

Visualizing Multidimensional Data with Glyph SPLOMs

A. Yates*, A. Webb*, M. Sharpnack*, H. Chamberlin*, K. Huang*, and R. Machiraju*

* The Ohio State University, USA

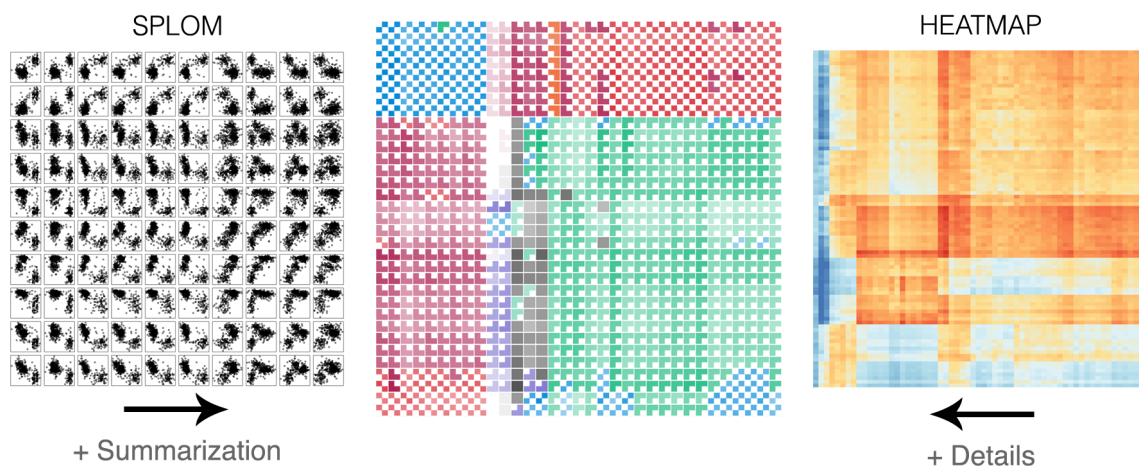


Figure 1: *Glyph SPLOM uses color-coded, areal glyphs to summarize a scatterplot matrix.*

Abstract

Scatterplot matrices or SPLOMs provide a feasible method of visualizing and representing multi-dimensional data especially for a small number of dimensions. For very high dimensional data, we introduce a novel technique to summarize a SPLOM, as a clustered matrix of glyphs, or a Glyph SPLOM. Each glyph visually encodes a general measure of dependency strength, distance correlation, and a logical dependency class based on the occupancy of the scatterplot quadrants. We present the Glyph SPLOM as a general alternative to the traditional correlation based heatmap and the scatterplot matrix in two examples: demography data from the World Health Organization (WHO), and gene expression data from developmental biology. By using both, dependency class and strength, the Glyph SPLOM illustrates high dimensional data in more detail than a heatmap but with more summarization than a SPLOM. More importantly, the summarization capabilities of Glyph SPLOM allow for the assertion of “necessity” causal relationships in the data and the reconstruction of interaction networks in various dynamic systems.

Categories and Subject Descriptors (according to ACM CCS): H.5.0 [Information Interfaces]: General—

1. Introduction

Much of data science can be distilled to a single question: *what relates to what?* Current practice in quantitative high dimensional analysis typically focuses on the *what*: what variables or features are interesting, which are relevant, and what grouping of these fit best? In contrast, comparatively little attention has been spent on the *relates* part of the data science question: effectively, the patterns evident in scatterplots. It is these patterns of relations that imply a topology of causal relations in data, and it is the methodological treatment of these that motivates this work.

Currently, visual analysts use two complementary approaches to explore all-pairs relations in large datasets: *maximize summarization* by computing representative values such as correlation, or *maximize detail* by inspecting individual scatterplots. These values or scatterplots are then organized into an all-pairs matrix and plotted to form a heatmap or a scatterplot matrix (SPLOM) respectively. In the former case, details about patterns evident in the scatterplot are discarded, while in the latter case, inspecting individual scatterplots is infeasible for more than a few dozen scatterplots.

New approaches are now possible given the advances in

non-linear statistical dependency testing [RN13]. However, there is a lack of methods to understand, classify, and organize the non-linear dependencies that can now be detected with these new, more powerful statistical methods. These non-linear dependencies are apparent in scatterplots, but they are not well captured by heatmaps.

To address this, we have developed a methodology in which we richly describe pairwise relations using both dependency class and a non-linear dependency strength measure. To visualize these richer descriptions, we present Glyph Scatterplot Matrix, or Glyph SPLOM (see Figure 1, center). Glyph SPLOM combines a general measure of dependency strength, distance correlation [SRB07], with a dependency class that is based on the occupancy of scatterplot quadrants [SDG*08]. Together, dependency strength and dependency class reveal important patterns that separately they fail to distinguish, see Figure 2. We map each relationship description to a distinct areal and orientated glyph and arrange these glyphs in a matrix ordered by hierarchical clustering. It is evident that the Glyph SPLOM reveals more detail than both the plain SPLOM and the heatmap when deployed on the same dataset. In brief, by summarizing the plots in a scatterplot matrix with glyphs, even higher dimensional data can be considered. The Glyph SPLOM organizes glyphs to reveal patterns and produces a visually compelling summary of relationships not obtained by existing methods.

