# Guiding the Exploration of Scatter Plot Data Using Local Interest Measures

Lin Shao, Timo Schleicher, Michael Behrisch, Tobias Schreck and Daniel A. Keim
University of Konstanz
Konstanz, Germany

***Abstract*—Scatter plots are effective diagrams to visualize distributions, clusters or correlations in two-dimensional data space. For high-dimensional data, scatter plot matrices can be formed to show relationships in all pairwise combinations of dimensions. Finding interesting patterns in large scatter plot spaces is a challenging problem and becomes even worse with an increasing number of dimensions. Previous approaches for exploring large scatter plot spaces have focused on ranking scatter plots based on their global interestingness. However, often local patterns contribute significantly to the interestingness of a scatter plot and cannot be reflected appropriately by global quality measures. In this paper, we present an approach to automatically segment significant local scatter plot patterns, also called scatter plot motifs. Based on the frequency properties of the local motifs, we compute local and eventually global quality measures to filter, rank and compare scatter plots and their incorporated motifs. We demonstrate the usefulness of our approach with one synthetic and one real-world data set and showcase a data exploration tool that visualizes the distribution of local scatter plot motifs in relation to the overall scatter plot space.**

## I. Introduction

Nowadays, vast amounts of data are rapidly created in many application domains and thus the problem of effective and efficient access to large multivariate and high-dimensional data arises. While in the past, the storage capacity was the primary problem, today the challenges comprise tasks like detecting interesting patterns or correlations in large data sets. One solution is to apply suitable visualization techniques and search for hidden information within the data. *Scatter plot* visualizations are one of the most widely used and well-understood visual representations for bivariate data. They can also be applied for high-dimensional data via dimensionality reduction or the scatter plot matrix representation [1]. However, perceiving and finding interesting scatter plots in large scatter plot collections constitutes a severe challenge, especially when working with scatter plot matrices.

Since expert interests are usually diverse in nature (e.g., application-dependent and potentially subjective), one of the most outstanding exploration challenges is how to support users in gaining insights from large data sets. Manually searching through large amounts of data views or -sections is exhaustive and may become infeasible for high-dimensional data sets. Recent work in Visual Analytics has focused on computing interestingness measures, which can be used to filter and rank large data spaces to present the user a good starting point for exploration. Specifically, several previous approaches, such as [2], [3], have focused on interestingness measures based on *global* properties of scatter plots for ranking and filtering. However, global interesting scores neglect the intuitive understanding that local motifs –or combination thereof– contribute significantly to the global interestingness of a view.

In this paper, we present a novel approach to automatically discover interesting scatter plot views, which opposed to current quality metrics focuses on scatter plot interestingness derived from *local* properties. Therefore, we enhance a minimum spanning tree-based clustering technique for a non-parametric segmentation of scatter plots and derive local areas-of-interest. In a second step, we deploy ideas from the image analysis domain –specifically interest point detection– to scatter plots. Next, we derive the interestingness of the scatter plots in terms of its Bag-of-Visual-Words vocabulary. The idea is that visually discriminatory motifs are considered of interest, since they can be quickly recognized by the human. Therefore, we introduce the Bag-of-Visual-Words concept for scatter plots and transfer the idea of term-based TF-IDF to this domain. Thus, we can derive the interestingness of a local scatter plot motif based its occurrence among and within the scatter plot corpus. We make use of these local motif-based measurements to rank and filter large scatter plot spaces.

We claim the following technical contributions:

- We adapt the minimum spanning tree-based clustering technique for a non-parametric segmentation of scatter plot motifs.
- We introduce a motif based dictionary to assess the interestingness of local scatter plot patterns.
- We define a global interestingness score based on the occurrence and similarity of local motifs.

The remainder of this paper is structured as follows: In Section II, we discuss related work and show commonalities and highlight differences. Section III gives an overview of our general idea to use local motif analysis for computing local and global interestingness measures. In Section IV, we will present our approach with a technical focus. Next, in Section V we apply our implementation to different data sets and showcase a local motif driven exploration. Finally, Section VI concludes the paper with a discussion and points to future work.

## II. Related Work

Several works enhance the exploration of large scatter plot data sets by means of ranking, filtering and searching functionalities. We next review a selection of works in the context of our approach.

### A. Visualization of Scatter Plot Patterns

Visualizations of scatter plots need to have an appropriate aspect ratio and scale to reveal correlations, patterns, trends and clusters. This is challenging since the identification of patterns in scatter plots, and the notion of interestingness, is subjective in nature. Most of the existing aspect ratio optimization methods rely on properties of line segments displayed in a plot. In [4] the authors suggested to use segments of a virtual polyline that connects all existing data points of a scatter plot, or the segments of a regression line through the plot. Talbot et al. [5] showed that this approach is tailored for data containing trends, but is inappropriate for data, which do not have this kind of functional relationship. Hence, they proposed a method based on contour lines resulting from a kernel density estimation, which is able to deal with pairs of variables without functional relationship. In a recent approach, Fink et al. [6] present a scatter plot aspect ratio calculation that is based on the Delaunay triangulation of the data points. The authors claim that the aspect ratio is appropriate if the edges of the Delaunay triangulation have nice geometric properties. More generally, a study on perceptional factors, which links scatter plot properties with perceived interestingness and interpretability is given in [7].

