Volume xx (200y), Number z, pp. 1–12

Selecting Coherent and Relevant Plots in Large Scatterplot Matrices

Dirk J. Lehmann[1], Georgia Albuquerque[2], Martin Eisemann[2], Marcus Magnor[2], and Holger Theisel[1]

[1]Department of Simulation and Graphics, University of Magdeburg, Germany
[2]Computer Graphics Lab, TU Braunschweig, Germany

Abstract

The scatterplot matrix (SPLOM) is a well-established technique to visually explore high-dimensional data sets. It is characterized by the number of scatterplots (plots) of which it consists of. Unfortunately, this number quadratically grows with the number of the data set's dimensions. Thus, a SPLOM scales very poorly. Consequently, the usefulness of SPLOMs is restricted to a small number of dimensions. For this, several approaches already exist to explore such "small" SPLOMs. Those approaches address the scalability problem just indirectly and without solving it. Therefore, we introduce a new greedy approach to manage "large" SPLOMs with more than hundred dimensions. We establish a combined visualization and interaction scheme that produces intuitively interpretable SPLOMs by combining known quality measures, a pre-process reordering, and a perception-based abstraction. With this scheme, the user can interactively find large amounts of relevant plots in large SPLOMs.

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Information Search and Retrieval]: Information filtering— I.5.0 [Pattern Recognition]: Structural— I.3.8 [Computer Graphics]: Applications—

1. Introduction

High-dimensional data sets arise in various real situations, e.g., as a result of a physical measurement. In order to analyze them, the user requires to visualize the complete data set at once to get visual access to the underlying data. There is a variety of visualization approaches that aim at this problem, e.g., *Parallel Coordinates* [Ins85,Ins09], *RadViz* [HGM*97], or *Scatterplot Matrices* [CLN86, Cle93]. They all have different advantages/disadvantages. Thus, a convenient technique depends on the visualization goals of the user. However, all techniques have one common disadvantage: they scale poorly.

In this paper, we focus on the scalability problem in terms of the scatterplot matrix (*SPLOM*), which is a very powerful, intuitive, and well-established technique: A scatterplot (*plot*) is a discrete bivariate visualization of pairwise data dimensions generated by a projection from the data set's population onto an arbitrary plane. A SPLOM is a matrix of all such plots, and it gives a complete overview of the data. A data set with n dimensions links to an $n \times n$ symmetric SPLOM S with $S(i,j) = S(j,i); i,j = 1,\ldots n$. Therefore, it is sufficient to represent it either as an upper-triangular or a lower-triangular matrix. SPLOMs are widely used, and they are very popular due to many advantages:

- they can be easily combined with other visualization and interaction techniques, like linking & brushing [BC87],
- they are simple to evaluate, e.g., w.r.t. bivariate correlations, classifications, clusters, or trends,
- the experienced user is able to form hypotheses about multivariate relations between different dimensions of the underlying data set, and
- they are simple to implement, intuitively interpretable and also appropriate for unexperienced users.

Consequently, SPLOMs are a convenient access point for new users to enter visual analytics. Due to the advantages, the users/analysts would highly appreciate if SPLOMs could be used in any situation. Unfortunately, this is not the case because the number d of plots grows quadratically with the number n of the dimensions: $d = \frac{1}{2}(n^2 - n)$. Due to this scalability problem, the user is not able to evaluate the huge amount of plots, to detect relevant plots, or to examine all plots at all. Thus, in practice a SPLOM is only adequate if the number of dimensions is rather small. From our experience, we know that about 40 dimensions are the limit for

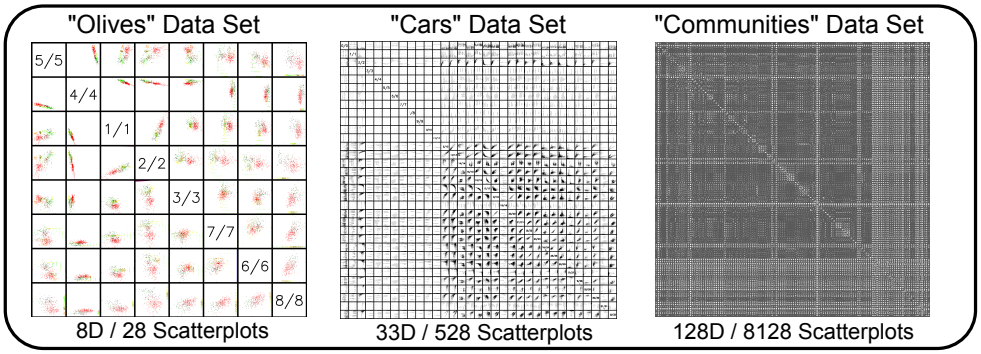

Figure 1: *Scalability Problem: (left to right) SPLOM with 8, 33 and 128 dimensions from different data sets. An increasing number of dimensions produces visualizations which are hard to interpret.*

their usefulness. To illustrate this effect, Figure 1 shows several SPLOMs: the larger the number of dimensions gets, the more difficult it becomes to identify relevant plots. This scalability problem is already known and there exist different strategies to deal with it:

- A subset of dimensions (e.g., with a large entropy) can be preselected in order to downsize the SPLOM. This provides the advantage that the SPLOM becomes smaller and the disadvantage that important information might be lost.
- Different navigation strategies can be used to reach all plots successively over time, which provides the advantage that a user views all plots and the disadvantage that a long time passes by doing so. This way, associations are not recognizable anymore.

However, these strategies are developed for SPLOMs with a dimensionality much smaller than hundred dimensions, and they do not scale. Therefore, with this paper we address the scalability problem for SPLOMs in terms of data sets larger than one hundred dimensions.

At first, we introduce five *selection criteria* that should be satisfied by an approach for selecting plots from large SPLOMs.

1. Scalability (SC): The approach has to scale with the number of dimensions.
2. Coherence of Plots (CC): The selected plots need to be directly comparable with each other, i.e., they have to be coherent w.r.t. their pairwise dimensions.
3. Explorative Analysis (EC): The approach should be appropriate to support an analysis without the need of an a priori knowledge of the data set.
4. Number of Plots (NC): The user has to be able to steer the number of selected plots that he/she wants to simultaneously analyze, in order to avoid to be overburdened.
5. Relevance of Plots (RC): The user has to be able to select only those plots that are relevant for a particular visualization goal.

This paper introduces an approach for the selection of plots within a large SPLOM that satisfies these criteria. To accomplish this goal, we combine a visualization scheme with an interaction scheme: At first, our approach generates an abstract and intuitively interpretable SPLOM (*A-SPLOM*) based on both a reordering technique and the ordinary SPLOM, and supported by quality measures that measure the relevance of a plot. Subsequently, several of those A-SPLOMs are integrated within an interaction scheme to select, to view, and to analyze relevant plots, without losing the context in terms of the whole SPLOM. In detail, our contributions are:

- We introduce an approach to abstract large SPLOMs in order to reduce their visual complexity.
- We present an interactive framework.
- We test our concept for real data sets with more than one hundred dimensions.

2. Related Work: Background and Issues

In order to analyze high-dimensional data sets with a large number of dimensions, appropriate navigation strategies can be classified into *Hierarchical Navigation* and *Data Space Navigation*. These approaches also occur combined.

Hierarchical Navigation A popular approach of hierarchy dimension ordering, spacing and filtering was presented by Yang et al. [YPWR03]. This approach forms a hierarchical tree of dimensions, steered by a threshold w.r.t. a particular clustering measure. The tree is formed during a successive clustering which groups dimensions together. In addition, there exist strategies for navigation within the tree and for filtering in order to reduce the visual clutter. The height of the tree grows with the number of dimensions. In contrast, our A-SPLOMs (cf. Section 4) look similar, independent of the number of dimensions. Furthermore, a cluster of the hierarchy level i is not that easily comparable with a cluster of hierarchy level $i \pm t$. Therefore, we do not further focus on hierarchical approaches within this paper.