Further, a Glyph SPLOM matrix can be transformed into an adjacency matrix of a directed network using the “necessary for” relations summarized by our technique. The resulting directed *necessity network* or *NecNet* implies a partial order of variables that can be used to infer causal relations from data. In other words, the summarization capabilities of the Glyph SPLOM allows for the reconstruction of interaction networks in various dynamical systems. We present a proof of concept of this idea using a *C. elegans* gene regulatory network in Subsection 4.2, but this technique and its variations have potentially many applications in the domains of social, physical, and other biological networks.

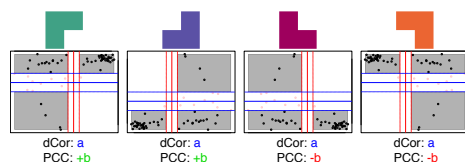


Figure 2: An example of different logical implication classes with the same dependency strength. PCC: Pearson's correlation coefficient.

Roadmap: In Section 2, we describe previous work on the visualization of multidimensional data. Later in Section 3, we describe our rationale for the Glyph SPLOM and outline the approach for constructing and visualizing *NecNet*. In Section 4, we describe the application of the Glyph SPLOM on two datasets. Finally in the last two sections we first

discuss limitations of the Glyph SPLOM and then summarize our contributions and point to the future. We publish the source code and corresponding documentation for Glyph SPLOM as an open source, Bioconductor-style R package at: [github.com/andrewdyates/gsplom.rpackage].

2. Related Work

We highlight related work in the survey below, ranging from multidimensional data visualization to the statistical estimation of dependencies. It should be noted that much of the reported work does not offer direct solutions to visualize relationships in the data. Techniques to effectively handle more than a few dozen variables in a scatterplot matrix is an active area of research [HBO10, CM84, LMK07]. Current work in this area primarily applies two strategies: limit or reduce the number of variables (dimensions) considered, or summarize pairwise relationships using an abstraction, such as a graph, and visualize the abstraction instead [MG13].

Dimensionality reduction includes techniques of feature mapping, variable or feature selection strategies, and linear and non-linear projections and embeddings. Scagnostic based approaches [WAG05, TT85] summarize each scatterplot as a set of visually apparent features, like “skewed” or “clumpy”. In this way, one maps an N -variable SPLOM of M samples to a K -feature SPLOM of N samples, $K \ll N$. This technique summarizes gross patterns in scatterplots and can highlight outlying relationships in the data. However, the original measurements and variables are obscured in the K -feature SPLOM, and the relevance of scagnostic dimensions in statistical inference is unclear. Other work [Sam69, YXZ*07] mathematically projects an N dimensional variable space into K representative dimensions using the methods of Singular Value Decomposition (SVD), and Multidimensional Scaling (MDS). However, interpretation of features in projected space can be challenging, and not all data can be reasonably reduced in this way. Finally, many methods and tools select a subset of “best” or “most interesting” variables to plot in a SPLOM based on a variety of criteria such as a domain-based “quality measure.” [AEL*09, EDF08, VMCJ10, War94, LAE*12].

Relationship summaries map each pair of variables to some abstraction including a statistical measure or a set of features. These summaries are then visualized in place of the data. Friendly [Fri02] compiles several techniques of displaying an all-pairs matrix of one popular relationship measure, correlation. Wilkinson and Friendly [WF09] document the long history of visualizing matrices as clustered heatmaps of various measures. One can consider a heatmap to be a “maximally summarized” SPLOM in the sense that each scatterplot is condensed to a single number and then visualized. However, many details about the relationship cannot be captured by existing applications of this method. For enhanced visual display, Keim [Kei00] outlines the advantages of “pixel-perfect” visualization strategies. When

adapted to Glyph SPLOMs, they can be very effective as we show later.

A matrix of scalar relationship summaries can be interpreted as an adjacency matrix of a graph or network, and graphs themselves can be visualized by a wide variety of graph layout techniques [HMM00]. Alternatively, one can visualize the adjacency matrix directly [DWvW12]. Such network abstractions of relationship summaries can be far removed from a direct examination of all pairs of variables in the data. Brandes and Nick [BN11] present a matrix of highly detailed glyphs that represent complex pairwise relationships in a social network over time. However, the design, layout, and interpretation of these glyphs are specific to its domain, and it is not clear how this technique can be applied to other domains. In biological network reconstruction, Schneidman *et al.* [SBSB06] show that pairwise correlations can explain much of the biological function observed in a neuronal network and suggest that all-pairs dependency frameworks similar to ours can describe biological networks.

Finally, we survey methods that capture or estimate dependencies. Sahoo *et al.* [SDG*08] present a general method of classifying dependencies as “Boolean implications”, which we apply and extend. Sahoo *et al.* also demonstrate an application of “necessary for” dependency class mining in developmental biology [SSB*10]. However, the primary focus of these works is assigning pairs of genes to invariant “boolean implications” as tabulated in a database. It does not include notions of dependency strength, clustered matrices of all-pairs dependency class, a systematic treatment of all “necessary for” relations as a network, or visualizations. Finally, Reshef *et al.* [RRF*11] present maximal information-based non-parametric exploration (MINE) statistics to quantify particular qualities of dependencies like “asymmetry” and “non-linearity”, which in turn could be used as features in other classification approaches and visualization techniques.

