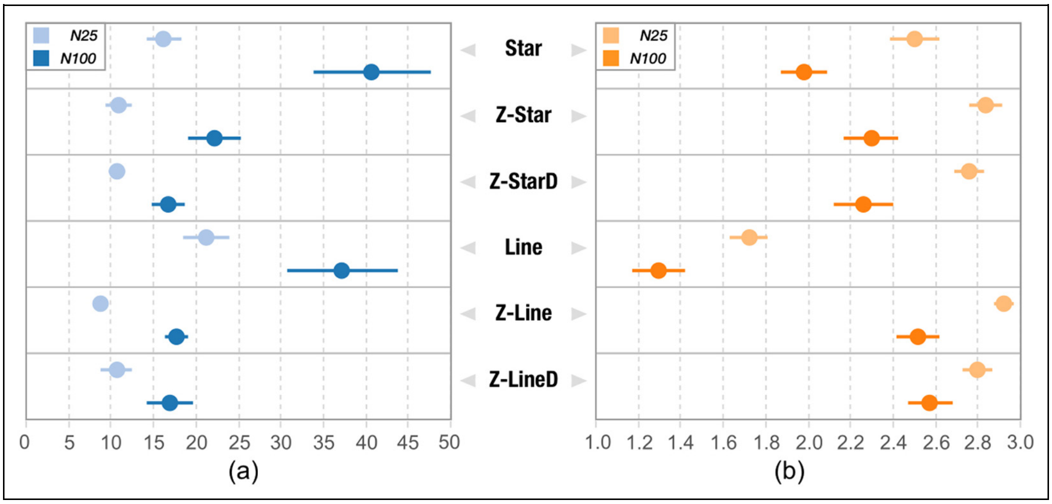


Figure 8. The effect of data size when dimensionality is 50 (high-dimensional): (a) mean time (D50) and (b) mean accuracy (D50).

factor which significantly affected users' performance ($F(1, 17) = 13.974$, $p < .05$) across all kinds of glyphs with faster speed for smaller datasets (N25). In terms of task accuracy (Figure 7(b)), Z-StarD and Z-LineD both proved most robust (less sensitive) to changes in dataset size, and RM-ANOVA test showed that overall, there was no significant change in users' ability to correctly identify outliers.

As in the low-dimensional case, high-dimensional data resulted in significantly slower performance ($F(1, 17) = 84.884, p < .05$) over all types of glyph designs (Figure 8). In this configuration, the impact on accuracy was also statistically significant ($F(1, 17) = 60.472, p < .05$). However, Z-Glyph family showed generally smaller impacts (i.e. has relatively less difference in accuracy when dimension is changed as shown in Figure 8), and Z-LineD glyph is the least impacted over all the glyphs.

When the data size was small (Figure 9), the task-completion time of the Z-Glyph family was affected less by changes in dimensionality compared to the baseline glyphs, although the overall drop in performance was statistically significant for all glyphs ($F(1, 17) = 62.813, p < .05$). For task accuracy, the baseline star glyphs suffered a large drop in performance, while the Z-Star family proved most robust.

Similar to the small data size case, task completion times for large datasets were significantly impacted ($F(1, 17) = 62.153, p < .05$) by changes in dimensionality (Figure 10). Moreover, in contrast to the small

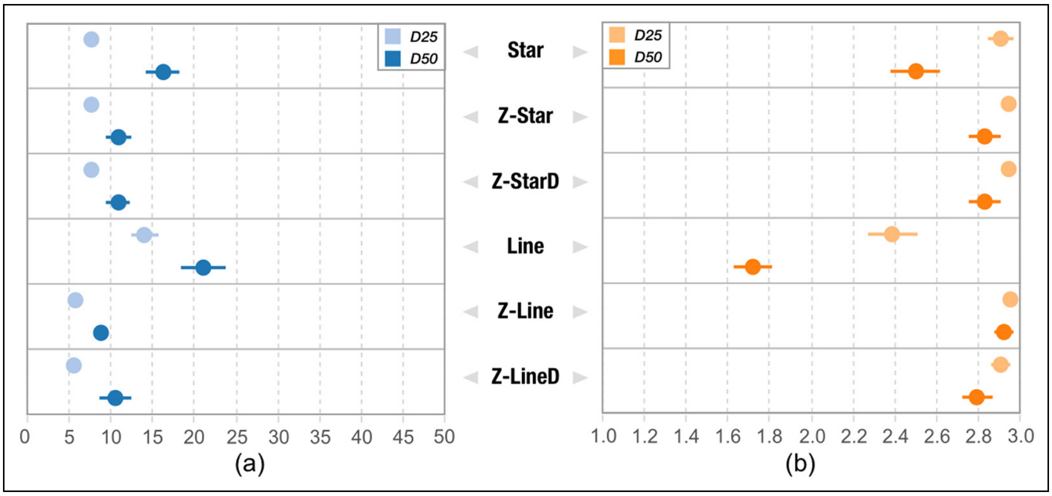


Figure 9. The effect of dimensionality when data size is 25 (small): (a) mean time (N25) and (b) mean accuracy (N25).