### B. Feature-Based Analysis of Scatter Plots

Automatic identification of interesting candidates within large sets of scatter plots has recently been an active field of research. The Scagnostics method [8] is a well-known feature-based approach, which proposes a set of graph-based measures for scatter plots, to describe the data properties. While the Scagnostics method does not require classified data, consistency measures [9] can further improve the identification of informative scatter plots for the case that class labels are available. In [10], a multi-step analysis of large scatter plot matrix spaces was introduced. The approach is based on visual quality measures, matrix reordering, and visual abstraction, and supports navigation and analysis in large scatter plot data.

Often, different scatter plot views need to be compared. In [11], two-dimensional color-coding was applied to compare sets of scatter plots for topological relationships. Other works supported the comparison of sets of scatter plots by automatic and interactive approaches. Albuquerque et al. [12] introduced an importance-aware sorting algorithm to find good projections in scatter plot matrices. A recently tackled problem is the identification of interesting subspaces in high-dimensional data, using scatter plots of projected subspaces. In [13], a sampling approach was shown that identifies interesting subspace projections for high-dimensional data sets. In [14], a visual approach for the identification of interesting subspaces was proposed. It relies on a clustering-based subspace search method to compute the interestingness score from density and class-separation measures.

### C. Navigation in Scatter Plot Space

The effectiveness of analyzing large scatter plot data also depends on appropriate navigation facilities. Animated navigation and extrusion-based transitions between views was proposed in [15] to navigate in scatter plot matrix spaces. Scherer et al. [16] introduced a search and navigation interface that is based on the scatter plots global regression features. Furthermore, in [17] an experimental study compared the effectiveness of global features for ranking scatter plot similarities.

### D. Delineation of our Approach and Further Relations

Our work uses a feature-based approach for an interestingness ranking of scatter plots based on their contained local motifs. Other than previous approaches, which concentrate on global features (i.e., derived from the entire plot), we here consider *local* properties of interest in scatter plots. Therefore, we complement global approaches. Our work is inspired by techniques from image processing and in particular the segmentation of local areas-of-interest in images. We employ the idea of a minimum spanning tree-based clustering, as introduced by Jana and Naik [18], to segment scatter plots into scatter plot motifs. In order to derive a motif's interestingness with respect to the entire data set, we apply the Bag-of-Words concept from information retrieval and the TF-IDF scheme [19]. Thus, we can derive the interestingness of a local scatter plot motif based on its frequency among and within the scatter plot corpus.

## III. Approach Overview

The main goal of our approach is to guide the analysts through the exploration process by providing scatter plots with interesting local motifs as starting point. The interest measure is derived from a local motif dictionary that distinguishes significant motifs by their frequency of occurrence in the data set. Figure 1 shows our iterative pipeline that consists of three steps to generate the dictionary: 1) automatic segmentation of scatter plot motifs, 2) motif feature extraction, and 3) feature-based clustering.

The automatic segmentation of scatter plots builds the basis of our interesting measure and hence requires special attention. Since each scatter plot may contain a distinct set of characteristics regarding its motifs (e.g., number, shape), its points (e.g., density) and the input scale of the dimensions, a flexible segmentation method is needed. Because our data sets under consideration contain possibly hundreds of scatter plots, a manual adjustment of segmentation parameters or the incorporation of domain knowledge in the segmentation process is often not feasible. Therefore, the segmentation technique must be parameter-free and capable of finding motifs regardless of their shape. Since most of the available segmentation techniques (e.g., DBSCAN or $k$-means) do not satisfy these requirements, we extended a minimum spanning tree (MST) based clustering technique. Another important prerequisite for the segmentation technique is to extract meaningful motifs that have a strong connection to the human perception. Experiments in [20] have shown that the MST method produces similar structures in the constellation of connection pairs of points (stars) as humans. The idea of the segmentation technique is to represent the data by means of a minimum spanning tree and iteratively remove the longest edge to derive an appropriate amount of clusters. Recent research on MST clustering has been conducted by Jana et al. in [18]. While their MST approach assesses the clustering quality after each edge removal by an internal validity criterion [21], we follow-up on their research by introducing an outlier-insensitive technique that focuses on larger clusters containing more than one point. Also