3. Methods

In this section we first describe our guiding principles of visual design. Then, we will describe the steps in constructing and visualizing a GSPLOM. Converting a scatterplot matrix (SPLOM) to a Glyph SPLOM involves two steps. First, we summarize each scatterplot by dependency strength and dependency class using distance correlation and logical implication class respectively. Second, we order the rows and columns of the SPLOM to cluster similar scatterplots together using hierarchical clustering resulting in the Glyph SPLOM. Lastly, we describe visual enhancements to the Glyph SPLOM rendering it efficient and scalable.

3.1. Glyph SPLOM: Underlying Visual Design

The Glyph SPLOM was created to answer the following questions about interacting entities in a social, physical or biological network: (i) How can one easily infer pair-wise relationships between two variables? (ii) How many separate functional blocks or groups exist? (iii) Can the size

of each of block be easily estimated? (iv) What “group” or block interactions can be inferred from the representation? (v) Can all these tasks be completed without any help from an expert and/or with little background knowledge of the domain?

The mainstay of Glyph SPLOMs includes the encoding of pair-wise or bi-variate relationships as visually discriminative areal and orientable glyphs. In essence, each glyph is a robust summary of the underlying bivariate relationship and hence provides a direct response to the very question (i). It is also necessary that the glyph visually depict a certain relationship between variables. Consider this, if X (horizontal axis) is necessary for Y (vertical axis), then an increase in X should be accompanied by an increase in Y . The corresponding glyph should clearly capture this relationship. We call this the dependency class XY (Section 3.3) and it actually suggests an “increasing” relationship between two interacting pairs. Further, the chosen glyphs should depict all possible classes of relationships between two variables. In this study we employ a 2×2 grid of dimension-less pixels to visually describe all possible glyphs for relationships between two variables (Section 3.3). Each glyph and its pixel map uniquely denote a specific class of relationship; therefore, no glyph can be obtained by scaling a different glyph. This is useful because, in practice, a Glyph SPLOM can be stretched or scaled during the final step of visual mapping, but this property will prevent ambiguity in stretched or scaled glyphs. It should be noted that the glyphs are not defined in terms of pixel sizes. In the final visualization, they will be mapped to a 2×2 grid of pixels. Further, after clustering the elements using our methods, a global summary view can be obtained that can be used to group elements and establish a partial order. Thus the final Glyph SPLOM will help answer the remaining questions listed above. It should be noted that each scatterplot and in turn a glyph is indeed a summarized edge and therefore will lead to the reconstruction of the dual directed graph. Many of the above tasks cannot be completed through the use of either a heatmap or a plain SPLOM. The glyphs visually enrich and annotate the SPLOM with extra information.

3.2. Step I-A: Dependency Strength - Distance Correlation

Szekely *et al.* [SRB07] rigorously define distance correlation and expound its use in various applications. New statistical tests include multivariate distance rank association measure HHG (Heller-Heller-Gorfine) [HHG12] and the maximal information coefficient (MIC) [RRF*11]. Both are computationally expensive; however, HHG is useful in the presence of small sample sizes while MIC is not. Simon and Tibshirani [ST11] succinctly comment on issues with MIC, particularly for sample sizes of $n < 100$. While these tests measure dependency strength for many types of dependencies, they do not distinguish different types of dependencies from each other. On the other hand, Snijder *et al.* [SLF*13] propose the hierarchical interaction score (HIS), an asymmet-

ric, non-linear dependency statistic that predicts functional relationships better than existing methods. HIS is similar to Glyph SPLOMs “necessary for” summary relations XY and YX (see Section 3.3).

Distance correlation (dCor) is a general measure of statistical dependence that is equal to zero if and only if two vectors are independent [SRB07]. It has many attractive features: it is between 0 and 1, it has good statistical discriminative power, it has an explicit formulation that can be efficiently computed as matrix operations, and in the bivariate normal case, dCor is a function of Pearson’s correlation, (ρ) where $\frac{dCor}{|\rho|} \approx 0.9$ and $dCor = |\rho|$ when $\rho = \pm 1$. We use a simplified distance correlation formulation using Euclidean distance on single dimensional vectors as follows.

Let X and Y be two vectors of n values. Compute all n^2 pairs of absolute differences (Euclidean distance in one dimension) between all pairs of n values in X . Repeat for Y .

$$\begin{aligned} A_{i,j} &= |X_i - X_j|, & i, j &= 1, 2, \dots, n \\ B_{i,j} &= |Y_i - Y_j|, & i, j &= 1, 2, \dots, n \end{aligned} \quad (1)$$

Center all-pairs distance matrices A and B over rows and columns where $\bar{A}_{i,*}$ is the mean of i^{th} row, $\bar{A}_{*,j}$ is the mean of j^{th} column, and $\bar{A}_{*,*}$ is the mean of all of the values in A . Repeat for B .

$$\begin{aligned} \hat{A}_{i,j} &= A_{i,j} - \bar{A}_{i,*} - \bar{A}_{*,j} + \bar{A}_{*,*} \\ \hat{B}_{i,j} &= B_{i,j} - \bar{B}_{i,*} - \bar{B}_{*,j} + \bar{B}_{*,*} \end{aligned} \quad (2)$$

Distance covariance (dCov) squared is the average product of corresponding matrix values in \hat{A} and \hat{B} . Distance variance (dVar) is the distance covariance of a variable with itself. Then, Distance correlation (dCor) is $dCov(X, Y)$ divided by the geometric mean of $dVar(X)$ and $dVar(Y)$; see Eq. 3.