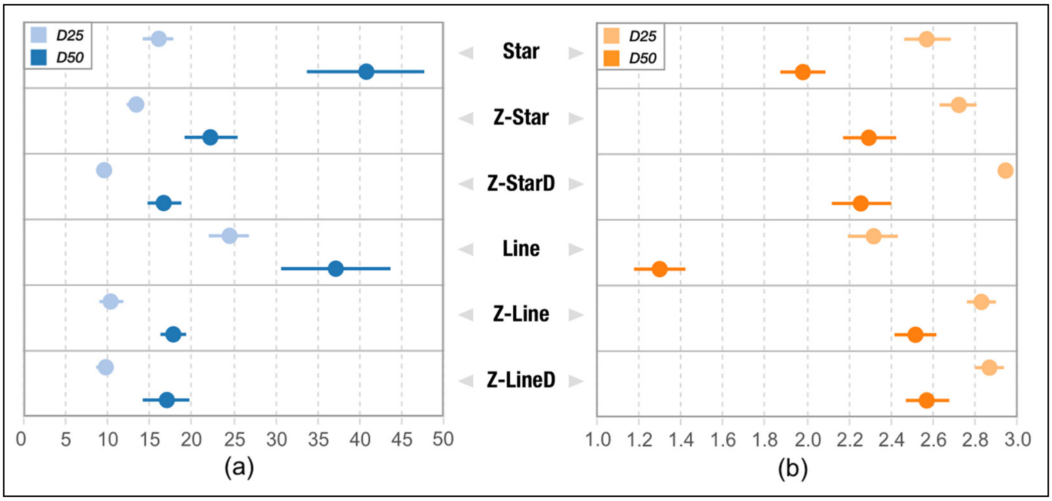


Figure 10. The effect of dimensionality when data size is 100 (large): (a) mean time (N100) and (b) mean accuracy (N100).

data case, task accuracy was also significantly impacted ($F(1, 17) = 143.5, p < .05$). However, as shown in Figure 10, the increase in time and decrease in accuracy were most strongly felt in the baseline designs.

In summary, both dimensionality and data size are key factors that may significantly affect task performance for all types of glyphs. The affection is expectable, i.e., the later the data size is or the higher the dimensionality is, the slower the performance will be. Comparatively speaking, Z-Glyph designs are performed more robust than that of the baseline glyphs.

Comparison of glyphs. While the results above show that data size and dimensionality broadly impact performance, there are also differences between specific designs. To quantify these differences, we compared the Z-glyph family to the two baseline glyphs (Star and Line) under different conditions using RM-ANOVA and analyzed the pairwise comparisons using Bonferroni correction. With respect to the null hypothesis, we assume that there is no difference in means between Z-Glyph family and baseline glyphs in terms of both task completion time and accuracy. Similar to the above analysis, the normality and homogeneity assumption were also tested and the data were transformed or the degree of freedom was corrected when the corresponding assumptions were violated. The analysis results are summarized in Figures 11 and 12 and described below in more detail.

T1 (N25-D25): finding outliers in 25 25-dimensional glyphs. The tests of within-subjects effect showed that these glyphs are significant different in terms of both task completion time ($F(5, 85) = 16.746, p < .01$) and accuracy ($F(5, 85) = 14.504, p < .01$). When compared to

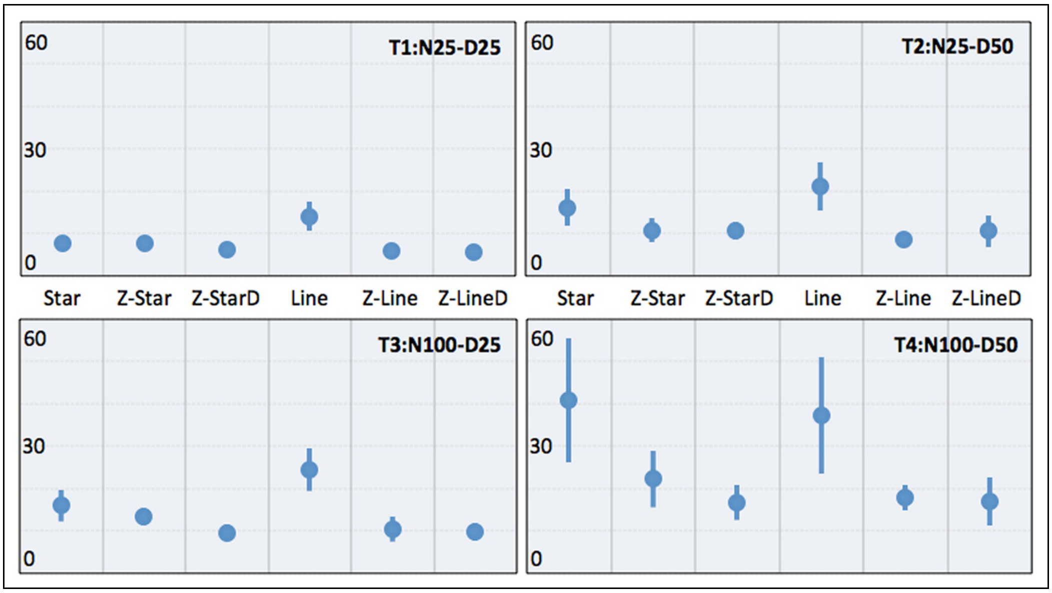


Figure 11. Comparing the mean task-completion time of six glyph designs under different conditions.

the baseline Line glyph, the whole Z-Glyph family was significantly better in terms of both time ($p < .05$) and accuracy ($p < .05$), which rejects the null hypothesis. When compared to the baseline Star glyph, however, the benefits of the Z-Glyphs were not significant, where null hypothesis is true.

T2 (N25-D50): finding outliers in 25 50-dimensional glyphs. The tests of within-subjects effect showed that these glyphs are significant different in terms of both task completion time ($F(5, 85) = 7.910, p < .01$) and accuracy ($F(5, 85) = 30.581, p < .01$). When compared to the Line glyph, the whole Z-Glyph family was significantly better than the Line glyph in terms of both time ($p < .05$) and accuracy ($p < .01$), but no significance was found between Z-Glyph family and the Star glyph.