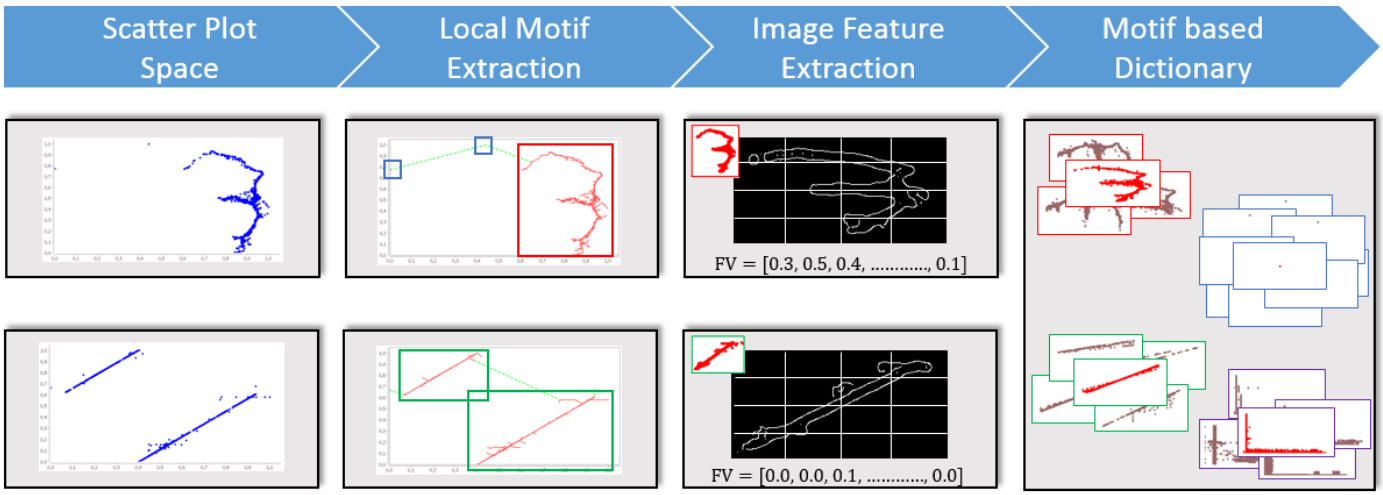


Fig. 1. Our proposed pipeline to generate a motif-based dictionary and to derive a local interestingness score. The first step is to extract the local motifs of each scatter plot by an adapted minimum spanning tree clustering approach. Next, we transform the motif shape into an image-based feature vector and, finally group all similar motifs by a $k$-means clustering to achieve the dictionary.

we reason on the impact of different quality criteria and their influence on the exploration process and the final result set in the following Section. The final scatter plot segmentation is achieved by considering the clustering with the overall best assessment.

As already mentioned, we attach great importance to the human perception of local motifs and therefore steer the interest measure based on interesting motifs. To do this, it is necessary to give a definition of interestingness for motifs. With regards to data exploration, we believe that besides complex and significant motif shapes also the frequency of occurrence plays a great role. Complex motifs attract the attention of humans during the exploration and may be less frequent –harder to find in large data– than common scatter plot patterns, such as negative or positive correlations. Furthermore, we assume that retrieving noise is not in the search scope of analysts and therefore should be penalized by our interest measure. In order to investigate these properties, we have to abstract the motif shapes and cluster similar motifs into a dictionary.

## IV. Global Interest-Measure Based on Local Motifs

This section provides a technical overview of our implementation for detecting interesting scatter plot segments and presents our aggregation scheme from local to global interestingness scores.

### A. Automatic Motif Segmentation in Scatter Plots

We present an enhanced, parameter-free minimum spanning tree based clustering method. A core part of the method is the assessment of the clustering quality, typically defined by simultaneously achieving a high intra-cluster and low inter-cluster similarity. Following up on the research of Jana and Naik [18], we conduct a performance evaluation with different well-known internal validity scores (*F-Ratio* [22], *Inter-Intra Ratio* [23], *Davies-Bouldin Index* [24], *Silhouette Coefficient* [25]) to qualitatively discuss the choice of an appropriate clustering quality measure. Each measurement was applied together with the MST clustering technique to eight distinct pre-classified, ground truth scatter plots[1]. In order to challenge the segmentation approach, we chose scatter plots that differ in the number of contained clusters, cluster shapes and cluster densities. The performance of the quality indices was measured by means of an external cluster evaluation measurement, the *Rand Index* [26] [MiB: (cf. result table in Appendix)] . In our experiments, the *F-Ratio* validity index outperformed the others with an average clustering accuracy of 89.3%. Due to this result, we employ the *F-Ratio* index as a clustering quality measure. Since the global clustering quality optimum would be inherently achieved after removing all edges from the MST and every point would constitute its own cluster we limit the number of overall clusters to $\lceil \sqrt{\frac{n}{2}} \rceil$ with $n$ being the number of points [27].

By limiting the number of clusters to this threshold, another problematic issue arises: The removal of an edge, which connects an outlying point to the MST, can lead to a cluster of size one. By considering each cluster regardless of its content, the threshold might be reached prematurely. One naïve way to avoid this problem would be to simply ignore cluster of size one. Although this would eliminate the case, non-outlier points with only one connection within the MST would be discarded as well. Thus, the point should be taken into consideration by reducing the clustering quality validity score. In order to distinguish outliers from cluster points, an outlier detection method is applied if the removal of an edge results in a cluster of size one. We rely on a distance-based outlier detection which considers the length of the last removed edge. If the edge length is high, compared to the edge lengths within an user-defined neighborhood area, then the outer vertex of the edge is marked as an outlier. Detected outliers are ignored in subsequent iterations of the algorithm.

[1]Shape sets. Collected by the 'School of Computing, University of Eastern Finland' (http://cs.joensuu.fi/sipu/datasets/)

Our final segmentation approach is depicted in Algorithm 1. As a result, we achieve a set of connected components ($|MST| = \#components$) for the initial minimum spanning tree. Each component thereby represents one cluster.