$$\begin{aligned} dCov(X, Y) &= \sqrt{\frac{\hat{A} \bullet \hat{B}}{n^2}} \\ dVar(X) &= dCov(X, X) \\ dCor(X, Y) &= \frac{dCov(X, Y)}{\sqrt{dVar(X) dVar(Y)}} \end{aligned} \quad (3)$$

3.3. Step I-B: Dependency Class

Dependency classes are the patterns produced by dividing a scatterplot into quadrants. Based on which quadrants are filled or empty, we assign one of the seven possible dependency classes. Each of these patterns can be mapped to a logical function of two high/low Boolean variables, hence the name “logical implication” dependency class. Table 1 lists all logical implication classes. In our classification scheme, we do not allow an empty left/right or top/bottom margin because it is typically uninformative to discretize a variable to a single level. However, to allow for the odd applications re-

quiring empty margins and other sundry exceptions of classification, we define an eighth “null” class, NA.

Class	Sign	Bias	Glyph	Binary	Implication
YX	+	Left		1110	Y Necessary for X
PC	+	None		1010	$X \Leftrightarrow Y$
XY	+	Right		1011	X Necessary for Y
UNL	0	None		1111	Non-linear or Ind.
MX	−	Left		0111	Mutual Exclusion
NC	−	None		0101	$\bar{X} \Leftrightarrow Y$
OR	−	Right		1101	$X \text{ OR } Y$
NA				----	Null

Table 1: Logical Implication Classes with Glyphs

We assign each class a sign and bias to more easily discuss and describe them. Sign corresponds to positive or negative correlation of filled quadrants. Bias corresponds to an asymmetrically filled quadrant on the left or right. Symmetric classes have no bias, and the UNL class has no sign or bias. Sign and bias are undefined for NA. See Table 1.

Classifying scatterplots as logical implication dependency classes has the following workflow:

1. Discretize each variable:
 - a. Find high/low threshold.
 - b. Find uncertainty interval.
 - c. Discretize each sample as low, uncertain, or high.
2. Assign classes to scatterplots:
 - a. Test if each quadrant is filled or empty.
 - b. Assign a class by which quadrants are filled.

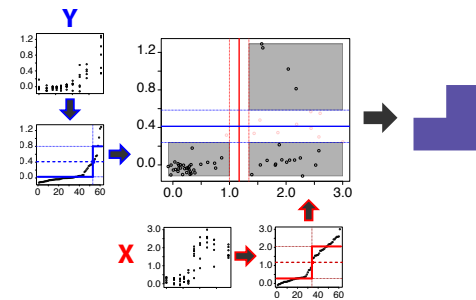


Figure 3: Workflow for Logical Implication Classification.

See the classification workflow diagram in Figure 3 for an illustration. Our method handles uneven distributions of high and low values, and it robustly determines which quadrants are “empty.” We describe below our workflow in some detail.

3.3.1. Discretization For The Glyph SPLOM

We describe pre-processing data transformation steps that allows for the mapping of each scatterplot into a glyph.

Find high/low threshold: To accommodate uneven distributions of high and low values, we sort the values from low to high and fit a step function to the sorted values. The high/low threshold is the average of both the high and low steps of the fitted function.

Find uncertainty interval: The uncertainty interval is a range centered on the high/low threshold in which samples are discretized as neither high nor low. By default, we assume that the variables with the lowest variance in a multidimensional dataset are signal-free Gaussian measurement noise, and as is norm we use twice the 3rd-percentile standard deviation of all variables in the given data as a “global uncertainty” estimate.

Discretize each sample as low, uncertain, or high: We discretize each variable as “low,” “uncertain,” or “high” depending on if a sample value is less than, within, or greater than the uncertainty interval bounds respectively. Values discretized as “uncertain” are ignored in logical implication classification.

3.3.2. Assign Classes To Scatterplots

This is the stage where the pairwise relationships and an appropriate glyph is assigned to every scatterplot. Although an implicit visual shape is implied from the occupancy patterns in the scatter plot, the actual visual presentation is only completed in the last stage.

Test if each quadrant is filled or empty: A scatterplot is divided into four quadrants plus a cross-shaped uncertainty region centered on the high/low thresholds. Simply, a quadrant is empty if there are no samples in it. Otherwise, it is filled. We allow for some error by counting a quadrant as empty if a statistical test based on the difference between the expected and observed number of samples in a quadrant is above the user parameter threshold Z ; see Equation 4. $|Q_{xy}|$ is the number of samples in a particular quadrant, $|Q_{x*}|$ and $|Q_{*y}|$ are the number of samples in each of the margins that contain that quadrant, and $|Q_{**}|$ is the total number of samples in any quadrant.

$$\text{Expected}(Q_{xy}) = \frac{|Q_{x*}| |Q_{*y}|}{|Q_{**}|} \quad (4)$$

$$\text{SparsityTest} : Z < \frac{\text{Expected}(Q_{xy}) - |Q_{xy}|}{\sqrt{\text{Expected}(Q_{xy})}}$$

Assign a class by which quadrants are filled: Based on the pattern of filled and empty quadrants, we assign the scatterplot to one of the seven classes in Table 1. Class binary strings represent which quadrants are filled (1) or empty (0) in standard quadrant order: counter-clockwise from the “X high, Y high” quadrant. In the case of a classification exception, for example, there are too many samples in the uncertainty region or there is an empty margin, we assign the null class, NA.