T3 (N100-D25): finding outliers in 100 25-dimensional glyphs. The tests of within-subjects effect showed that these glyphs are significant different in terms of both task completion time ($F(5, 85) = 16.741$, $p < .01$) and accuracy ($F(5, 85) = 16.741, p < .01$). In particular, pairwise comparisons showed that the following cases reject the null hypothesis. When compared to the Line glyph, the Z-Glyph family was significantly better in terms of both task completion time (with all $p < .05$) and accuracy (with all $p < .05$). When compared to the Star glyph, the Z-StarD and Z-LineD glyphs were both significantly better in terms of task completion time (with $p < .05$). Z-StarD also had a significantly better accuracy (with $p < .05$).

T4 (N100-D50): finding outliers in 100 50-dimensional glyphs. The tests of within-subjects effect showed that these glyphs are significant different in terms of both task completion time ($F(5, 85) = 6.519, p < .01$) and accuracy ($F(5, 85) = 22.651, p < .01$). In particular, pairwise comparisons showed that the following cases reject the null hypothesis. When compared to the Line glyph, the whole Z-Glyph family produced significantly better task completion times (with all $p < .05$) and accuracy (with all $p < .05$). When compared to the Star glyph, the whole Z-Glyph family was significantly better in terms of task completion time (with all $p < .05$). The Z-Line and Z-LineD glyphs were significantly better than the Star glyph ($p < .05$) in accuracy.

Considering all four configurations, the Z-Glyph family outperformed the baseline glyphs by a wide margin for both task completion times and accuracy rates. Moreover, the effects were stronger as the datasets grew in size and dimensionality. There was no statistically significant difference between the different Z-Glyph designs. However, Z-Line and Z-LineD glyphs performed the best overall, and they outperformed the baseline glyphs in both time and accuracy under most conditions. The results suggest that the Line glyph is the worst option for the studied outlier detection tasks. However, the baseline Star glyph—contrary to our initial hypothesis—produced relatively strong performance results when the data size was small or data dimension was low. However, its limitations were revealed in the more complex conditions.

Post-study questionnaire. Users completed a post-study questionnaire with 13 questions designed to capture qualitative feedback. The first two questions in the survey asked users to choose which glyph type was most useful and easy-to-use for outlier detection. The

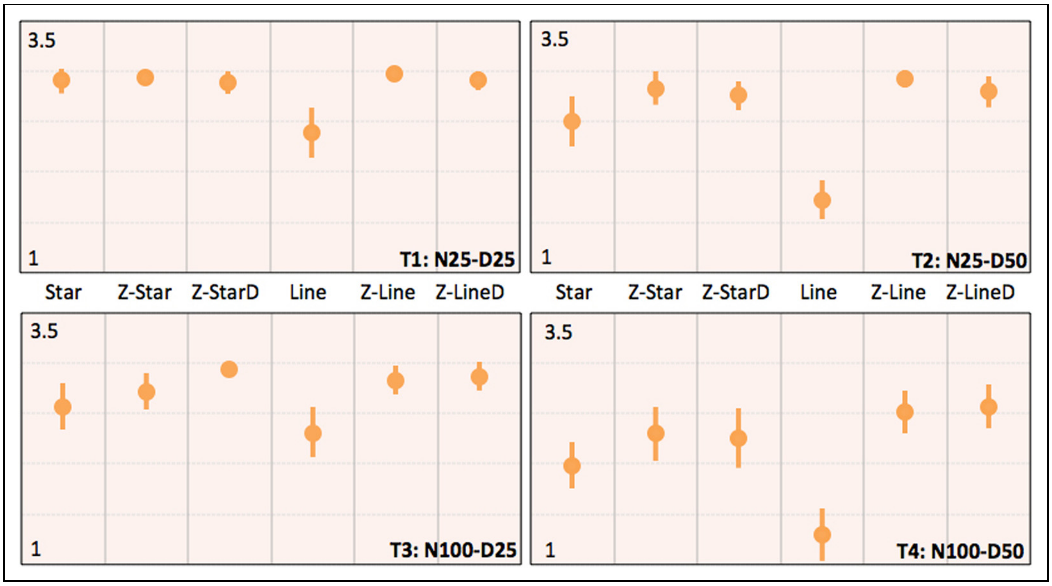


Figure 12. Comparing the mean of the numbers of correct answers (maximum is 3, the number of repetitions in our study design) reported based on different glyphs under different conditions.

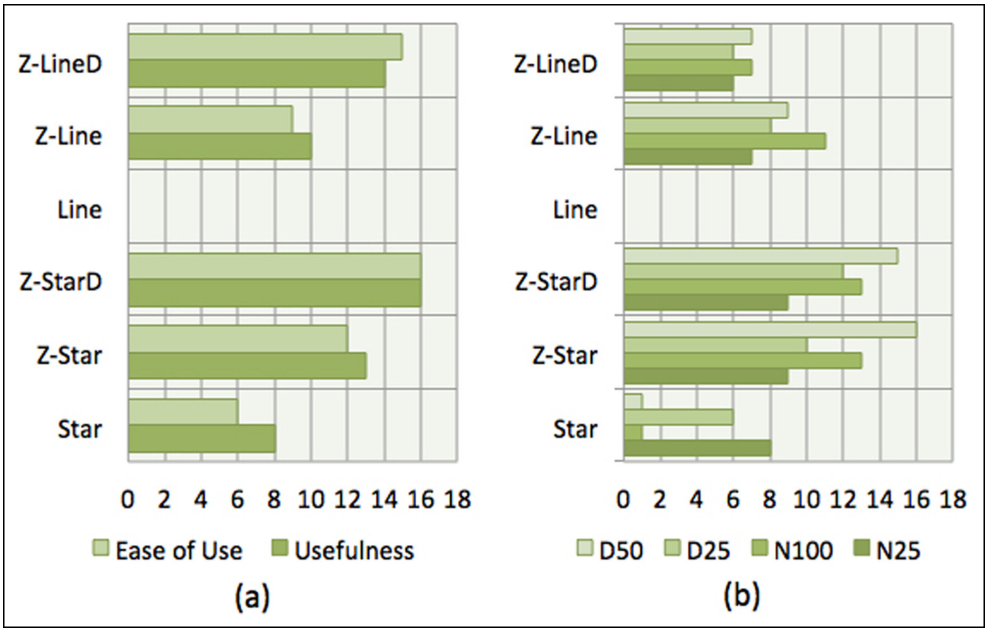


Figure 13. Users' ratings of different glyphs by considering (a) their usability and (b) their efficiency under different conditions. In the figure, *x*-axis indicates the number of ratings. A user was allowed to rate multiple glyphs at the same time.