**Algorithm 1:** MinimumSpanningTreeSegmentation()

**Input**: $D = \{p_1, \ldots, p_n\}$ with $p_i \in \mathbb{R}^2$
**Result**: A set of connected components
$MST = constructMST(D)$
$bestMST = MST$
$k = \lceil \sqrt{\frac{n}{2}} \rceil$
**while** $|MST| \leq k$ **do**
  $e_{max} = longestEdge(MST)$
  **if** $hasOutlier(e_{max})$ **then**
    ignore outlier in all subsequent iterations.
  $MST = MST \setminus e_{max}$
  **if** $assess(MST)$ *better than* $assess(bestMST)$ **then**
    $bestMST = MST$
return $bestMST$

For the construction of the minimum spanning tree, several algorithms are available (e.g., Kruskal's [28] or Prim's algorithm [29]). The MST algorithms take a connected, undirected graph for which we suggest to use the graph resulting from a Delaunay triangulation. This preprocessing step has the advantage to minimize the memory consumption to a linear level, which would otherwise be quadratic for complete graphs. Since the Delaunay triangulation is a supergraph of the MST, no relevant information is lost.

### B. Dictionary-Based Interestingness Score

After we identified all sets of connected components (motifs), we group all similar motifs in the scatter plot space and build a similarity-based dictionary. The dictionary contains information about the distribution and frequency of motifs, and is used to determine the local interesting score. Therefore, the characteristics of each motif need to be considered and transfered into a uniform feature vector. To achieve this goal, we generate an image for each motif independent from their original scales and compute the feature vector based on image descriptors. In the end, each dictionary entry represents a set of similar motifs and a corresponding interestingness score can be computed individually.

Previous work [17] discovered that image-based feature approaches perform well to compare scatter plots. Hence, we compare all identified motifs by the point density distribution and their general shape derived through an edge detection approach. To generate the feature vector, we subdivide the local motif into a regular $16 \times 16$ grid and compute the grid's point density, and a histogram of edge orientations. In order to extract the grid-wise edge orientations, point clouds will be blurred by applying a Kernel-filter and converted into an edge image with the help of a Laplacian image filtering technique [30]. This guarantees that the descriptor merely considers the proper shape of a motif. By means of these *visual* features, we may cluster the motifs regardless of position or axes scales. The motif dictionary is formed by a $k$-means clustering on the feature vectors of all local motifs, as illustrated in the last step of Figure 1.

An essential step here is the parameter setting $k$ for the number of dictionary entries, since it influences the quality of the dictionary and consequently the local interestingness score. To determine an appropriate setting for $k$, we developed a visual exploration tool for experimental tests (cf. Section V-A), which displays the clustered motifs for different $k$ in a tabular and matrix representation. The local interestingness score expresses the uniqueness of a motif and how discriminant the motif is regarding the entire scatter plot space. Accordingly, scatter plots containing rare or conspicuous motifs are higher ranked and gain more attention for the global exploration in the scatter plot data.

$$MU_{score}(q) = \frac{1}{|\{m \in Dict[q]\}|} \tag{1}$$

Equation 1 shows the *Motif Uniqueness* score (MU) and how we measure the local interestingness for a given motif $q$. We divide one by the total number of similar motifs $m$ that were clustered to the dictionary entry of $q$.

### C. Global Interest Measure

The overall goal of our approach is to find interesting scatter plots for the exploration, containing discriminative local motifs. The global interest measure should reflect the interestingness of a given scatter plot based on the frequency of its local motifs in the entire scatter plot space. It is comparable to the text mining approach *tf-idf* [19], which uses the importance of a word to rank a document in a corpus. Instead of using the term frequency (*tf*), that computes the frequency of a term in a document, we use the *motif uniqueness* score from Section IV-B. It reflects how interesting and discriminant a motif is with respect to the corpus/scatter plot space. The basic idea of this local score is to weight frequent motifs (e.g., single dots or stripes) lower, and vice versa to weight discriminant motifs (e.g., complex patterns) higher.

The global interestingness measure is derived from these local factors in combination with an overall interestingness score. It corresponds to the inverse document frequency (*idf*) in text mining. The inverse document frequency is a measure to compute the overall importance of a term across all documents and follows the same idea as our second weighting factor that we call *inverse scatter plot frequency*. The difference to our approach is that we take the dictionary information and visual features into account and measure whether a motif is common or rare across all scatter plots. As shown in Equation 2, this score is obtained by dividing the total number of scatter plots $N$ by the number of scatter plots $sp$ containing one of the motifs in the dictionary cluster, and then taking the logarithm of that quotient. The substantial idea of this second weighting factor is to identify if a dictionary entry is based on many scatter plots containing such a motif, or e.g., just one scatter plot that contains many identical motifs.

$$ISPF_{score}(q) = \log \frac{N}{|\{sp \in Dict[q]\}|} \tag{2}$$

All local motif scores of a scatter plot are accumulated to produce the global interestingness score. Thus, scatter plots containing different and infrequent motifs achieve a higher score and are thereby considered as more interesting. Our proposed aggregation scheme for this interest measure is specified in Algorithm 2. For comparison reasons, we divide the aggregated global scatter plot interest score by the number of local motifs. Alternatively, analysts can use a range factor to prioritize the number of desired motifs and can penalize scatter plots containing more or less motifs. By means of this interest measure approach, we are able to extract automatically interesting scatter plots for exploration of large scatter plot spaces.