Data Space Navigation Two-dimensional projections have been widely used as a starting point in the visual exploration of high-dimensional data sets. To address the scalability problem, Asimov proposed the Grand Tour [Asi85]

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

as a navigation method that provides a complete overview of the data set by generating sequences of two-dimensional projections. Different approaches for ranking/selecting the best low-dimensional projections were further proposed. One is Projection Pursuit [FT74], a statistical technique to search for low-dimensional projections that expose interesting structures of the data sets. Different Projection Pursuit indices [FFT75, Hub85] as well as a combination of the Grand Tour and Projection Pursuit [CBCH95] as visual exploration systems were developed. However, within the class of Data Space Navigation, the approaches that deal with SPLOM-based navigation form an own subclass. Due to the importance for our work, we subsequently treat this subclass.

SPLOM-based Navigation An approach to interactive exploration was proposed as "Rolling the Dice" [EDF08]. An interactive visual exploration that uses SPLOMs, 3D transitions between the plots, and dimension reordering of the matrix. The Parallel Scatterplot Matrix was introduced in [VMCJ10]. Here, the ideas of parallel coordinates and SPLOM are combined to a hybrid approach of rotating from plots to the corresponding parallel coordinates and vice versa. The drawback of such approaches is that they do not scale well when the number of dimensions grows.

Some alternative navigation strategies of SPLOMs are based on pattern recognition in order to detect relevant patterns. The relevance itself is then expressed as a real number. Such *quality measures* work similarly to the human perception system, as shown empirically, e.g., in [TBB*10]. The trivial approach for a quality measure-based navigation is thresholding. However, until now there is no practical approach due to the following reasons: only the value of relevance can be steered. From this it follows that the number of the resulting plots is not directly controllable and might be unmanageably large, and the coherence of the plots is not controllable, too. Thus, thresholding is probably not a practical approach.

The possibility of combining quality measures and dimension reduction has been investigated, e.g., by Johansson & Johansson [JJ09]: their approach reduces the number of dimensions in order to emphasize relevant ones. This is done via a user-based interactive process through combining and weighting several quality measures. The general problem of dimension reduction is the loss of information. Without the user's knowledge of important information it is difficult to decide which dimensions should be removed. Thus, this approach is appropriate rather for a confirmative data analysis instead of an explorative analysis.

The Scagnostics method [TT85, WAG05] was presented as an alternative to Projection Pursuit. The main idea is to compute different scagnostics indices (e.g. Convexity, Skinny, etc.) for each plot based on quality measures, and to show the results in a SPLOM of the indices, called Scagnostics SPLOM, which creates a high level of abstraction so that the original dimensions are not represented anymore. We use a lower level of abstraction to support the understandability of the user. Our method still resembles the original SPLOM, i.e., the dimensions are represented on the rows and columns of the matrix, preventing any occlusion artifacts that might appear in the Scagnostics representation.

Several SPLOM reordering approaches are already known: the Rank-by-Feature Framework [SS05] proposes a method to rank all dimensions according to a selected criterion. The dimensions are reordered with the aid of this criterion and presented to the user. Reordering a SPLOM based on the dimensions is equal to finding an appropriate sequence of permutations. This permutation problem is NP-hard. Thus, the Rank-by-Feature Framework is just appropriate for small SPLOMs where all permutations are testable.

An information-aware reordering algorithm for visualization matrices was proposed by [AEL*09] to overcome this permutation problem: the dimensions are sorted with the aid of a quality measure, and they are presented such that the best plots are concentrated on the upper left corner of the SPLOM. The weak point of this approach is that for a large number of dimensions it is not always possible to cluster the best plots together. To overcome this problem, our reordering method concentrates on the best plots in different regions, based on the similarity of the dimensions.

Peng et al. [PWR04] presented another reordering approach for SPLOMs aiming at clutter reduction. They define a clutter measure for a permutation of a SPLOM and tested which of the permutations would minimize it. The algorithm's complexity is $O(n^2 \cdot n!)$, whereas n is the number of dimensions. Thus, the approach is also not practical for large SPLOMS. Regarding this, Keim [Kei] already described that reordering problems in general seem to be complex optimizations problems which are NP-complete and are therefore only solvable by using appropriate heuristics, e.g., the Kohonen map [Koh90]. Thus, our algorithm avoids a complete search for the benefit to find at least one good (greedy) solution. Consequently, our reordering approach extends the described approaches with the difference that our approach runs with a complexity of $O(n^3)$. In addition, we introduce a navigation metaphor and tests with more than 100 dimensions.

Figure 2 shows a subjective comparison[†] of the selection criteria in terms of the SPLOM-based related work: there is currently no approach that completely fulfills all criteria. To conclude, strategies either navigate through levels of a hierarchy or reduce the number of plots due to the scalability problem. In our paper, we combine and advance both concepts to reach the goal of purposefully selecting relevant plots in large SPLOMs.

One possibility was already mentioned, e.g., with the

[†] The comparison is based on the author's experience.

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

Approach	uses Quality Measure	uses Reordering	SC	CC	EC	NC	RC
Rolling the Dice[EDF08]	No	Yes	−	−	+	−	−
Parallel Scatterplots Matrix[VMCJ10]	No	No	−	−	+	−	−
Thresholding	Yes	No	−	−	+	−	++
Interactive Dimension Reduction[JJ09]	Yes	No	+	+	−	−	+
Scagnostics[TT85, WAG05]	Yes	No	−	+	+	−	+
Rank-by-Feature[SS05]	Yes	Yes	−	+	+	−	+
Information Aware[AEL*09]	Yes	Yes	−	+	+	−	+
Clutter Reduction[PWR04]	Yes	Yes	−	+	−	−	+

CC...Coherence SC...Scalability EC...Explorative Analysis ++ very good + good − bad
NC...Number of plots controllable RC...Relevance of plots controllable

Figure 2: *Selection criteria for the SPLOM-based (and directly related) work.*

Scagnostics Framework, namely the quality measures. Therefore, for our approach the idea beyond the quality measures is important, subsequently explained in more detail.

3. Introduction of Quality Measures

A high-dimensional data set D encodes m properties $A_i(D), i = 1,\ldots,m$ as correlations or clusters by its attributes. Additionally, a two-dimensional visualization v_j of the data set D is generated by an operator Φ_j (,e.g., a projection): $v_j = \Phi_j(D)$. A specific property $A_i(D)$ of the data set D partially remains within the visualization v_j as a property $A_i'(v_j)$ of it (dependent on $\Phi_j(D)$). Thus, the existence of property $A_i'(v_j)$ implies the possibility of the existence of a property $A_i(D)$. Consequently, the data set can be analyzed by analyzing its visualizations.

This is the main idea of quality measures: instead of analyzing a data set itself, the visualizations of it are analyzed. The advantage is that only lower-dimensional projections need to be handled, well-known image process methods can be applied, and the human perception process can be mimicked. The definition of a quality measure follows directly from this comprehensive concept: a quality measure $Q_{A_i'}(v_j) \in \mathbb{R}$ of a visualization v_j appraises a property $A_i'(v_j)$ via a real number or a scalar value, respectively. For example, a correlation (that is such a property A_i') within visualization v_j can be more distinct than in visualization v_k. Then, $Q_{A_i'}(v_j)$ is larger than $Q_{A_i'}(v_k)$. A lot of such quality measures have been introduced within the last years, e.g., [TT85], [WAG05], [SSK06], [TAE*09], [SNLH09], or [AEL*10]. With the aid of those quality measures, we now present a concept for the efficient selection of plots.

4. Approach

This section explains our approach in detail. In order to handle large SPLOMs, our concept is separated into two parts. Figure 3 shows a schematic overview: At first, our approach maps a SPLOM onto a visually abstracted representation (A-SPLOM). Then, different A-SPLOMs are generated based on certain quality measures. Such A-SPLOMs support the user in the interaction part of our concept in order to interactively select large amounts of relevant plots without losing the context.

4.1. Visualization

Within the first step, the plots $v_j; j = 1,\ldots,d$ are measured by a quality measure $Q_{A_i'}(v_j)$: the larger the calculated quality measure value is, the more relevant is a plot. The second step is to reorder this measured SPLOM so that clusters are formed maximizing the likelihood to find plots of similar relevance. Finally, the last step abstracts the reordered SPLOM so that the user is able to recognize regions with a large likelihood of relevant plots.