3.4. Step II: Ordering SPLOM Using Hierarchical Clustering

In a typical heatmap, the rows and columns are reordered using hierarchical clustering to reveal patterns in the data. This clustering is based on similarities between measures of dependency strength—typically correlation or mutual information. We refine this technique by also incorporating similarities in dependency class in the hierarchical clustering. This is because different classes may have the same dependency strength, and so clustering on dependency strength alone is insufficient to distinguish different dependency classes as demonstrated in Figure 2.

To proceed, we need a measure of similarity (inverse distance) between logical implication classes. We use the Hamming distance between class binary strings. In other words, the distance between two classes is the minimum number of “quadrants” that must change to transform one class to the other. For completeness, the distance from NA to all other classes is the average Hamming distance between any two different classes which evaluates to 2.05. Let m be the number of variables in the dataset, and let $DCOR$ and CLS be m by m all-pairs-variables distance correlation and logical implication class matrices respectively (Eq. 5). We can now compute all-pairs-rows distance matrices of both $DCOR$ and CLS as Δ and Γ respectively (Eq. 6). We normalize the ranges of these distance matrices, add them together, and proceed with ordinary average linkage hierarchical clustering of the SPLOM rows and columns using the sum distance matrix $SumDist$ (Eq. 7).

$$DCOR_{i,j} = dCor(X_i, X_j), \quad i, j = 1, 2, \dots, m \quad (5)$$

$$CLS_{i,j} = LogicClass(X_i, X_j)$$

$$\Delta_{i,j} = \|DCOR_{i,*} - DCOR_{j,*}\|, \quad i, j = 1, 2, \dots, m$$


$$\Gamma_{i,j} = \sqrt{\sum_k^m \text{HammingDist}(CLS_{i,k}, CLS_{j,k})} \quad (6)$$

$$SumDist = \Delta + \frac{\Gamma}{2} \quad (7)$$

3.5. Visual Enhancements For The Glyph SPLOM

We map filled pixels to filled quadrants as described earlier; see Table 1 for all glyphs and their logical implication class mappings. To improve visual acuity and appeal, we map dependency measurements to various visual attributes namely tint and hue [Kei00].

We tint glyphs by their dependency strength on an inverse linear scale: strong dependencies are brightly colored, and weak dependencies are tinted nearly white. Maximum tint (white) is set at the distance correlation statistical significance threshold, a user parameter. This parameter, the highest $dCor$ value expected by chance if there is no dependency, can be estimated by permutation testing.

We choose opposing primary color hues for the two opposing symmetric glyphs of opposite signs: blue for  (PC)

and red for \blacksquare (NC). Asymmetric glyphs are adjacent in hue on either side of the symmetric glyph of the same sign. This groups glyphs of the same sign by hue while clearly distinguishing glyphs of opposite bias. We assign black (no hue) to the UNL glyph \blacksquare to emphasize that the UNL class has no sign or bias and is equidistant from glyphs of opposite signs.

For large Glyph SPLOMs, 2x2 pixel glyphs can be difficult to interpret; our preliminary user study reveals this limitation (see Section 4.3). To address this, one can further compress the Glyph SPLOM by replacing each 2x2 pixel glyph with a single pixel of its corresponding color. The summarized figure eventually becomes more difficult to interpret. However, it encodes the same information as the full Glyph SPLOM with only one-fourth of the “data ink.” We recommend the single pixel compression option for Glyph SPLOMs of over about 300 variables.

4. Applications

We demonstrate the generality and usefulness of the Glyph SPLOM and its advantages over existing techniques by applying it to two collections of multi-dimensional data. In Section 4.1, we first use a dataset from the World Health Organization (WHO) to show that Glyph SPLOM presents and organizes underlying relationships in a multidimensional data of over 200 variables. Later, in Section 4.2, we use *C. elegans* genomic data to show that a Glyph SPLOM summarizes meaningful patterns and can be extended to produce directed *necessity networks* that represent functional inferences (gene regulation in this case). Finally, we report a preliminary user study that demonstrates the relative efficacy of the Glyph SPLOM over competing techniques.



Figure 4: The complete Glyph SPLOM for the WHO data.

4.1. Application: World Health Organization Data

Reshef et al. [RRF*11] compiled a dataset of World Health Organization (WHO) demographic and sociological indicators for use in a demonstration of the general utility of

a novel dependency statistic in data exploration. Likewise, we use this dataset to demonstrate the general utility of the Glyph SPLOM. The data were downloaded from `exploredata.net` and we removed variables with fewer than 100 samples. At 258 variables measured in 202 countries, this dataset is too large to be feasibly visualized as a scatterplot matrix.

We present advantages of the Glyph SPLOM (Figure 4) over signed correlation using selected examples of one of each logical dependency class that includes “Years of Life Expectancy from Birth” (longevity) as shown in Figure 5. Signed correlation does not distinguish relations with the same sign in each row, but the Glyph SPLOM reveals meaningful differences that are apparent in the corresponding scatterplots. The complete Glyph SPLOM in Figure 4 preserves these distinctions and organizes these examples into groups with other WHO indicators, revealing various patterns of dependency classes. We include Figure 4 in the supplementary materials in higher resolution and with labels. Further, we make some observations pertaining to different classes.