**Algorithm 2:** GlobalInterestMeasure()

**Input**: $motifDict, S$
**Result**: List of global interest measures
**foreach** *scatterplot in* $S(s_1, ...s_n)$ **do**
  $localMotifs =$ get motifs of $scatterplot$
  **foreach** $m$ *in* $localMotifs$ **do**
    $dictIndex =$ get dict index of $m$
    $localScore =$
    $MU(dictIndex) \cdot ISPF(dictIndex)$
    $globalScore$ += $localScore$
  $globalScore = globalScore/$size of $localMotifs$
  add $globalScore$ to $resultList$
return $resultList$

## V. Application of Motif-based Exploration

We now demonstrate the usefulness of our interest measure and the global scatter plot ranking by means of a visual exploration tool. First, we introduce the exploration interface and show how it supports the selection of an appropriate dictionary size. Then, we use a synthetic data set as a proof-of-concept to showcase our proposed interest measure. Finally, we make use of the interest measure on a real-world data set and analyze the suggested scatter plots for exploration.

### A. Visual Exploration: Identification of Similar Local Motifs

Selecting an appropriate dictionary size is difficult and has an impact on the subsequent process of finding similar local motifs and interesting global scatter plots. Especially for large and complex data, it is crucial to define a good cluster parameter $k$. Therefore, we developed a visual exploration tool to support analysts in the search process, find appropriate parameter settings, and finally suggest interesting scatter plots for exploration.

The tool involves a global overview in the form of a scatter plot matrix and a detailed dictionary view of all clustered motifs, as depicted in Figure 2. It allows the analysts to experiment with different image-based descriptors and clustering settings for a given data set. The dictionary view provides insights into the quality of the parameter setting and shows core information like cluster representatives and cluster size. The cluster size indicates the frequency of a particular representative motif in the scatter plot space. Moreover it hints on the practicability of the chosen clustering parameter $k$. To represent the cluster, we chose the local motif, which is the nearest neighbor to the center of mass in the feature vector space. By clicking on a dictionary entry, all cluster members will be highlighted in the linked scatter plot matrix. Conversely, it is possible to highlight all related motifs by clicking on a given motif in the scatter plot space. Moreover, we distinguish a multi-selection by using different color codings for each dictionary and motif respectively. Thus, users can quickly recognize the dictionary quality, as well as the motif distribution. [MiB: CHECK next sentence] A further benefit of this overview is that users can estimate whether the cluster extraction threshold (c.f. Section xxx) is configured appropriately.

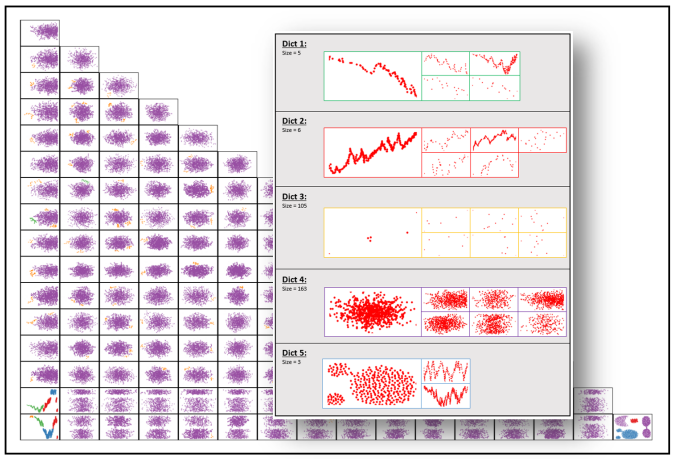


Fig. 2. Scatter plot matrix overview of the synthetic data set and the resulting dictionary with five entries. The first 15 dimensions consist of Gaussian clusters and the last rows are combinations with the *Aggregation* data set. By means of the displayed dictionary view and the corresponding color coding, analysts can easily determine a good dictionary size for a given data set.

### B. Synthetic Data: Interestingness Measure

We created a synthetic data set by merging 15-dimensional Gaussian clusters with the two-dimensional *Aggregation* data set presented in [31]. Since, the aggregation data set consists of a small sample size (788 record), we randomly created Gaussian clusters with the same size and merged the data, as illustrated in the background of Figure 2. The original scatter plot of the aggregation data set is located at the bottom right corner of the matrix. The experiment was designed to depict that motifs of the Gaussian dimensions (purple motifs), which appear more often will also result in a low local and overall interestingness score. In contrast, scatter plots that were merged with one of the aggregation data dimensions (last two rows) contain more complex and outstanding motifs, and will thus be rated more interesting. [MiB: Brauchst du den nchsten satz?] Furthermore, the aggregation data set contains several significant motifs that we want to detect by our segmentation approach and rank for global exploration.