4.1.1. Step 1: Measuring the Relevance

In the first step, each plot $v_{i,j}$ $(i, j = 1,\ldots,n)$ of our SPLOM $\mathbf{S}$ is mapped to a real number with the aid of a quality measure $Q_{A_i'}(v_{i,j})$. Hence, we obtain a symmetric $n \times n$ matrix:

$$Q_{A_i'}(\mathbf{S})_{n\times n} = \begin{pmatrix} Q_{A_i'}(v_{1,1}) & \cdots & Q_{A_i'}(v_{1,n}) \\ \vdots & \ddots & \vdots \\ Q_{A_i'}(v_{n,1}) & \cdots & Q_{A_i'}(v_{n,n}) \end{pmatrix}.$$

We need to describe the terms *visualization goal*, *relevance* and *property* in detail because of their different meaning: a visualization goal of a visualization means that this visualization should visualize a certain property that a user wants to investigate. If this property A_i' might be measured by a quality measure $Q(A_i')$ the measured value exactly equates to the relevance of the visualization. We assume that a quality measure in terms of a visualization goal is available and do not further focus on this discussion. Thus, the measured SPLOM $Q_{A_i'}(\mathbf{S})$ describes the relevance of the visualizations.

4.1.2. Step 2: Reordering

The relevances of the plots within the measured SPLOM $Q_{A_i'}(\mathbf{S})$ are randomly distributed and form a filigree texture, as Figure 4 (a) demonstrates. Thus, a reordering as shown in Figure 4 (b) is helpful and will be explained in the following.

The filigree texture makes it difficult to find relevant plots visually because the human visual system is subjected to several optical illusions that apply especially to filigree textures, like an afterimage [Ant78], the Chubb effect [CSS89] or simultaneous contrast [Bur]. An afterimage is a virtual image that occurs after a user's view changes from one part of the SPLOM to another part. The Chubb effect describes that the visual perception of textures is influenced by the neighboring textures. A similar effect is the simultaneous contrast, which describes that different adjacent colors influence each other with respect to their perception. Altogether, the optical illusions of SPLOM's filigree textures mislead the user. To avoid this, it is advisable to abstract and to coarsen the SPLOM, which our approach does. Furthermore, it is not trivial to define convenient interaction methods to select randomly distributed plots reasonably. Due to all these reasons, the approach reorders the measured SPLOM with the goal to cluster plots with similar relevance together. To

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

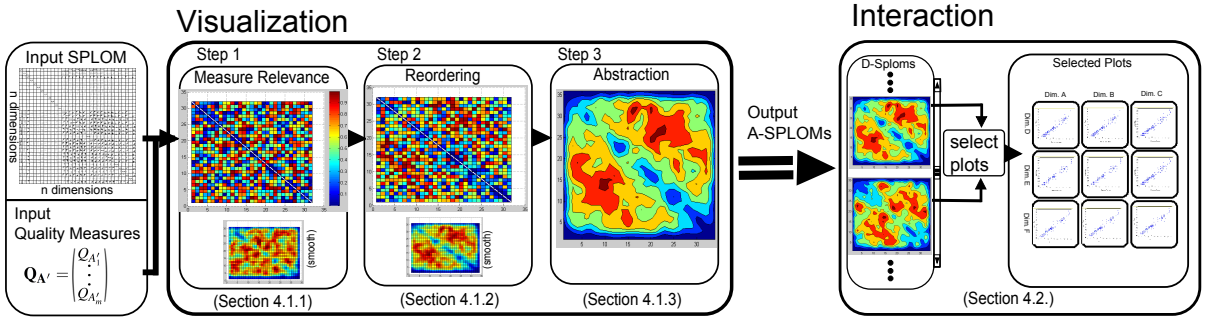

Figure 3: *Schematic overview of the approach in order to select relevant plots in large SPLOMs.*

avoid losing the associations between the dimensions, we are restricted to transposing entire rows/columns. Thus, reordering the matrix $\Omega = Q_{A'_i}(\mathbf{S})$ is the same problem as looking for the permutation matrix $\mathbf{P}$ that minimizes an appropriate permutation measure $\lambda(\Omega)$, i.e. finding $\text{argmin}_{\lambda}(\lambda(\Omega \cdot \mathbf{P}))$. Regarding this, the measures that are certainly not suitable

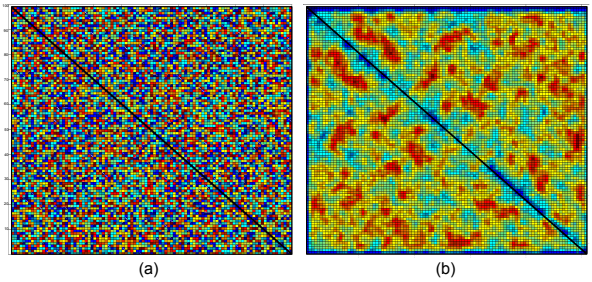

Figure 4: *Effect of the Reordering: (a) SPLOM before the reordering. The color-coded relevances are approximately uniformly distributed and form a filigree texture, which might cause optical illusions. (b) SPLOM after the reordering. Clusters of similar relevance can be perceived simply.*

are measures that are invariant to permutation, as, e.g. the entropy, or the signal-to-noise ratio. Therefore, it is reasonable to measure the matrix locally for regions of similar relevance. Considering a $w \times w$ neighborhood (window size) within the matrix gives us the required locality, where w is an odd integer. We tested three (local) measures for the matrix Ω, which measure the cumulative difference within a certain neighborhood given by:

$$\lambda_s = \sum_{x,y=1}^{n} \underbrace{\sqrt{\sum_{i,j=-w/2}^{w/2} \big(\Omega(x+i,y+j) - \mu(x,y)\big)^2}}_{\text{local standard deviation}};$$

$$\mu(x,y) = \frac{1}{w^2} \sum_{i,j=-w/2}^{w/2} \Omega(x+i,y+j),$$

$$\lambda_d = \sum_{x,y=1}^{n} \underbrace{\sum_{i,j=-w/2 \,\wedge\, i \neq j}^{w/2} (\Omega(x,y) - \Omega(x+i,y+j))^2}_{\text{local square differences}},$$

$$\lambda_r = \sum_{x,y=1}^{n} \underbrace{|\max_{w\times w}(\Omega(x,y)) - \min_{w\times w}(\Omega(x,y))|}_{\text{local min-max differences}};$$

$$\max_{w\times w}(x,y) = \max_{\substack{-w/2 \le i \le w/2 \\ -w/2 \le j \le w/2}} \{\Omega(x+i,y+j)\},$$

$$\min_{w\times w}(x,y) = \min_{\substack{-w/2 \le i \le w/2 \\ -w/2 \le j \le w/2}} \{\Omega(x+i,y+j)\}.$$

All measures calculate a certain difference property: the local standard deviation λ_s, the local squared differences λ_d, and the difference between the minimum and the maximum value λ_r. To handle bordering issues, the borders of the matrix are connected to each other periodically.

Furthermore, the approach requires a sorting strategy. Note that finding an optimal permutation $\mathbf{P}$ for the matrix Ω with n dimensions is an NP-hard problem. For instance, even a small matrix that represents 30 dimensions would cause $30! \approx 265 \cdot 10^{30}$ different permutations. Thus, a brute-force approach that tests all combinations is not feasible.

This permutation problem is well known, and we have to rely on a heuristic optimization algorithm. Our algorithm applies a pairwise transpose of all dimensions to detect which pairs of dimensions require transposition to minimize the measure λ at best. This step is repeated k-times until a minimum is reached. For this, we use a Hill-Climbing algorithm. The pseudocode is given in Figure 5. If the goal function has a (global) monotonic behavior then the worst case complexity is $\mathcal{O}(n!)$. If the initial permutation already yields the best measure with respect to all pairwise permutations, we obtain the best case complexity with $\Omega(n^2)$; else, the algorithm needs k iterations which yield a complexity of $O(n^2 \cdot k)$. However, from our experiences, $k \le n$ applies mostly. Thus, the complexity is $O(n^3)$. Nevertheless, in order to get a trade-off between the reorder improvement and the run time, the algorithm stops at the latest after $k \ge 2 \cdot n$ iterations. Note that in our experiments the algorithm has never been stopped caused by this condition, since an optimum has always been reached before.