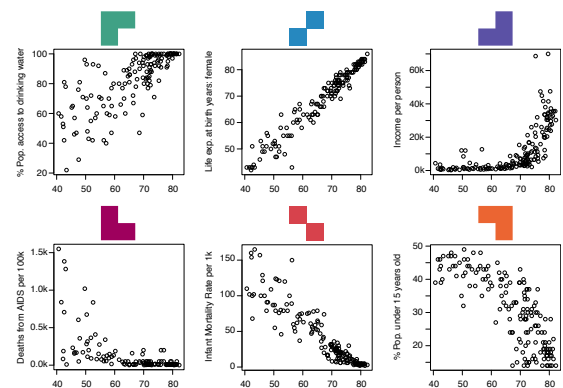


Figure 5: WHO Per Class Examples. Life Span in Years (from birth) versus other measures per country; one example per signed logical implication class.

Positive Correlations with Longevity in Figure 5

1. \blacksquare % Population access to drinking water is necessary but not sufficient for longevity.
2. \blacksquare Female life expectancy at birth (years) is symmetrically correlated with overall longevity. This is an example of a strong but trivial correlation that could be filtered from a ranked list of dependencies by dependency class.
3. \blacksquare Income per person: Longevity is necessary but not sufficient for individual wealth.

Negative Correlations with Longevity in Figure 5

1. \blacksquare Deaths from AIDS per 100k: Mutual exclusion: no country has an AIDS epidemic and high longevity, but not all countries with low longevity have an AIDS epidemic.
2. \blacksquare Infant Mortality Rate per 1k: Infant mortality and

longevity are symmetrically and inversely related. No country has excellent infant health outcomes but poor overall longevity and *vice versa*.

3. ■ % Pop. under 15 years old: Logical OR: While a young population is inversely correlated with longevity, there are countries with many children and high longevity.

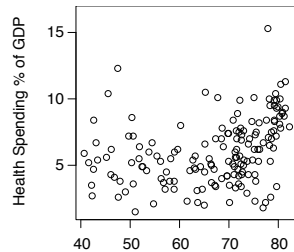
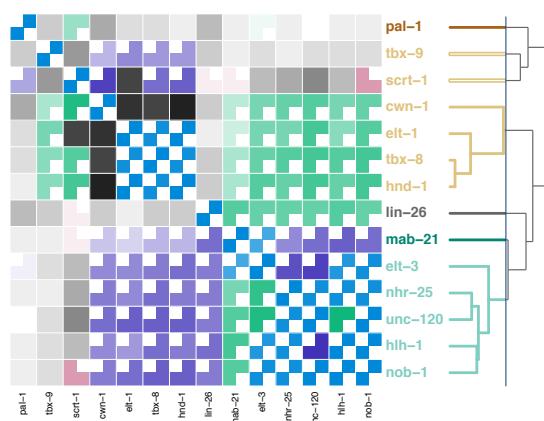


Figure 6: WHO Non-linear Dependency Example; Life Span in Years (from birth) versus Health spending as % GDP.

Non-linear and Dependent Relationships

The Glyph SPLOM also reveals and organizes non-linear dependencies in the WHO dataset using the ■ glyph. In Figure 6, we observe a significant non-linear dependency, potentially a “U-shaped” relationship or a multi-modal function, between longevity and health spending as % GDP. Presumably, countries with the lowest life expectancies have low overall GDP yet must spend comparatively more of it to manage public health crises. Otherwise, health spending as % GDP is positively correlated with longevity. This is an observation that we make from the Glyph SPLOM which must be validated by policy makers after consulting with other sources of data. Note that not all significant ■ glyphs necessarily represent non-linear dependencies; they may also represent noisy linear relations where enough samples are in all four quadrants.



Dendrogram from hierarchical clustering. Genes colored by biological phase: **Phase I**, **Phase II**, **Phase III**, **Phase IV**, Not Classified.

Figure 7: A Glyph SPLOM for the *C. elegans* PAL-1 genes.

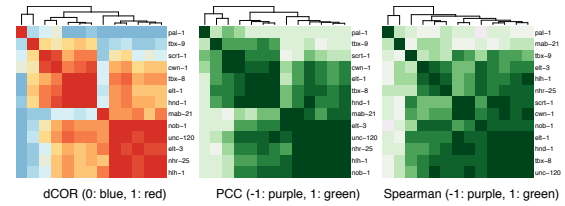


Figure 8: *C. elegans* heatmaps for PAL-1 genes.

4.2. Application: *C. elegans* Gene Regulatory Network

Caenorhabditis elegans (*C. elegans*) is a model organism widely studied in developmental biology. In the data presented in [BHC*05], gene expression levels were measured over time as embryonic stem cells differentiate. Here, we apply the Glyph SPLOM to visualize relationships between genes that encode a class of important proteins known as *transcription factors* or TFs. TFs regulate expression of other genes and thus play an essential role in cell growth and development. Therefore, it is critical to understand the regulatory relationships among the TFs in order to decipher the developmental process. Raw *C. elegans* microarray gene data were downloaded from NCBI Gene Expression Omnibus with accession number GSE2180 and were normalized using the SCAN-UPC algorithm [PSC*12]. Only wild-type and mutant samples were selected for this study.