The first step of our approach is to determine the interesting motifs by running our adapted MST segmentation approach (see Section IV-A). After the segmentation step, we have extracted 282 local motifs of 136 scatter plots. When looking at the scatter plot matrix, we can see that the data set contains only a few kinds of motifs. In this case, we recommend choosing a small $k$ (e.g., between three and five) to keep the quality of the dictionary high and clearly separate the different

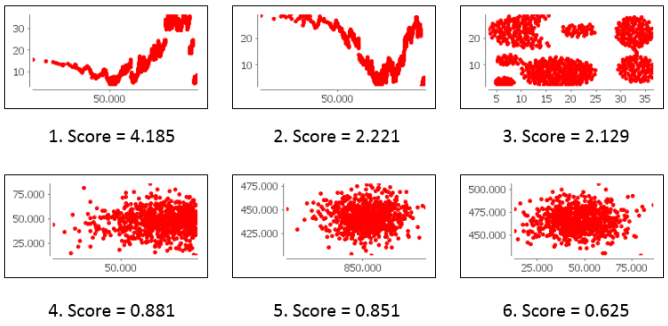


Fig. 3. The six most interesting scatter plots of the synthetic data set for analyzing local motifs. The global scores are obtained by aggregating all local motif scores ($MU \times ISPF$) of a given scatter plot.

motif shapes. Choosing a too large dictionary size would lead to splitting up the homogeneous motifs of the Gaussian clusters into several dictionary entries and thus will distort the local interestingness score. On the contrary, a too small dictionary size will merge dissimilar motifs and also negatively influence the ranking.

In the experiment, depicted in Figure 2, we found a good dictionary setting by using the combined image descriptor, which takes the edge direction and density of a motif into account (c.f. Section IV-B) and chose a dictionary size of five. Thus, we received a dictionary with five well-separated clusters, containing a negative trend motifs (green), positive trend motifs (red), sparse point clouds (orange), dense point clouds (purple) and a motif cluster with wide-spread distributions (blue). The largest dictionary entry is the cluster represented by the purple color with 163 similar motifs, followed by the orange cluster with 105 motifs. As one can see, all motifs are highly similar except those from the original aggregation scatter plot and the two scatter plots in combination with the first dimension. Consequently, our interest measure ranks the scatter plots with the purple and orange motifs less interesting than the other motif groups. A result overview of the highest ranked six scatter plots is shown in Figure 3. As expected, the three scatter plots containing deviant motifs achieve the highest global scores. The reason why the first three scatter plots received significantly higher scores is due to their higher ranked local motifs. The original aggregation scatter plot has been ranked third after the two line shaped scatter plots, because of the poor local score of the two purple motifs. Very unexpected was the incident that all other scatter plots derived from the two aggregation data dimensions are not in the top six results. This can be explained by the fact that scatter plots on rank four to six also contain the higher ranked motifs of the orange and green dictionary entry.

### C. Real-World Data: Interestingness Analysis

The second evaluation data set is retrieved from the *eurostat*[2] data repository. The data repository provides in total 5500 data sets each containing information about a European related topic, such as economy, population and industry. We extracted a data set containing 27 statistical attributes from 28 EU countries that show temporal changes over the last decades.

[2]Statistical Office of the European Union (http://ec.europa.eu/eurostat). Accessed 05/2015.

We created a scatter plot matrix with 351 unique scatter plots from these 27 dimensions in which each data instance (point) represents one country at a specific year. The corresponding scatter plot matrix is illustrated in Figure 4.

As in the previous example, we start the interestingness search process with segmenting the scatter plot space into local motifs and select a good setting for the dictionary. Our segmentation approach returned 1549 local motifs of the 351 scatter plots. We are using the combined image descriptor for characterizing the motifs in this experiment. As dictionary size, we found that appropriate results were achieved by using a size between 10 and 15. We iteratively highlighted the most considerable motifs in the scatter plot matrix to identify the similarity of a dictionary entry and thus proof the quality of the settings. Finally, we decided to choose a dictionary size of 11 for further analysis.
As Figure 4 depicts, one can clearly recognize the dependencies between similar motifs and the dimensions in the scatter plot matrix. For instance, if we consider the dictionary entry seven (strips colored in brown), we are able to identify all the dictionary items in column two, four and five. The same applies to the orange motif class with sparse negative trend direction (dictionary cluster ID 10), which are mostly located in row 16 and 18. Finding such properties in the scatter plot matrix may lead to first insights in the local motif analysis.

The top ranked scatter plots of our chosen setting are outlined at the bottom left corner of our visual exploration tool. An enlarged excerpt of the best six rankings is also shown in Figure 4. [MiB: Willst du diesen satz wirklich drinne haben!?!?] At this point, we would like to stress again that these results are intended to give a guidance for interesting scatter plot exploration containing significant local motifs. On closer inspection, we can see that all suggested scatter plots contain significant motifs, which may be interesting to analyze. As an analysis example, we want to focus on the scatter plot ranked on the third place. The scatter plot shows separated motifs with several positive trend directions shifted on both axes. [MiB: Bitte neue feature beschreibungen finden!] These motifs describe the relation between the average duration of working life against the average age of mother at birth of the population in all EU countries. [MiB: Neu formulieren... Das ist deutsch...] It becomes clear that the total work duration of womens decreases when they become a mother earlier and thus can not longer work. Additionally, it would be interesting to analyze these different groupings in relation to other non-numerical attributes, such as geolocation and see which countries share similar characteristics and how they change over time.