results are shown in Figure 13(a). Questions 3–6 asked users to choose the glyph type most effective for outlier detection under specific conditions (large vs small datasets; low vs high dimensionality). The results are shown in Figure 13(b).

The baseline Line and Star glyphs were the least popular, mapping to the results, mirroring to some extent the performance measurements for these glyph types. However, surprisingly, however, the results show that the Z-Star and Z-StarD glyphs were most popular, even though the Z-Line and Z-LineD glyphs generally performed better in our quantitative evaluation.

In question 7, we investigated which visual attribute, shape or color, was considered most useful for

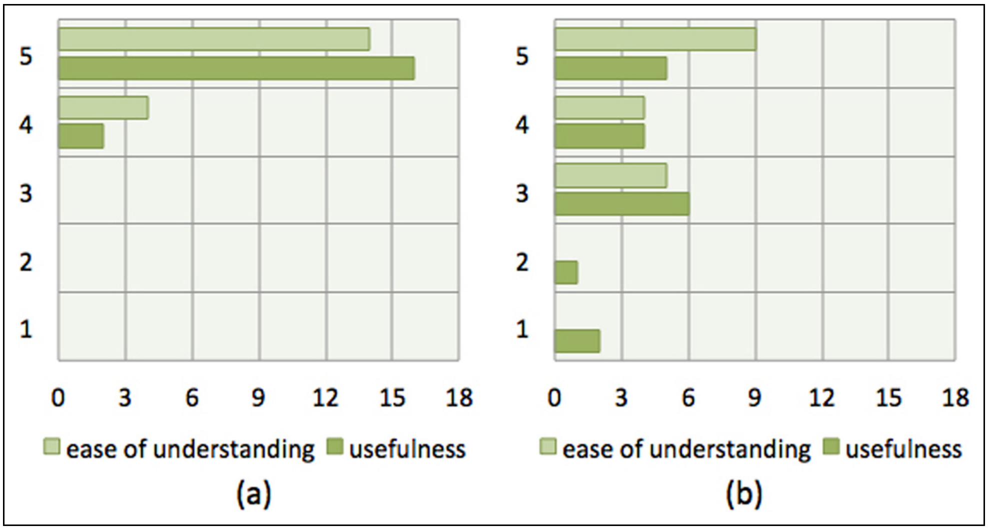


Figure 14. The usability of the two design factors: (a) standardization and (b) color enhancement, in Z-Glyphs. In this figure, *y*-axis indicates the rating score where 5 means very useful or very easy to understand, in opposite, 1 means not useful at all or very difficult to understand; *x*-axis indicates the number of ratings.

detecting outliers. The results show that all 18 users detected outliers by comparing glyph shapes of data items, but only 8 (less than half) reported taking color comparison into consideration.

Questions 8–11 focused on the utility and ease-of-use of the two key elements in the Z-Glyph construction process: standardization and color enhancement. The results (Figure 14) show that standardization was considered very useful by all users (Figure 14(a)). Color enhancement, in contrast, received less support, although the responses were still positive overall.

The final two questions were free response questions asking for feedback as to the advantages and disadvantages of the Z-Glyph design. The most valuable feedback from these questions is reported in the “Discussion” section.

Discussions

Both the user study statistics and the questionnaire results provide valuable insights into when and how the Z-Glyph design is useful.

When should Z-Glyphs be used?. The Z-Glyph is designed to support outlier detection tasks for all types of multivariate data in which (1) the data are normal deviate or (2) the data can be transformed to be close to the location-scale distributions. The study results showed that the Z-Glyph family of designs produced faster performance times with more higher accuracy rates when compared to the baseline designs. This held true nearly universally across the evaluated variable space (small vs large; low-dimensional vs high-dimensional), with increasing benefits as the visualized data grew more complex. More specifically, within the Z-Glyph family of designs, the Z-Line and Z-LineD glyphs outperformed the others in most cases. These are recommended as a first choice in most real-world applications.

Why was the Star glyph family popular?. While the Z-Line and Z-LineD glyphs produced the quantitative results for speed and accuracy, users reported a preference in their post-study feedback for the Star family over the Line family of glyphs (see Figure 13(b)).

The reasons were found in users’ comments collected in the questionnaire. Users’ free responses in the questionnaire help explain this apparent discrepancy in aesthetic terms: “They [the star glyph family] are in a circular shape, making the design more compact and also making the eyes more comfortable when looking at those images for a long time.”

Why did Z-Line(D) outperform Z-Star(D)?. Clues to the benefits of the Line-based version of this glyph were found in feedback gathered from the study users. In particular, two users reported a critical problem: the circular shape of the star-based glyphs produced a “smoothing” of the irregular shape patterns that serve as a primary encoding for outlier detection within the Z-Glyph design. Echoing this challenge, another user said, “when the number of features is very large, the differences of the shapes are limited in Z-Star glyphs.” Yet another user mentioned that “all the zigzag shapes become unclear in the circular arrangement. Picking up outliers from a large set of data thus becomes difficult." Similarly, another reason by users was the "need to calculate the area in my mind to figure out the outliers, the circular ones making this calculation a little bit difficult."