In Figure 7, we show Glyph SPLOM clusters of 14 selected TF genes in the PAL-1 TF dependent lineage over the four developmental phases described by Baugh et al. [BHC*05]. Notably, the “X is necessary for Y” glyph ■ often corresponds to “X is activated before Y” in the developmental phase order, a directed inference that cannot be made using correlation or symmetric measures of dependency strength alone; see Figure 8.

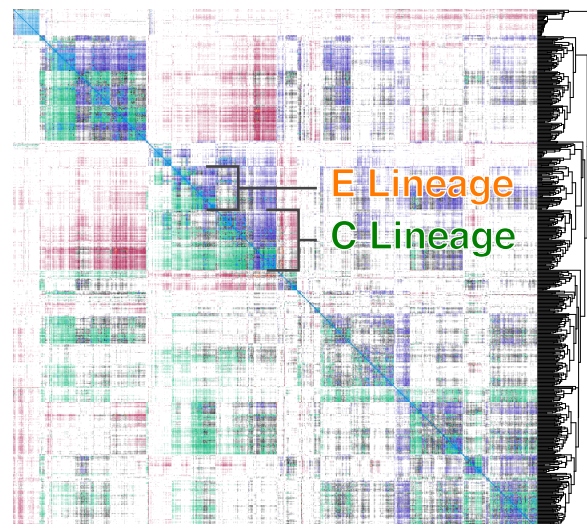




Figure 9: A Glyph SPLOM for all 655 *C. elegans* TFs.

In Figure 9, we demonstrate Glyph SPLOM's ability to scale by applying it to all 655 TF genes measured in this dataset—a dataset too large for a scatterplot matrix. In this figure, the Glyph SPLOM clusters together most of the 14 PAL-1 dependent lineage genes in a C lineage cluster and separates them from genes in a different stem cell lineage, the E lineage. These clusters are separated by the  dependency class rather than by dependency strength; the clusters both consist of positive correlations and are also positively correlated with each other. Thus, heatmaps, which do not consider dependency class, do not separate these clusters well. Together, these examples show that a Glyph SPLOM can visualize relations in data and be applied to generate novel biological hypotheses, even in large datasets with hundreds of variables. We include Figure 9 in the supplementary materials in high resolution and with labels.

To better visualize the statistically significant necessity relations between PAL-1 dependent genes apparent in the Glyph SPLOM, we organize the TFs into a *NecNet* by mapping the necessity relations to network edges as shown in Figure 10. To simplify the network diagram, we preserve the Glyph SPLOM clusters, and edges between clusters are mapped from the most frequent glyph between genes in those clusters. To add direction to edges mapped to significant  glyphs, we use a simplified variant of hierarchical interaction score [SLF* 13] with a single threshold.

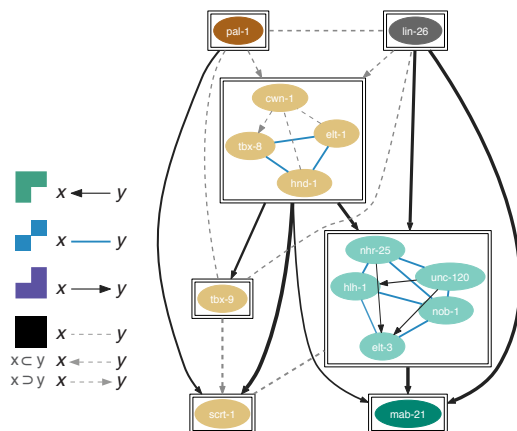


Figure 10: *NecNet* of *C. elegans* PAL-1 genes.

	STIGLER:MIM	NecNet
True Positives (TP)	27	33
True Negatives (TN)	16	19
False Positives (FP)	17	10
False Negatives (FN)	20	33
Accuracy	53%	55%
Precision	57%	77%
Sensitivity	55%	50%
Specificity	48%	66%

Table 2: *NecNet* Vs *Gold Standard*

4.3. Validation

We validated the biological findings produced by Glyph SPLOM with the help of a domain-expert and a co-author (Helen Chamberlin). We validate the biological relevance of our necessity network edges by comparing them with a tentative “Gold Standard” of true regulatory relations and an alternative method (STIGLER:MIM) of inferring a regulatory network using the same data [SC12] (also co-authored by Chamberlin). In Table 2, it is clear that necessity network edges predict true regulatory relations with higher precision and sensitivity than the alternative method.

Preliminary User Study: Further we, conducted an informal study where we asked performance-related questions pertaining to Glyph SPLOM visualization. We invited a total of *sixteen* subjects to participate in our user study. All participants in this preliminary study are either research staff (including a faculty member) or graduate students at our institution of higher learning. The majority of them were computationally adept and possessed a background in data analysis and bioinformatics, while a significant but smaller proportion were well-versed on topics of system and developmental biology. In addition, a majority of them were not familiar with graph-based visualization tools provided by various social networking sites. While they can be considered typical users of our tools, it is certainly the case that *none* of the participants are visualization researchers. In the study, a brief introduction to the topic and the challenges of visualizing high-dimensional data was first presented. The participants were not given access to the manuscript or relevant supplementary material. The subjects were then presented visualizations that included the Glyph SPLOM depicted in Figure 9 and heat maps derived from correlation functions (Pearson, Spearman, and distance correlation) and were asked to assess the effectiveness of visualizations in the context of the questions listed in Section 3.1. We added one more question to the mix; we also asked the subjects to compare the GSPLOM visualization to those obtained from heatmaps derived from correlations.