The start permutation influences the quality of the results for similar optimization problems and is often determined

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

Require: random permutation **P**, initial matrix Ω
  $\psi \leftarrow \mathbf{P} \cdot \Omega$
  $n = width(\Omega)$
  $k = 0$
  repeat
    $\lambda_\psi = \lambda(\psi)$
    for $i, j = 0; i \neq j$ to n **do**
      permute the dimensions i, j for ψ:
      $\psi_{i,j} \leftarrow permute(\psi, i, j)$
      calculate sorting measure $\lambda_{i,j} \leftarrow \lambda(\psi_{i,j})$
      $\Delta\lambda_{i,j} \leftarrow (\lambda_\psi - \lambda_{i,j})$
    end for
    best permutation: $(i_b, j_b) \leftarrow argmax_{i,j}(\Delta\lambda_{i,j})$
    $\psi \leftarrow permute(\psi, i_b, j_b)$
    k++
  until $(\max(\Delta\lambda_{i,j}) \leq 0)$ || $k \leq 2 \cdot n)$
  return $\Omega \leftarrow \psi$

Figure 5: *Hill-Climbing Algorithm for Heuristic Reordering*

by a pre-processing step. Our approach does not apply such a step, because from our experience a good start solution is a random permutation, due to the fact that an ordered start solution is deadlocked very early in the optimization process.

Reorder Measures. Figure 6 shows that all measures improve the order of the reordered SPLOM. Nevertheless, the measure λ_r tends to distribute relevant plots over the whole SPLOM and it causes an inhomogeneous distribution. This is reasonable because this measure is strongly influenced by outliers within a window w. Furthermore, the measure λ_r "tries" to estimate the quality of the independent relevances just by using two of them (per window w) and it consequently fails. Thus, λ_r is not appropriate. The measures λ_d and λ_s perform better and form larger clusters of similar relevance. The benchmark test of Table 1 confirms that the measures λ_d and λ_s seem to be appropriate. So, which one should be preferred? λ_d penalizes outliers more strongly than λ_s due to the fact that differences of the neighbors are directly measured. Thus, it is more stable in terms of generating smooth clusters of relevance. Additionally, λ_d needs fewer arithmetic operations than λ_s, which reduces the reordering time. Therefore, we use measure λ_d within this paper.

Size of Neighborhood. The measure λ_d depends on the window size w. If the dimensionality n grows and the window size w stays constant, the clusters would have (on average) the same size and would appear smaller in relation to the growing SPLOM. Thus, larger SPLOMs need to form larger clusters in order to facilitate/ease that the user is still able to perceive the clusters. Consequently, the window size w needs to depend on the dimensionality n. We determined this correlation empirically.

For a SPLOM of a dimensionality n there is one window size w_n (out of all possible w) which gives an optimal reordering result. Therefore, for a certain dimension

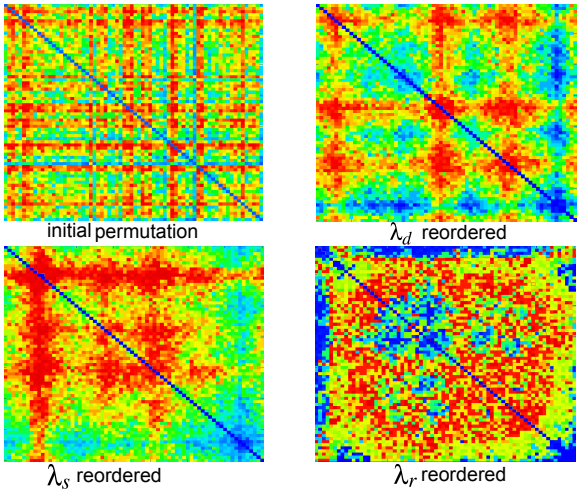

Figure 6: *Example of a qualitative comparison of the reordering results for a SPLOM with 70 dimensions and based on the measures λ_d, λ_s and λ_r. It can be seen that the reordering improves the initial permutation for all measures.*

Dimensions	Measure	λ_s	λ_r	λ_d
50	μ (sec)	66.93	115.5	54.43
	σ (sec)	12.81	17.84	8.54
	μ_{red}	0.724	0.889	0.610
	σ_{red}	0.236	0.210	0.172
70	μ (sec)	508.3	1117	429.5
	σ (sec)	97.40	174.2	60.42
	μ_{red}	0.691	0.813	0.612
	σ_{red}	0.199	0.240	0.161
150	μ (sec)	4228	9762	3702
	σ (sec)	881.7	3420	698.1
	μ_{red}	0.711	0.809	0.633
	σ_{red}	0.232	0.275	0.138

Table 1: *Benchmark Test: mean μ and standard deviation σ for the reorder runtime for 50, 70, and 150 dimensions. The value μ_{red}/σ_{red} describes the mean/standard deviation of the reduction from the measure λ_i from the initial permutation to the final permutation. Each time, 100 reorder runs with each measure (total 900 runs) have been computed with randomly initial permutation (on 3.4 GHz Intel Quad, WinXP, 4 GB RAM in single core). λ_d performed best in all categories.*

n, e.g., $n = 150$, we investigated the optimal window size $w_{n_1}, w_{n_2}, \ldots, w_{n_{10}}$ for 10 different matrices. A group of 3 users – who are familiar with data visualization – agreed on that n_i giving the best result w_{n_i}. The optimal window size is $w_n = 0.1 \cdot \sum_{i=1}^{10} w_{n_i}$ rounded to the next odd integer value. This way, we determined 7 different dimensions, illustrated in Figure 7. From empirical observations and linear regression we obtain w as a linear function in n as: $w(n) = 4.3 \cdot 10^{-2} n + 1.3$.

Now we denote the SPLOM as *reordered SPLOM*. It contains more structural information than before (cf. Fig. 6).

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

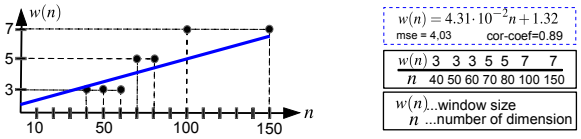

Figure 7: *Dependency between the number n of dimensions and the neighborhood size w determined by an empirical test.*

4.1.3. Step 3: Abstraction

The process of reordering already reduces texture-based optical illusions. In addition, there are further effects in terms of the human cognition that also negatively affect the reordered SPLOM's perceptibility. Thus, such effects must be treated, too. An important one is the *Stroop effect* [Str35]. It describes that the more stimuli are presented at once, the more difficult it is for a user to carry out a certain task (e.g., a task of visual search). This applies particularly to such stimuli that are processed less automatically than others. Therefore, the number of stimuli (colors, shapes) in our reordered SPLOM needs to be reduced, also to avoid mental fatigue. Another effect confirms this, namely the user's need for a *moderately abstract conceptual representation* [Zei97] in order to be able to build memories and to improve interpretation processes, respectively. For both reasons, we subsequently introduce an abstraction approach for the reordered SPLOM.

The reordered SPLOM is a real matrix, which can be visualized as a plot of density values (=density plot). In order to suppress high-frequency noise, a noise filter operation is applied by a 3×3 binomial filter [Jäh05]. However, in terms of visualization, the density values have to map onto an appropriate gray value range space, which is denoted as *normalization*. One drawback of normalization is that important visual information might be lost, depending on the underlying mapping function [SSK06].

A suitable map function $g(\sigma)$ depends on the distribution of its density values σ. This distribution is described by the histogram $H(\sigma)$ of the reordered SPLOM. Several map functions based on the histogram exist, e.g., [BGS07, EAM11]. They differ in the way of choosing and weighting the bins of the histogram. In our case, the mapping function shall fulfill the following properties:

- hide outliers, since they would cause optical illusions and mislead the user (cf. Section 4.1.2 & 4.1.3),
- map density values σ on a certain range $g(\sigma) \in [0, g_{max}]$ (,e.g, $g_{max} = 1$), and
- maximize the visual contrast.