Why did colors provide little help?. The lack of effectiveness for color-coding was especially surprising. Using color to highlight differences from the norm was a major part of the Z-Glyph design, and the expectation was that it would be valuable for the outlier detection task.

However, as one user said explicitly, the shape is the dominant feature used to make judgements and the color often proved distracting:

> the shapes come first, then the color helps. But when there are a large number of features, the color seems to dazzle the eyes and makes it very tired. Also, it doesn't help to distinguish the shape when the features are too many and each one is too small; the color makes it harder to distinguish the shape differences. The Z-Star glyph seems better here.

Another user mentioned that

> focus on colors [meant] I was looking at outliers with respect to the color distributions of all glyphs, rather than being able to detect outliers with respect to the provided baselines in each glyph. I [felt] that this lead to a high false positive rate.

Despite these reservations, a majority of users still believed that using colors was useful, and that it resulted in a more aesthetically pleasing visualization. There was also a suggestion that colors would be more useful for larger glyphs where more pixels were available to depict the graphics.

Domain expert interview

We conducted interviews with two domain experts to further evaluate the proposed Z-Glyph designs. The first is an expert in information security and the second is a medical doctor with dual certification in internal medicine and pediatrics. In this section, we report our procedure and present the interviews' results.

Procedure

The two interviews were both conducted in the form of a short-term case study, during which the expert was asked to identify outliers from a dataset relevant to their expertise. Each interview started with a tutorial period. The tutorial explained the outlier detection concept, described the various glyph designs, presented an overview of the outlier explorer system, and had the experts begin interacting with the system on their own. Once the experts were proficient with the prototype system, they were asked to find outliers in a prepared dataset appropriate to their area of expertise. During this procedure, we conducted a semi-structured interview that included questions about various aspects of the glyph designs, overall usefulness, ease of use, and general pros and cons of the approach taken. Each interview lasted about 1 h and was recorded and notes were taken.

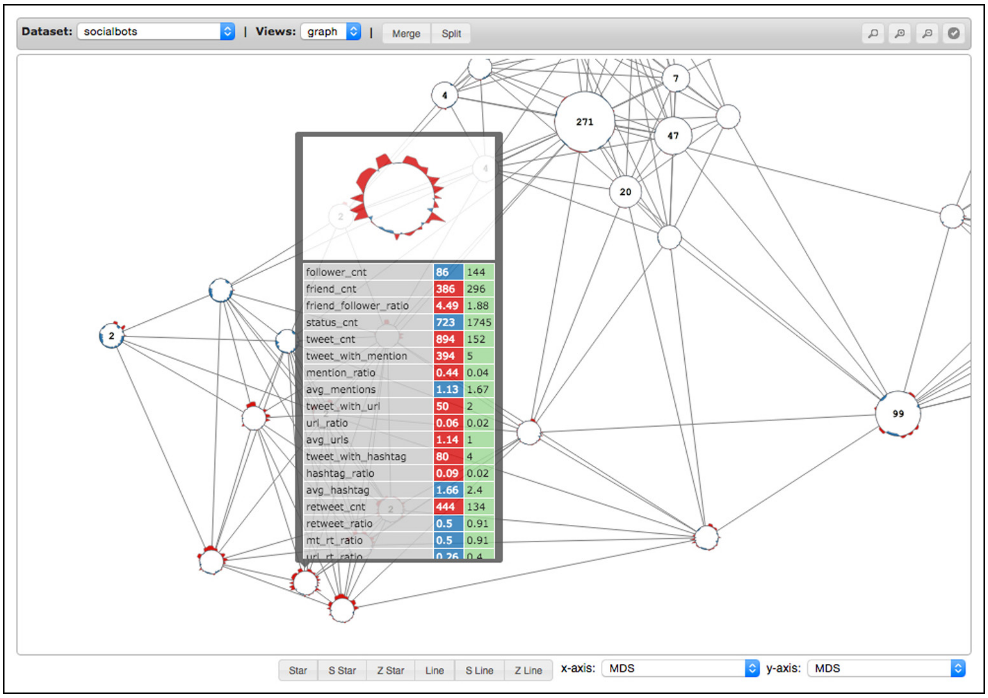
Outlier explorer

To support the interview, we developed a prototype *Outlier Explorer* in which data points are visualized as the glyphs using the designs outlined in this article and arranged using graph layout algorithms or MDS projection depending on the structure of the data (Figure 15). The system is highly interactive, allowing users to zoom in and out, and to pan their view to focus on specific sections of the dataset. To prevent occlusions when zooming out, glyphs are automatically aggregated into meta-glyphs based on averaged feature values when the boundaries of two or more glyphs begin to overlap. Similarly, the meta-glyphs are then split into multiple smaller glyphs when zooming in, which provides more room. The expert users were also able to switch between different glyph styles, with Z-StarD used as the default.

Interview I: detecting suspicious users in Twitter

The first interview was conducted with an expert in information security. The expert is a male professor at a highly ranked US University with more than 20 years of experience in the field. The dataset for this interview contained statistics for 500 Twitter accounts, 30 of which were social bots rather than normal users. These 500 accounts were sampled from a larger Twitter dataset in which each account was described by a 58-dimensional feature vector capturing various social behaviors (for details about the features and the dataset, see Cao et al.[8]). The data were rendered as a graph, with nodes representing user accounts and links representing communication paths (mentions, retweets, and so on). A screenshot of the explorer is shown in Figure 15. The information security expert was asked to examine these data to identify the bot accounts.

The expert identified a group of the most suspicious users with just a first glance at the outlier explorer.