## VI. Conclusion

We introduced a novel pipeline approach in which we analyze the interestingness of automatically extracted local motifs to guide the exploration in scatter plot data. Therefore, we enhanced a minimum spanning tree-based clustering technique for a non-parametric segmentation in order to derive local areas-of-interest. To assess the overall interestingness, we adapted the Bag-of-Words concept and the TF-IDF scheme from information retrieval to the domain of scatter plot motifs. We derive the interestingness of local scatter plot motifs based on its occurrence among and within the scatter plot space.

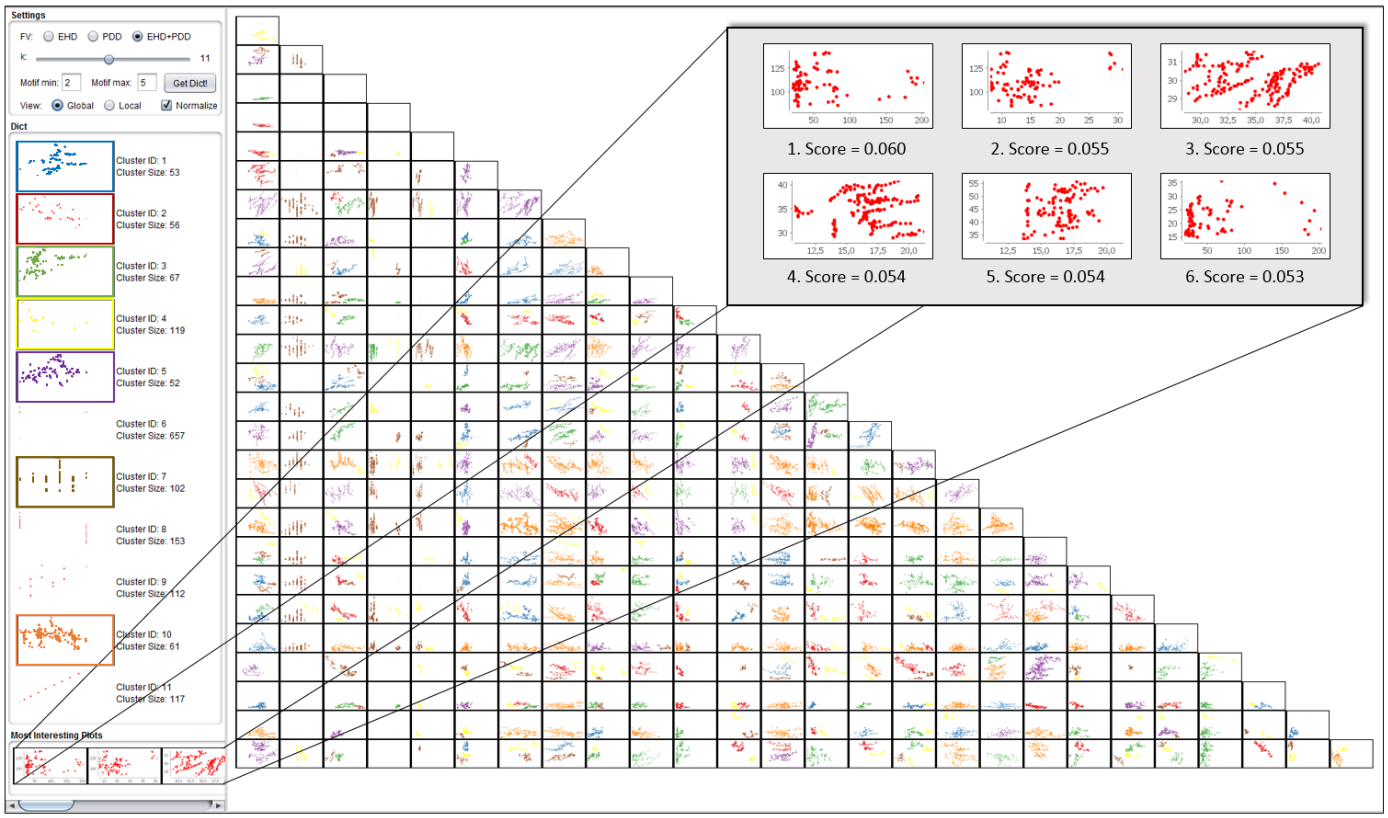


Fig. 4. Our visual exploration tool for global and local scatter plot analysis. By means of this tool, analysts can derive different motif based dictionaries by adjusting the parameter settings and thus achieve various interesting scatter plots suggestions for exploration. The parameter settings, dictionary view and the resulting global interest ranking are located on the left hand side. Local motifs of a given setting can be highlighted in the scatter plot matrix to assess the dictionary quality.

Furthermore, we developed an interactive visual exploration tool with brushing and linking that supports analysts to find appropriate motif dictionaries and suggests interesting scatter plots for exploration. Finally, we applied the pipeline on a synthetic and real-world data set to demonstrate how it can efficiently lead to interesting discoveries of local motifs.

While our technique has proven useful, we have identified several areas where improvements or alternatives can be explored. Firstly, we want to research alternative scatter plot segmentation techniques. Generally, two fundamental approaches have to be distinguished: First, data space segmentation approaches and, second, image-based segmentation approaches. While implemented approach is based on partitioning distributions in the data space, other data space segmentation approaches are possible. For example, a regression tree can be learned for finding a (non-)linear partitioning. Alternatively, a wide range of options from the image analysis community are available and not yet explored on scatter plots. We are planing to experiment with convex hull calculations on the rendered scatter plot images to find local motifs. One advantage of this approach is that data space axis ranges are normalized by definition, such that their treatment is not important anymore.