Following [SSK06, BGS07, EAM11], these conditions are satisfied by applying the histogram equalization of Equation (1) as a map from the density values σ to the gray value $g(\sigma)$ (Figure 8 (up)).

$$g(\sigma) = \underbrace{g_{max} \cdot \int_0^{\sigma} H(\rho)\, d\rho}_{\text{continuous case}} := \underbrace{\left\lfloor \frac{g_{max}}{\sigma_{max}} \sum_{\rho=0}^{\sigma} H(\rho) \right\rfloor}_{\text{discrete case}} \quad (1)$$

Furthermore, in compliance with [Hea96] the user is not able to efficiently distinguish more than $s = 7$ different colors, thus the approach quantizes the gray space into $s = 7$ steps[‡] by:

$$g_q(g(\sigma)) = \left\lfloor \frac{g(\sigma)}{\omega} \right\rfloor \cdot \omega \text{ with } \omega = \frac{g_{max}}{s}.$$

Finally, the quantification g_q needs to be assigned to an appropriate color map. The perceptional properties of different color maps are well investigated in [BHH03, HB03]. Such color maps can be *colorblind safe* – which means they are also suitable for users who are colorblind – and *sequential* or *diverging*. A sequential color map visually emphasizes either the high end or the low end of a spectrum of values. A diverging color map emphasizes both ends of a spectrum. Both schemes yield colors that are well distinguishable within the color spaces. This minimizes color-based ambiguity errors by the user. Thus, the question arises which scheme is more appropriate? First, it is reasonable to use only color maps that are colorblind safe. Second, quality measures stably detect plots where a property is maximally correlated and maximally uncorrelated as well. For a user who wants to visually detect both, a diverging color map is more appropriate, for a user who only wants to detect the plots with the highest/lowest value of the quality measure, a sequential color map should be chosen. Our system supports both kinds of color maps, as illustrated in Figure 8 (down) (see also the case study in Section 5).

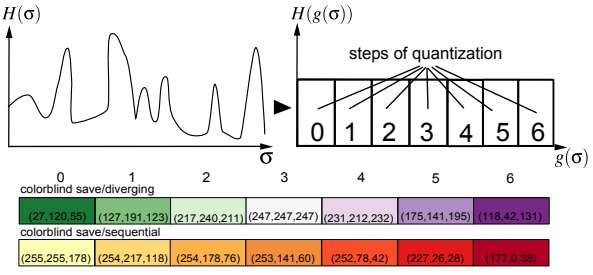

Figure 8: *Histogram equalization and quantization of the density values of the reordered SPLOM (up) and potential color maps as RGB scheme (down).*

The result of this abstraction, based on the reordered SPLOM, is an image denoted as A-SPLOM, which represents (clusters of) plots by their characteristic trends. This

[‡] Using 7 steps of quantization is arguable, but from the authors' experiences – who are users of visual analytic systems – this is a convenient number in terms of the perception of clusters.

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

is shown in Figure 9. Note that an A-SPLOM is generated exactly once within a pre-process. The colored regions

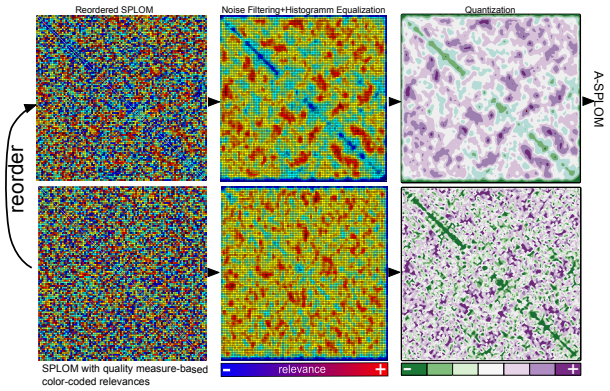

Figure 9: *Scheme of abstraction to reveal the A-SPLOM: (upper row) reordered SPLOM is equalized/quantized and yields the A-SPLOM, (lower row) unordered SPLOM is equalized/quantized and yields an abstraction with smaller/more disorganized clusters in comparison to the A-SPLOM.*

within an A-SPLOM are a metaphor for the average relevance of the underlying plots. With respect to several quality measures, also several A-SPLOMs can be obtained. Consequently, tools to select and to interact with them are required.

4.2. Interaction

For the visual exploration the user requires an interactive system, introduced in the following.

Selection. The clusters within an A-SPLOM often tend to have curvy boundaries, caused by the reordering. To address this, a circle metaphor is used that fits well to the curvy shape of the boundaries. A circle metaphor avoids selecting a lot of false positive plots as, e.g., a rectangle metaphor might cause. Thus, a circle metaphor with dynamically adjustable radius is used to select a region of interest within an A-SPLOM, as Figure 10 (up) shows. This facilitates to directly adjust the number and relevance of the underlying (coherent) plots during the user's navigation over an A-SPLOM.

An alternative selection approach arises by exploiting that a region is bounded by a constant color c. Thus, a flood fill algorithm (seeded by a point $\mathbf{p}$ in the A-SPLOM S) can be used for selection, which gives the selected region $S(\mathbf{p}) = c$. A practical extension is the selection of all adjacent regions, given by $S(\mathbf{p}) \geq c$ and $S(\mathbf{p}) \leq c$, respectively. This facilitates the selection of a certain range of relevance, as shown in Figure 10 (down). On the other hand, by using this approach the number of plots cannot be controlled directly.

Presentation of Relevant Plots. A selected region links to an amount of plots of the original SPLOM. Therefore, the

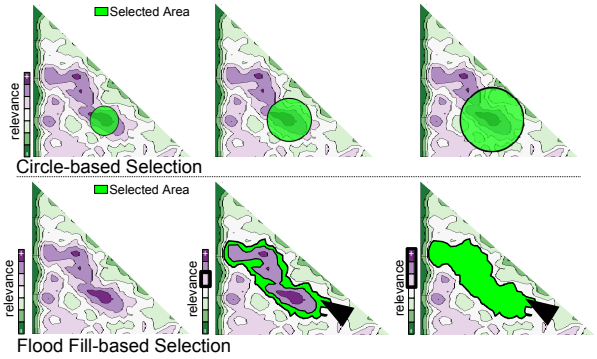

Figure 10: *Interactive Selection. (up) The circle-based selection: the number of plots can be steered by using a circle with adjustable radius, (down) the flood fill-based selection: selected region by a food fill which is seeded at the black arrow (middle), additionally selected adjacent regions in order to integrate more relevance at once (right).*

approach back-projects the bounding box of this region onto a reordered version of the SPLOM. This way, the corresponding sub-SPLOM $S_{sub}(a)$ of the region a is detected. Plots which are not within the back-projected region are faded out. Thus, the sub-SPLOM solely contains plots that correspond to the interactively selected region and selected relevance, respectively. The background color of a plot is set on a color representation of their relevance. This eases the perceptual recognition. Such a sub-SPLOM is now presented to the user who is able to visually explore the plots and their dimensional relations by interaction techniques as zooming and translation. Figure 11 schematically illustrates this.

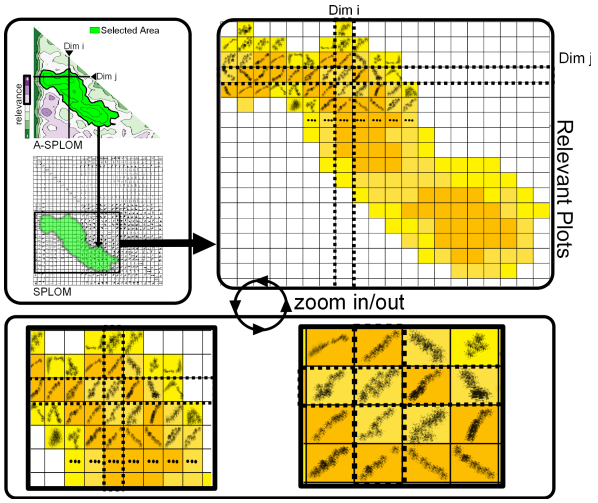

Figure 11: *Presentation of sub-SPLOMs/relevant plots: the selected region of an A-SPLOM is projected onto the corresponding region of the reordered SPLOM. Plots within the region are presented, which can be interactively explored.*

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

Visual Exploration Process. Depending on the number of used quality measures, the corresponding A-SPLOMs are integrated into an interactive system, to allow rapid analysis of the overall structure of the data set by a simple selection and navigation metaphor (cf. Figure 11). A direct comparison between associated plots within different A-SPLOMs is supported by linking and highlighting the associated plots within all other A-SPLOMs, based on the selected region of the A-SPLOM that currently has the focus (cf. Figure 12).