Dataset: socialbots | Views: graph

follower_cnt	86	144
friend_cnt	386	296
friend_follower_ratio	4.49	1.88
status_cnt	723	1745
tweet_cnt	894	152
tweet_with_mention	394	5
mention_ratio	0.44	0.04
avg_mentions	1.13	1.67
tweet_with_url	50	2
url_ratio	0.06	0.02
avg_urls	1.14	1
tweet_with_hashtag	80	4
hashtag_ratio	0.09	0.02
avg_hashtag	1.66	2.4
retweet_cnt	444	134
retweet_ratio	0.5	0.91
mt_rt_ratio	0.5	0.91
url_rt_ratio	0.26	0.4

x-axis: MDS y-axis: MDS

Figure 15. Visualizing Twitter users' behaviors in Z-Star Glyphs.

"Oh, this is obvious" he said while identifying the group. "All the abnormal ones are already highlighted in colors" and the "shapes also provide some cue."

The expert then zoomed in to view the suspicious group in more detail. The accounts in this group had many feature values that were well above average. Hovering the mouse over each of the accounts, the expert investigated the detailed feature values which were shown via a tooltip. He found the most suspicious user account based on the glyph design. based on the glyph design (shown in Figure 15). He found that the account had a rather small number of followers (below average) but had a very high retweeting rate. In addition, the account's messages had a high ratio of mentions and contained many URL links. The expert felt that this appeared to be behavior typical of a spammer. It was confirmed later that the expert's suspicion was correct, and that he had indeed identified a bot.

After comparing views of the data using various glyph designs, the expert believed that the glyphs without colors (i.e. Star, Z-Star, Line, Z-Line) were difficult to read. "It is difficult to see these lines (referring to the polylines shown feature values in the glyphs) as they intersect with these graph links." He stated a preference for the Z-StarD and Z-LineD glyphs, and believed that both of them were well designed for the outlier identification task.

Overall, the expert felt that the Z-Glyph designs were "simple but informative" and expressed the desire to adopt the Z-StarD glyph design in some of this own work. However, he also provided valuable comments regarding limitations of the prototype explorer, which we present in the discussion later in this section.

Interview II: finding high-risk patients

The second interview was conducted with a medical doctor. The expert is a female clinician with dual specialty in internal medicine and pediatrics. The dataset for this interview contained data from a cohort of patients, some of whom were suffering from chronic kidney disease (CKD). The remaining patients were generally healthy. Each patient was represented by a 24-dimensional feature vector describing factors such as age, blood pressure, and various medical test results.[54] The doctor was asked to examine the patient population to identify patients most likely to have CKD.

Given the independence between patients (in contrast to the Twitter accounts, which interacted with each other), the data for this interview was visualized using a layout based on the MDS projection. MDS attempts to make distances in screen space reflect inter-item similarity measures, resulting in similar items appearing proximate to each other when the positions are used for visualization. A scatter-plot view was also included in the prototype, in which layout was driven by specific feature values (see Figure 16).

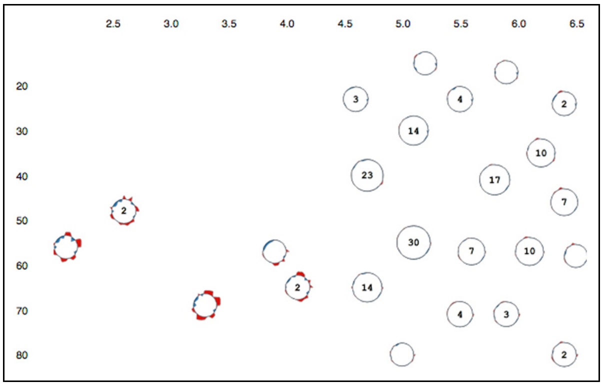


Figure 16. The scatter plot view of the patient dataset. X-axis shows the data dimension of "red-blood-cell-count" and Y-axis indicates the dimension "patient-age".

The doctor was able to immediately identify a number of suspicious glyphs. "These on the periphery. The ones with more red, or blue." She then used the tooltop to inspect the clinical indicators and verify her initial hypotheses. When asked her to compare different glyphs, she said "I liked [Z-StarD] the best." Continuing, she stated that "the others are harder to interpret at a glance," and that "Z-StarD is the easiest." When asked to explain the reason, she simply stated that "the other ones are just harder to look at." Moreover, in reference to the baseline glyphs, she suggested that "in a clinical context, I would worry that I would misinterpret. To get it wrong, not life or death, but [it] could really mess up the course of treatment."

Finally, the doctor felt that the system would be useful for population management. In particular, she discussed the job of assigning limited resources to challenging patients, and that this difficult job often falls on the shoulders of the actual physicians. She felt that the outlier explorer could help them figure out which patients were the best ones to select for special attention when allocating those resources.

Discussion

The expert interviews described above reinforced the idea that real-world outlier detection tasks are quite challenging. Detailed domain knowledge and human judgement were essential in correct data interpretation. With this in mind, the Z-Glyphs were designed to help embed a "human in the loop" within the outlier detection process to help address the two major challenges mentioned in the introduction. The current design was mostly well received by the domain experts. In particular, their feedback verified that Z-Glyphs are more effective than the baseline glyphs in assisting outlier detection. Interestingly, however, the first expert believed Z-Line glyphs were less effective when compared to Z-Star glyphs for graph visualization, where the lines may intersect with the graph links. This potentially introduced visual clutter that could affect users' judgment. This finding contradicts with our hypothesis and experimental results, but also provides a useful insight about how to make different design decisions given different conditions. In addition, all of the experts believed that the glyphs with color enhancement were more helpful. This verified our hypothesis but contradicted the experimental results. We believe this is due to the data items in outlier explorer are laid out according to their similarities. This approach produced a meaningful placement that proved helpful in revealing color patterns.