Secondly, we are planning to integrate further motif descriptors in our system. While currently, an edge-descriptor and a density-based descriptor proved useful for our case studies, we expect that a greater variety of scatter plot motifs can be described with other feature descriptors. Therefore, we are planning to incorporate a Hough descriptor (line-detection), as an alternative to the edge-detection descriptor. Other than that, regressional features, such as described in [16], can be integrated. These descriptors will bring an advanced semantics level in the motif detection, which needs to be understood and researched more extensively.

Lastly, we want to improve the visual exploration tool by including additional non-numerical attributes, such as geolocation or textual data, to gain further information about the extracted motifs. Thus, for instance, it could be examined why certain motifs differ in their shapes or are translated over the scatter plot axes. Another idea would be to extend other existing visualization techniques for motif analysis. The traditional parallel coordinate plot could be extended with visual motif axes to analyze the extracted motifs along other numerical or categorical axes.

## References

[1] M. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, MA, USA: A. K. Peters, Ltd., 2010.

[2] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan, "Selecting good views of high-dimensional data using class consistency," in *Computer Graphics Forum*, vol. 28, no. 3. Wiley Online Library, 2009, pp. 831–838.

[3] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim, "Automated analytical methods to support visual explo-

ration of high-dimensional data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 5, pp. 584–597, 2011.

[4] W. Cleveland, "The shape parameter of a two-variable graph," *Journal of the American Statistical Association*, vol. 83, no. 402, pp. 289–300, 1988.

[5] J. Talbot, J. Gerth, and P. Hanrahan, "Arc length-based aspect ratio selection." *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2276–82, Dec. 2011.

[6] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff, "Selecting the aspect ratio of a scatter plot based on its delaunay triangulation," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2326–35, dec 2013.

[7] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," *Computer Graphics Forum (Proc. EuroVis 2012)*, vol. 31(3), pp. 1335–1344, 2012.

[8] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *In Proceedings of the IEEE Symposium on Information Visualization*, Oct 2005, pp. 157–164.

[9] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan, "Selecting good views of high-dimensional data using class consistency," *Computer Graphics Forum (Proc. EuroVis 2009)*, vol. 28, no. 3, 2009.

[10] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, and H. Theisel, "Selecting coherent and relevant plots in large scatterplot matrices," *Computer Graphics Forum*, vol. 31, no. 6, pp. 1895–1908, Apr. 2012.

[11] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck, "Assisted descriptor selection based on visual comparative data analysis," *Wiley-Blackwell Computer Graphics Forum*, vol. 30, no. 3, pp. 891–900, 2011, proceedings of Eurographics/IEEE-VGTC Symposium on Visualization 2011).

[12] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. A. Magnor, "Quality-based visualization matrices." in *VMV*, 2009, pp. 341–350.

[13] A. Anand, L. Wilkinson, and D. T. Nhon, "Visual pattern discovery using random projections," in *IEEE VAST*, 2012, pp. 43–52.

[14] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, and D. A. Keim, "Subspace Search and Visualization to Make Sense of Alternative Clusterings in High-Dimensional Data," in *Procedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*. IEEE CS Press, 2012, pp. 63–72.

[15] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 2008)*, vol. 14, no. 6, pp. 1141–1148, 2008.

[16] M. Scherer, J. Bernard, and T. Schreck, "Retrieval and exploratory search in multivariate research data repositories using regressional features," in *Proc. ACM/IEEE Joint Conference on Digital Libraries*, 2011, pp. 363–372.

[17] M. Scherer, T. von Landesberger, and T. Schreck, "A Benchmark for Content-Based Retrieval in Bivariate Data Collections," in *Proc. Int. Conference on Theory and Practice of Digital Libraries*, 2012.

[18] P. Jana and A. Naik, "An efficient minimum spanning tree based clustering algorithm," in *Proc. Int. Conference on Methods and Models in Computer Science*, 2009.

[19] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[20] M. Dry, D. Navarro, A. Preiss, and M. Lee, "The perceptual organization of point constellations," 2009.

[21] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. ICDM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 911–916.

[22] L. A. Garcia-Escudero, A. Gordaliza, A. Mayo-Iscar, and C. Matran, "A robust maximal f-ratio statistic to detect clusters structure," in *Communications in Statistics - Theory and Methods*, 2009, pp. 682–694.

[23] S. Ray and R. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," in *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*. New Delhi, India: Narosa Publishing House, 1999, pp. 137–143.

[24] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-1, no. 2, pp. 224–227, April 1979.

[25] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987.

[26] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[27] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press; 1 edition (January 27, 1976), 1976.

[28] J. Kruskal, "On the shortest spanning subtree and the traveling salesman problem," in *Proceedings of the American Mathematical Society.*, 1956, pp. 45–50.

[29] R. C. Prim, "Shortest connection networks and some generalisations," *Bell System Technical Journal*, vol. 36, pp. 1389–1401, 1957.

[30] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proceedings of the 2000 ACM workshops on Multimedia*. New York, NY, USA: ACM, 2000, pp. 51–54.

[31] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 4–es, Mar. 2007.