5. Applications

This section presents applications of the Communities data set [FA10] with 128 dimensions (8128 plots), the Subway data set [xtsdsml] with 104 dimensions (5356 plots), and the Census data set [FA10] with 42 dimensions (861 plots). In order to generate appropriate A-SPLOMs, we use the Scagnostics [WAG05] quality measures which describe a plot by nine different categories: outliers (outlying), shape (convex, skinny, string), trend (monotonic), density (skewed, clumpy), and striated. We obtain 9 different A-SPLOMs per data set, describing the nine aspects.

Figure 12 shows our graphical user interface (GUI) to select and explore relevant plots, which facilitates the visual exploration process: On the left side there are different A-SPLOMs presented, available through a scroll bar. Within each A-SPLOM arbitrary sub-SPLOMs can be selected with the aid of the mentioned interaction techniques. On the right side there is a sub-SPLOM presented: the user is able to visually explore the plots interactively.

For the Communities and Subway data set, Figure 13 illustrates a series of selected sub-SPLOMs based on a certain A-SPLOM. It can be seen that the relevance of the corresponding region within the A-SPLOM decreases, from left to right. It seems that the uniformity of the plots (concerning their individual relevance) depends on the relevance of the corresponding region: plots are nearly uniform if they link to a region with a large or a small relevance. This effect – we denote it as *uniformity property* – is interesting. It is probably caused by the uneven distribution of large/small values of relevance in comparison to values between: quality measures are convenient to "recognize" plots which fit well/worse but the values of the relevance between this spectrum are measured rather randomly, because the measure is not designed for measuring those plots. The uniformity property effect is an expression of this.

It follows that it is more likely to obtain uniform plots for regions with either large or small relevance and that this likeliness decreases to the middle of the relevance spectrum. Furthermore, as expected, it can be seen that the mean relevance of the sub-SPLOMs correlates to the relevance of the corresponding region of the A-SPLOM. Hence, this approach implements the ideas of the Shneiderman Mantra [Shn96]: *overview, zoom and filter, details on demand*.

Figure 14 shows the initial matrices, the reordered matrices, and the A-SPLOMs for all measures applied to the Communities data in comparison to the lower dimensional Census data set. In both cases, the reordered matrices are more ordered. Furthermore, it can be seen that mainly regions with a large or a small relevance form clusters, within the A-SPLOMs.

In order to evaluate the usefulness of our concept, a controlled experiment was conducted to get feedback from the user. For this, 12 study participants – who are familiar with visual analytics — were asked to select plots by the circle metaphor[§] so that both the mean value of the plot's quality measure is maximal/minimal and the number of plots is as large as possible. It has been carried out for different types of SPLOM, namely A-SPLOM, unsorted A-SPLOM (cf. Fig. 9 (down-right)), sorted SPLOM (cf. Fig. 9 (up-middle)), and unsorted SPLOM (cf. Fig. 9 (down-middle)). Furthermore, the participants did this selection for the sequential color map as well as for the diverging color map, illustrated in Figure 8 (down). In preparation, the participants have been informed about the task of interactively selecting regions within a colored triangle, in which a color encodes the value of the relevance of a quality measure. In addition, each participant practiced the selection for one randomly selected example of each type of SPLOM of a third data set. After this preparation the experiment started. Each participant conducted the controlled experiment with one data set throughout, which required 30-50 minutes per person (without the preparation). Solely real data sets have been used (cf. Table 2) and the different types of SPLOM have been presented randomly selected (uniformly distributed). This way, we gained 48 values per participant/data set. From this evaluation data, we derived several statistical properties: μ_p and σ_p are the mean and the variance of the number of selected plots, μ_q and σ_q are the mean and the variance of the values of the (normalized) quality measure from the selected plots. In total, we collected 64 values per data set in order to compare the types of SPLOM, the type of color maps, and the amount of selected plots with each other. The results are given in Table 2. This table reveals that the number of selected plots as well as the mean value of the quality measure is on average the largest for the A-SPLOMs, in comparison with the other types of SPLOM. Furthermore, the ratio between the mean value of the quality measure and the number of selected plots is the best for the A-SPLOMs. Especially without the abstraction and reordering, an over-segmentation occur more likely and a too large number of plots is selected, which negatively influences the mean value of the quality measure. Finally, all participants confirmed that the selection from our A-SPLOM is subjectively least exhausting. The same statement applies on average for the diverging color map.

[§] In contrast, the flood fill-metaphor would give the same plots for the same cluster which would be independent from the participant.

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

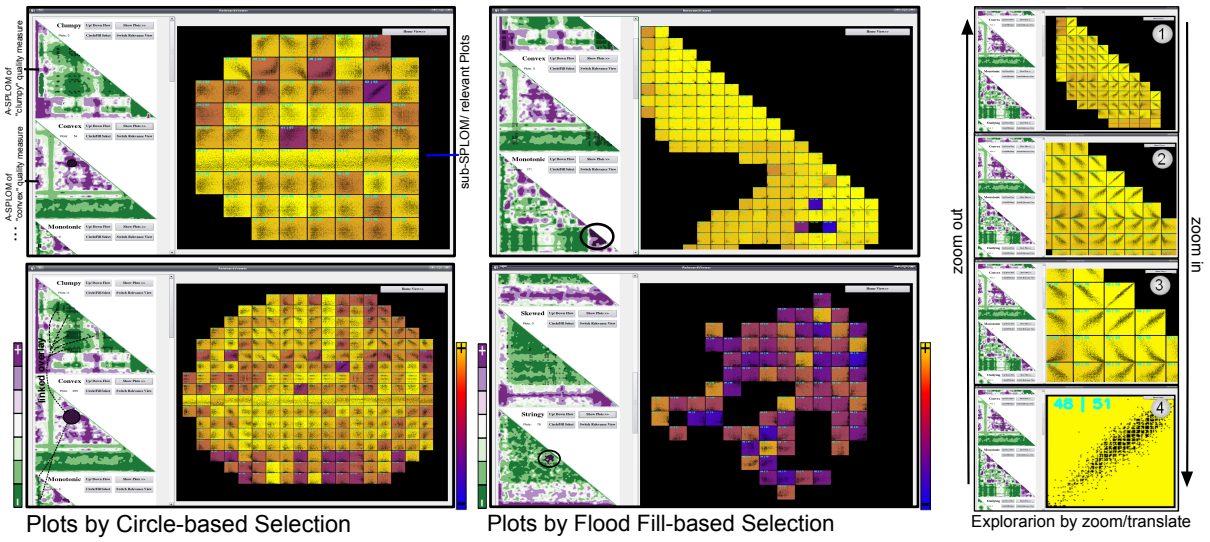

Figure 12: *GUI for selection and exploration: (left) Relevant plots selected by the circle metaphor for several radii (up-down) for the Communities data set: the number of plots can be steered well. (middle) Relevant plots with high relevance (up)/low relevance (down) selected by the flood fill metaphor (emphasized by black circles) for the Subway data set: the relevance of plots can be steered well. (right) Interactive exploration of selected plots/sub-SPLOMs.*