However, the experts also identified limitations. First, although it is a common practice to use Z-scores to identify possible outliers, this can be misleading (particularly for small sample sizes) due to the fact that the maximum Z-score is at most $(n-1)/\sqrt{n}$. To overcome this limitation, we allow users to manually set the baseline values based on their domain knowledge. For example, a doctor could enter a normal lab test value as the domain-appropriate baseline.

Second, the design of Z-Glyph are most suitable for data whose feature values are follow the normal distribution. If that condition does not hold, patterns may not emerge. To address this issue, data can be transformed to approximate a normal distribution. We have adopted this approach when appropriate using the Box–Cox transformation.[55]

Finally, baselines in the Z-Glyph design represent a single value where at times a range may be desired. This could be accomplished replacing the baseline with a "base-belt" whose thickness represents a value range.

Conclusion and future work

In this article, we introduced the family of Z-Glyphs, the first set of glyphs that were designed for revealing outliers in a multivariate dataset. We introduced a design scheme which converts a traditional glyph into Z-Glyphs in a procedure of standardization and color enhancement. We designed and conducted a controlled user study to test their performances in terms of revealing outliers under different conditions. Our results showed that the Z-Glyph family outperforms the baseline glyph designs when the data are large and dimensions are high. Among all our Z-Glyph implementations, Z-Line glyph has the best performance and Z-StarD glyph is the most favorite. We also conducted in-depth interviews with two domain experts from different areas. Their feedback further verified the effectiveness of our designs. The future work includes testing Z-Glyph's performance based on more tasks and applying Z-Glyph to solve real world problems in different application domains and keep developing the outlier explorer by adding more interactions as well as advanced active learning-based anomaly detection algorithms.

Acknowledgements

The authors would like to thank all the reviewers for their constructive comments. The authors would also like to thank all the users and domain experts who participated in their user study. Special thanks are given to Dr Wen-Ting Chung and Ms Jingjing Ding for their help on analyzing the user study results.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work is a part of the research supported from NSFC grant no. 61602306, NSF grant no. 1637067, and IBM 2016 SUR Award.

References

1. Chandola V, Banerjee A and Kumar V. Anomaly detection: a survey. *ACM Comput Surv* 2009; 41(3): 15.
2. Edgeworth FY. On discordant observations. *Philos Mag* 1887; 23(143): 364–375.
3. Aggarwal CC. *Outlier analysis*. New York: Springer Science + Business Media, 2013.
4. Axelsson S. Visualization for intrusion detection. In: *Proceedings of the 8th European symposium on research in computer security*, Gjøvik, 13–15 October 2003, pp. 309–325. Berlin: Springer.
5. Corchado E and Herrero A. Neural visualization of network traffic data for intrusion detection. *Appl Soft Comput* 2011; 11(2): 2042–2056.
6. Tsai CF, Hsu YF, Lin CY, et al. Intrusion detection by machine learning: a review. *Expert Syst Appl* 2009; 36(10): 11994–12000.
7. Teoh ST, Ma KL, Wu SF, et al. Case study: interactive visualization for internet security. In: *Proceedings of the information visualization*, Boston, MA, 27 October–1 November 2002, pp. 505–508. New York: IEEE.
8. Cao N, Shi C, Lin S, et al. TargetVue: visual analysis of anomalous user behaviors in online communication systems. *IEEE T Vis Comput Gr* 2016; 22(1): 280–289.
9. Zhao J, Cao N, Wen Z, et al. #FluxFlow: visual analysis of anomalous information spreading on social media. *IEEE T Vis Comput Gr* 2014; 20(12): 1773–1782.
10. Chen Y, Zhou XS and Huang TS. One-class SVM for learning in image retrieval. In: *Proceedings of the IEEE image processing*, Thessaloniki, 7–10 October 2001, vol. 1, pp. 34–37. New York: IEEE.
11. Song Y, Wen Z, Lin CY, et al. One-class conditional random fields for sequential anomaly detection. In: *Proceedings of the 23rd international joint conference on artificial intelligence*, Beijing, China, 3–9 August 2013, pp. 1685–1691. New York: ACM.
12. Angiulli F and Pizzuti C. Outlier mining in large high-dimensional data sets. *IEEE T Knowl Data En* 2005; 17(2): 203–215.
13. Kind A, Stoecklin MP and Dimitropoulos X. Histogram-based traffic anomaly detection. *IEEE T Netw Serv Manag* 2009; 6(2): 110–121.
14. Lin J, Keogh E and Lonardi S. Visualizing and discovering non-trivial patterns in large time series databases. *Inform Visual* 2005; 4(2): 61–82.
15. Laskov P, Rieck K, Schafer C, et al. Visualization of anomaly detection using prediction sensitivity (no. *2)*, 2005, pp. 197–208, https://koreauniv.pure.elsevier.com/en/publications/visualization-of-anomaly-detection-using-prediction-sensitivity
16. Haslett J, Bradley R, Craig P, et al. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *Am Stat* 1991; 45(3): 234–242.
17. Kruskal JB and Wish M. *Multidimensional scaling*, vol. 11. Newbury Park, CA: SAGE, 1978.
18. Jolliffe I. *Principal component analysis*. Hoboken, NJ: Wiley Online Library, 2002.
19. Inselberg A and Dimsdale B. Parallel coordinates. In:Klinger A (ed.) *Human-machine interactive systems*. New York: Springer, 1991, pp. 199–233.
20. Kandogan E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: *Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Francisco, CA, 26–29 August 2001, pp. 107–116. New York: ACM.
21. Muñoz A and Muruzábal J. Self-organizing maps for outlier detection. *Neurocomputing* 1998; 18(1): 33–60.
22. Novotny M and Hauser H. Outlier-preserving focus + context visualization in parallel coordinates. *IEEE T Vis Comput Gr* 2006; 12(5): 893–900.
23. Thom D, Bosch H, Koch S, et al. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In: *Proceedings of the IEEE Pacific visualization symposium*, Songdo, Korea, 28 February–2 March 2012, pp. 41–48. New York: IEEE.
24. Borgo R, Kehrer J, Chung DH, et al. Glyph-based visualization: foundations, design guidelines, techniques and applications. *Eurograph State Art Rep* 2013; 39–63.
25. Zhou H, Xu P, Yuan X, et al. Edge bundling in information visualization. *Tsinghua Sci Technol* 2013; 18(2): 145–156.
26. Zhou H, Xu P and Qu H. Visualization of bipartite relations between graphs and sets. *J Visual* 2015; 18(2): 159–172.
27. Xu P, Du F, Cao N, et al. Visual analysis of set relations in a graph. *Comput Graph Forum* 2013; 32: 61–70.
28. Cao N, Lin YR, Sun X, et al. Whisper: tracing the spatiotemporal process of information diffusion in real time. *IEEE T Vis Comput Gr* 2012; 18(12): 2649–2658.