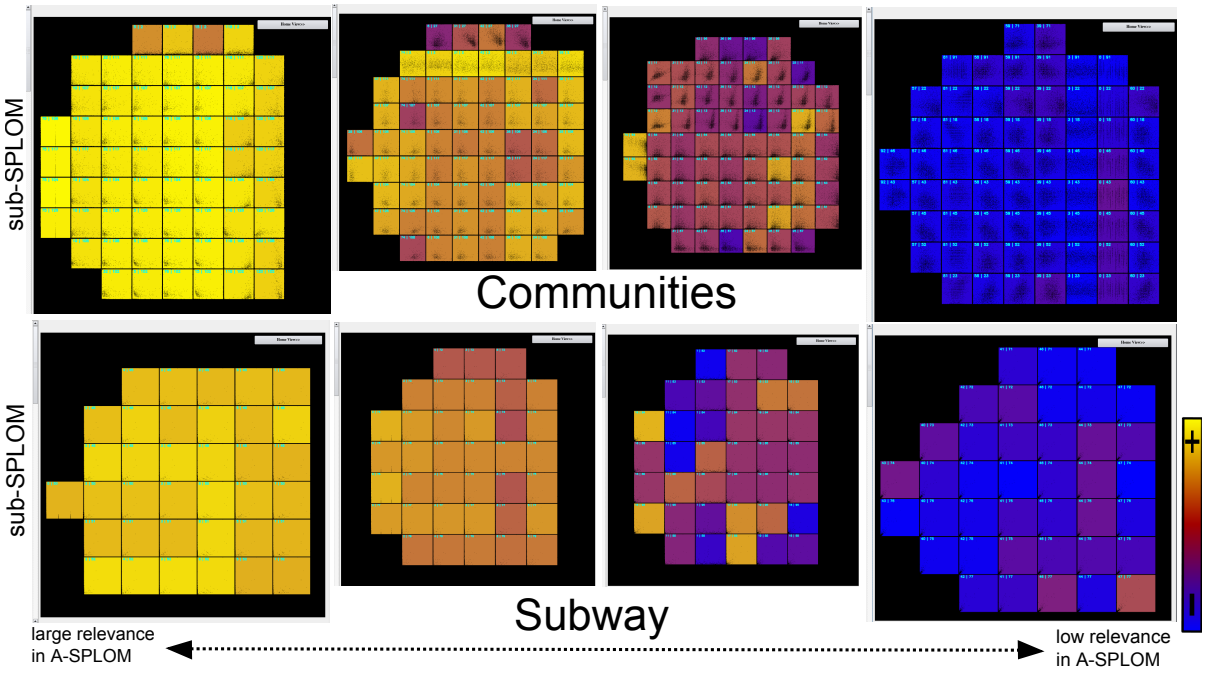

Figure 13: *Behavior of the distribution of relevance for the sub-SPLOMs of corresponding A-SPLOM's regions.*

6. Discussion

Our approach is a generic concept of selecting relevant plots in large visualization matrices and it implements the Shneiderman Mantra [Shn96]. We applied it to plots within large SPLOMs but it is also appropriate for further visualization techniques. The bottleneck of the approach is the runtime of the pre-processing step to generate the A-SPLOMs by reordering. However, the benefit of our approach is that it satisfies the required selection criteria (cf. Section 1), which is also emphasized by the case study (Table 2): the num-

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

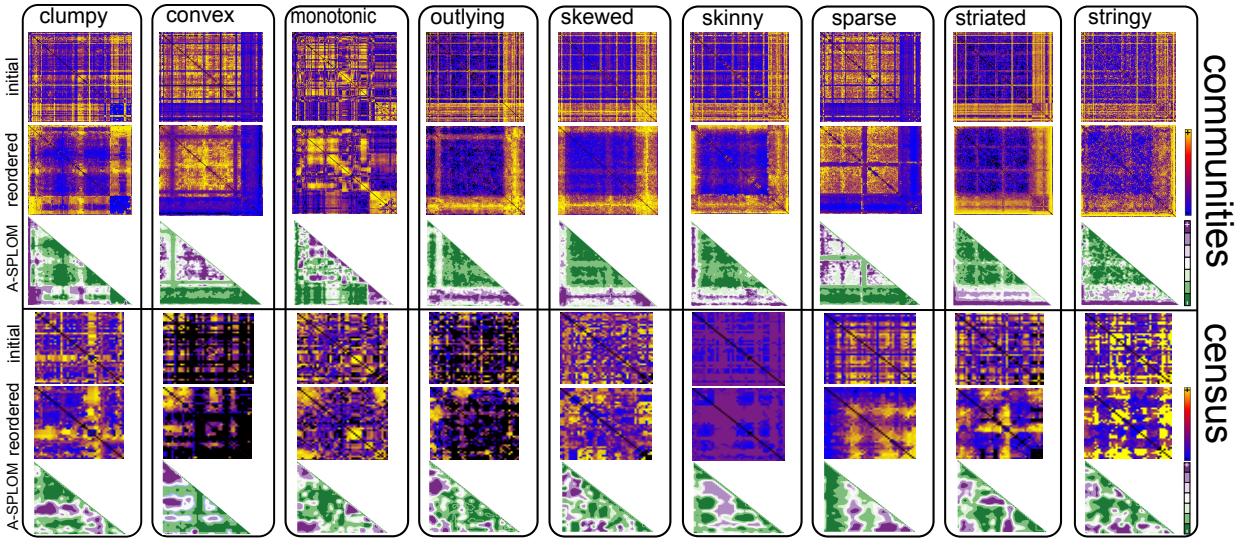

Figure 14: *Initial and reordered matrices of the Communities and Census data set for the Scagnostic measures.*

		A-SPLOM		A-SPLOM (unsorted)		SPLOM (sorted)		SPLOM (unsorted)	
		diverging colormap	sequential colormap	diverging colormap	sequential colormap	diverging colormap	sequential colormap	diverging colormap	sequential colormap
Communities (6128 Plots)	μ_p/σ_p	336,5/108,86	220,5/71,72	169,16/92,06	155,16/51,86	143,3/66,38	316,16/181,23	107,33/81,4	245,33/149,94
	μ_q/σ_q	0,72/0,12	0,55/0,13	0,65/0,11	0,69/0,088	0,53/0,14	0,39/0,11	0,43/0,14	0,49/0,06
	$\mu_p \cdot \mu_q$	**243,37**	**122,08**	**110,51**	**107,27**	**77,00**	**125,06**	**46,44**	**121,90**
	μ_p/σ_p	534,16/282,06	322,5/113,85	165,83/29,91	173,5/25,089	82,5/49,39	89,5/72,63	101,33/65,7	103.66/94,83
	μ_q/σ_q	0,13/0,04	0,07/0,05	0,04/0,02	0,04/0,03	0,1/0,12	0,1/0,15	0,05/0,02	0,02/0,02
	$\mu_p{}^{\alpha}(1-\mu_q)$	**462,17**	**298,66**	**158,60**	**166,30**	**74,23**	**79,23**	**95,95**	**101,53**
Census (861 Plots)	μ_p/σ_p	33,5/10,44	28,66/9,13	12,5/3,01	16,66/6,5	25,83/20,17	31,33/15,01	13/5,36	15,66/7,71
	μ_q/σ_q	0,81/0,17	0,81/0,10	0,65/0,27	0,55/0,30	0,77/0,13	0,65/0,19	0,64/0,25	0,62/0,26
	$\mu_p \cdot \mu_q$	**27,41**	**23,40**	**8,19**	**9,24**	**20,00**	**20,61**	**8,33**	**9,75**
	μ_p/σ_p	47/33,5	39,66/8,54	23,16/4,62	22,33/7,73	40,16/13,34	38,66/9,07	26/12,03	23,16/4,99
	μ_q/σ_q	0,15/0,37	0,14/0,11	0,11/0,13	0,089/0,06	0,15/0,014	0,003/0,016	0,005/0,02	0,0005/0,005
	$\mu_p{}^{\alpha}(1-\mu_q)$	**39,72**	**33,79**	**20,51**	**20,33**	**33,81**	**38,55**	**25,85**	**23,15**

least exhausting

Kind of SPLOM	Participants diverging	Participants sequential
A-Splom	7m 2f	1m 2f
A-Splom (unsorted) Splom (sorted) Splom (unsorted)	0	0

Legend
Study Participants...12 | 8 male (m) and 4 female (f) at the age of 22-30 years (mean 25,8)
... participant asked for plots with max. quality measure
... participant asked for plots with min. quality measure
μ_p/σ_p ... mean/ variance of number of plots
μ_q/σ_q ... mean/ variance of the values of the quality measure from the selected plots
$\mu_p \cdot \mu_q$ ... relation of the mean quality measure μ_q and the mean number of plots μ_p

Table 2: *Controlled Experiment: comparison between different colored SPLOMs in terms of diverging/sequential color maps.*

ber/relevance of plots can be steered, it is appropriate for an explorative visual search because there is no need for an a priori knowledge, and the approach scales better in the number of dimensions than comparable approaches (cf. Section 2). Finally, the associated plots of the clusters are coherent, which means that they are compact and directly comparable with each other in terms of the involved dimensions.

But there is also a drawback of the approach: one problem of the clusters of large relevance is that different initial orders of dimensions would yield different clusters of coherent plots within an A-SPLOM. In other words: it is chance which plots are coherent. We are just able to state that some plots are finally brought together, which have similar relevances.

What does this mean in terms of the visual analysis? From the amount of properties which are encoded by the underlying data set, we are just able to identify some of them by our approach because we only obtain one combination of the clusters from all combinations which are possible. On the other hand, this also means that we are able to carry out a scalable explorative visual search, in order to depict a subset of the properties of the data set. This is more than related approaches can achieve, especially concerning the scalability. Additionally, it follows that a **complete** explorative visual search cannot be carried out, as our approach is a greedy approach.

7. Conclusion & Future Work

We introduced a greedy approach to purposefully select relevant plots from large SPLOMs. The approach enables the quick selection of relevant plots without losing the context. For this, we use an abstraction approach based on quality measures called A-SPLOM to improve the user's recognition, and an interaction technique to handle A-SPLOMs and plots within a focus & context scheme. A visual search scales well without the need of an a priori knowledge on the data. On the other hand, the visual analysis depends on the reordering, and the uniformity of the resulting plots depends on the relevance of the regions. In future, we will include further visualization techniques into our prototype.

Acknowledgement

This work was supported by a grant from the German Science Foundation (DFG) from projects DFG MA2555/6-1 and DFG TH692/6-1.

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.

References

[AEL*09] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., THEISEL H., MAGNOR M.: Quality-based visualization matrices. In *Proc. of Vision, Modeling, & Visualization* (2009). 3

[AEL*10] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., THEISEL H., MAGNOR M.: Improving the visual analysis of high-dimensional datasets using quality-measures. In *IEEE Symp. on Visual Analytics Science and Technology* (2010). 4

[Ant78] ANTSTIS S.: Interactions between simultaneous contrast and colored afterimages. *Vision Research, 18* (1978), 899ff. 4

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *Journal on Scientific and Statistical Computing 6*, 1 (1985), 128–143. 2

[BC87] BECKER R., CLEVELAND W.: Brushing scatterplots. *Technometrics 29,2* (1987), 127–142. 1

[BGS07] BERTINI E., GIROLAMO A. D., SANTUCCI G.: See what you know: Analyzing data distribution to improve density map visualization. In *EuroVis* (2007), pp. 163–170. 7

[BHH03] BREWER C. A., HATCHARD G., HARROWER M.: Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science* (2003), 30(1), 5ff. 7

[Bur] BURGH P.: Peripheral viewing and simultaneous contrast. *Jour. of Experimental Psychology 16(3) (1964)*, 257–263. 4

[CBCH95] COOK D., BUJA A., CABRETA J., HURLEY C.: Grand tour and projection pursuit. *Journal of Computational and Statistical Computing 4*, 3 (1995), 155–172. 3

[Cle93] CLEVELAND W. S.: Visualizing data. *Hobart Press, Summit, New Jersey* (1993). 1

[CLN86] CARR D. B., LITTLEFIELD R. J., NICHLOSON W. L.: Scatterplot matrix techniques for large n. *Proc. Interface of Computer Sciences and Statistics* (1986), 297–306. 1

[CSS89] CHUBB C., SPERLING G., SOLOMON J.: Texture interactions determine perceived contrast. *Proc Natl Acad Sci, 86* (1989), 9631–9635. 4

[EAM11] EISEMANN M., ALBUQUERQUE G., MAGNOR M.: Data driven color mapping. In *Proc. EuroVA* (2011). 7

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE TVCG*, 14/6 (2008), 1539 et. seq. 3

[FA10] FRANK A., ASUNCION A.: UCI machine learning repository, University of California, Irvine, School of Information and Computer Sciences, 2010. 9

[FFT75] FISHERKELLER M. A., FRIEDMAN J. H., TUKEY J. W.: Prim-9: An interactive multi-dimensional data display and analysis system. In *ACM Pacific* (1975), pp. 140–145. 3

[FT74] FRIEDMAN J. H., TUKEY J. W.: A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput. 23* (1974), 881–890. 3

[HB03] HARROWER M. A., BREWER C. A.: Colorbrewer.org: An online tool for selecting color schemes for maps. In *The Cartographic Journal* (2003), pp. 40(1): 27–37. 7

[Hea96] HEALLEY C.: Choosing effective colours for data visualization. *Proc. of IEEE Visualization* (1996), 263ff. 7

[HGM*97] HOFFMAN P., GRINSTEIN G., MARX K., GROSSE I., STANLEY E.: Dna visual and analytic data mining. *In Proc. of 8th Conference on Visualization* (1997), 437ff. 1

[Hub85] HUBER P. J.: Projection pursuit. *The Annals of Statistics 13*, 2 (1985), 435–475. 3

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (1985), 69–91. 1

[Ins09] INSELBERG A.: Parallel coordinates. *Publisher Springer Berlin* (2009). 1

[Jäh05] JÄHNE B.: *Digital Image Processing*. 2005. 7

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE TVCG 15*, 6 (2009), 993–1000. 3

[Kei] KEIM D.: Designing pixel-oriented visualization techniques: Theory & applications. *IEEE TVCG 6 (2000)*, 59ff. 3

[Koh90] KOHONEN T.: The self-organizing map. *Proceedings of the IEEE 78*, 9 (1990), 1464–1480. 3

[PWR04] PENG W., WARD M., RUNDENSTEINER E.: Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proc. IEEE Info Vis* (2004), pp. 89–96. 3

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages* (1996), pp. 336–343. 9, 10

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum* (2009), vol. 28 (3), pp. 831–838. 4

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4* (July 2005), 96–113. 3

[SSK06] SCHNEIDEWIND J., SIPS M., KEIM D.: Pixnostics: Towards measuring the value of visualization. *Symp. on Visual Analytics Science And Technology* (2006), 199–206. 4, 7

[Str35] STROOP J.: Studies of interference in serial verbal reactions. *Jour. of Experimental Psychology 18* (1935), 643–662. 7

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. of IEEE Symposium on Visual Analytics Science & Technology* (2009). 4

[TBB*10] TATU A., BAK P., BERTINI E., KEIM D. A., SCHNEIDEWIND J.: Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *AVI* (2010), pp. 49–56. 3

[TT85] TUKEY J., TUKEY P.: Computing graphics and exploratory data analysis: An introduction. *In Proc. of Sixth Annual Conference & Exposition: Computer Graphics 85* (1985). 3, 4

[VMCJ10] VIAU C., MCGUFFIN M. J., CHIRICOTA Y., JURISICA I.: The flowvizmenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration. *IEEE TVCG 16* (2010), 1100–1108. 3

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. *IEEE Info Vis* (2005), 157–164. 3, 4, 9

[xtsdsml] XMDV TOOL: SUBWAY DATA SET:, http://davis.wpi.edu/xmdv/datasets/subway.html. 9

[YPWR03] YANG J., PENG W., WARD M. O., RUNDENSTEINER E. A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *IEEE Symp. on Information Visualization, 105–112* (2003). 2

[Zei97] ZEITZ C. M.: Some concrete advantages of abstraction: How experts representations facilitate reasoning. *Expertise in Context: Human & machine, AAAI Press/MIT Press* (1997). 7

© 2012 The Author(s)
Journal compilation © 2012 The Eurographics Association and Blackwell Publishing Ltd.