29. Cao N, Lu L, Lin YR, et al. SocialHelix: visual analysis of sentiment divergence in social media. *J Visual* 2015; 18(2): 221–235.
30. Wu Y, Wei F, Liu S, et al. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE T Vis Comput Gr* 2010; 16(6): 1109–1118.
31. Fuchs J, Isenberg P, Bezerianos A, et al. The influence of contour on similarity perception of star glyphs. *IEEE T Vis Comput Gr* 2014; 20(12): 2251–2260.
32. Abdul-Rahman A, Maguire E and Chen M. Comparing three designs of macro-glyphs for poetry visualization. In: *Proceedings of the Eurographics conference on visualization*, Swansea, 9–13 June 2014.
33. Chung DH, Legg PA, Parry ML, et al. Glyph sorting: interactive visualization for multi-dimensional data. *Inform Visual* 2015; 14(1): 76–90.
34. Duffy B, Thiyagalingam J, Walton S, et al. Glyph-based video visualization for semen analysis. *IEEE T Vis Comput Gr* 2015; 21(8): 980–993.
35. Ropinski T, Oeltze S and Preim B. Survey of glyph-based visualization techniques for spatial multivariate medical data. *Comput Graph* 2011; 35(2): 392–401.
36. Ropinski T and Preim B. Taxonomy and usage guidelines for glyph-based medical visualization. In: *Proceedings of the 19th conference on simulation and visualization (SimVis)*, 2008, pp. 121–138.
37. Fuchs J, Fischer F, Mansmann F, et al. Evaluation of alternative glyph designs for time series data in a small multiple setting. In: *Proceedings of the ACM SIGCHI conference on human factors in computing systems*, Paris, 27 April–2 May 2013, pp. 3237–3246. New York: ACM.
38. Maguire E, Rocca-Serra P, Sansone SA, et al. Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments. *IEEE T Vis Comput Gr* 2012; 18(12): 2603–2612.
39. Hlawatsch M, Sadlo F, Jang H, et al. Pathline glyphs. *Comput Graph Forum* 2014; 33: 497–506.
40. Jarema M, Demir I, Kehrer J, et al. Comparative visual analysis of vector field ensembles. In: *Proceedings of the IEEE conference on visual analytics science and technology*, Chicago, IL, 25–30 October 2015, pp. 81–88. New York: IEEE.
41. Jäckle D, Senaratne H, Buchmüller J, et al. Integrated spatial uncertainty visualization using off-screen aggregation. In: *Proceedings of the EuroVis workshop on visual analytics*, 2015.
42. Chan YH, Correa CD and Ma KL. The generalized sensitivity scatterplot. *IEEE T Vis Comput Gr* 2013; 19(10): 1768–1781.
43. Kachkaev A, Wood J and Dykes J. Glyphs for exploring crowd-sourced subjective survey classification. *Comput Graph Forum* 2014; 33: 311–320.
44. Erbacher RF, Walker KL and Frincke DA. Intrusion and misuse detection in large-scale systems. *IEEE Comput Graph* 2002; 22(1): 38–47.
45. Fry BJ. *Organic information design*. PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.
46. Xiong R and Donath J. PeopleGarden: creating data portraits for users. In: *Proceedings of the ACM symposium on user interface software and technology*, Asheville, NC, 7–10 November 1999, pp. 37–44. New York: ACM.
47. Gleicher M, Albers D, Walker R, et al. Visual comparison for information visualization. *Inform Visual* 2011; 10(4): 289–309.
48. Saito T, Miyamura HN, Yamamoto M, et al. Two-tone pseudo coloring: compact visualization for one-dimensional data. In: *Proceedings of the IEEE symposium on information visualization*, Minneapolis, MN, 23–25 October 2005, pp. 173–180. New York: IEEE.
49. Heer J, Kong N and Agrawala M. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Boston, MA, 4–9 April 2009, pp. 1303–1312. New York: ACM.
50. Heinrich J and Weiskopf D. State of the art of parallel coordinates. In: *STAR proceedings of Eurographics*, 2013, pp. 95–116.
51. Cornsweet T. *Visual perception*. New York: Academic Press, 2012.
52. Hulleman J. The mathematics of multiple object tracking: from proportions correct to number of objects tracked. *Vision Res* 2005; 45(17): 2298–2309.
53. Ellis G and Dix A. A taxonomy of clutter reduction for information visualisation. *IEEE T Vis Comput Gr* 2007; 13(6): 1216–1223.
54. Rubini L. Early stage of Indians chronic kidney disease dataset, 2015, http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
55. Sakia R. The Box-Cox transformation technique: a review. *J Roy Stat Soc D: Stat* 1992; 41: 169–178.