More than 60% of the participants found the GSPLOM quite useful (Question (i)) especially when compared to heatmaps. The ability to identify relationships was especially noted in a Glyph SPLOM. On the other hand, many noted a loss of significant detail and relatively poor discrimination between clusters in all versions of heatmaps. The participants did note some adverse operational issues when using the Glyph SPLOM. A major complaint was that one had to be reminded of the mappings of individual glyphs. There were also concerns about the use of color hues especially for large dimensional data when the various sub clusters of underlying data are difficult to delineate. Still, even those who were critical of the Glyph SPLOM felt that it potentially offered many benefits over techniques that relied solely on heatmaps especially when the underlying relationships were non-linear and the number of dimensions were very high.

When asked to identify the number of groups or blocks (Question (ii)), the variance in the estimates was smaller with the Glyph SPLOM than other methods. There was a tendency to overestimate with the Glyph SPLOM given its ability to delineate patterns in a SPLOM. On the other hand, the participants underestimated the number of clusters when using the heatmaps. The sizes of the blocks were difficult to estimate when the users were presented heatmap based visualizations (Question (iii)). It was certainly easier for users to complete these two particular tasks when Glyph SPLOMs were used. It should be noted that there was still a significant variance in the estimated sizes of blocks. This deficiency at first glance appears to stem from the difficulty that our choice of colors and the lack of clarity across similarly painted portions of the Glyph SPLOM posed to the participants.

The Glyph SPLOM enabled a majority of the participants to find sub-networks (Question (iv)). This suggests another promising use of the Glyph SPLOM. Although some felt that the heatmap based versions posed some difficulties for larger matrices, they felt that the Glyph SPLOM will fare poorly for smaller dimensional data given the confusion caused by glyph shapes. The role of background knowledge in obtaining insights from various visualizations seemed to be minimal (Question (v)). Finally, when asked to list the strengths and weaknesses of the Glyph SPLOM (Question (vi)), the participants clearly articulated similar comments that have been repeatedly stated here in this section. We plan to employ the valuable feedback from this study to improve the Glyph SPLOM as part of our future efforts such choice of colors and user option on if the glyph shape needs to be presented.

5. Discussion

In this section we discuss issues that are germane to the effective use of Glyph SPLOMs for visualizing multidimensional data. We only use seven classes in the Glyph SPLOM, but other classification schemes with more classes are possible. Currently, we chose these classes for reasons of simplicity and the ease of meaningful interpretation. However, more complicated and nuanced dependency class schemes are possible and we will address this in the subsequent section.

The Glyph SPLOM is also not entirely parameter free. In practice, we find that the default parameter settings work well for most data, but we allow the user to change these parameters to accommodate domain specific applications in our implementation. Finally, while distance correlation is statistically powerful, different types of functions with the same level of noise will produce different (albeit still significant) distance correlations. Therefore, when generating a ranked list of dependencies to inspect, we mitigate this issue by generating separate dCor rankings per logical implication class.

Glyph SPLOMs still suffer from scalability issues when

the number of dimensions is large. Computationally, a Glyph SPLOM is comparable in cost to other heatmap techniques. The most expensive computation in a Glyph SPLOM pertains to the estimation of the all-pairs distance correlation. We have provided an efficient R implementation that computes Glyph SPLOM using matrix computations where possible. In Section 3.5, we described a method to achieve single-pixel summarization for dimensions over 3K. For even larger matrices, representations based on super-pixels will need to be explored. Finally, we only map positive and unsigned glyphs (■ ■ ■ ■ ■) to edges in our Glyph SPLOM to necessity network transformation, but negative glyphs are not currently mapped to any edge. Continued exploration of Glyph SPLOM towards network reconstruction is a topic of future work.

6. Conclusion and Future Work

Glyph SPLOMs provide an intuitive way to increase the scalability of regular SPLOMs. Glyph SPLOM emphasizes interpretability and a compact representation, and it can be applied generally on any data for which a heatmap would be appropriate for a comparable computational expense. Further, the underlying Glyph SPLOM matrix can be transformed to an adjacency matrix of a directed network, leading to more advanced data exploration techniques.

As mentioned earlier, the Glyph SPLOM does suffer from limitations. In some Glyph SPLOM visualization examples it is often hard to delineate quadrants in each scatterplot especially in large submatrices that are all assigned the same logical class. Further, additional visual impairments arise for a large number of dimensions. A better interface with a more clarified color hues and annotations is likely to improve the user experience with a Glyph SPLOM. Appropriate summarization strategies will be pursued in addition to those which trade context and focus. We believe that an interactive Glyph SPLOM would help address its' scalability issues for high-dimensional and large datasets. Such an interactive tool would likely support the navigation of large Glyph SPLOMs using viable context and focus techniques. An interactive Glyph SPLOM tool is a topic of future work. Finally, another interesting avenue for future research is to use grids besides 2x2 to create Glyph SPLOMs. Their interpretation and visualization will pose new challenges.

Acknowledgments: Special thanks to (in alphabetical order) Andrew Fitzgerald, Jacob Hundley, and Mark Wahba of the Department of Computer Science and Engineering for assisting with this research and helping to produce results related to the WHO example application. We also thank Dr. Mohammadmahdi Rezaei Yousefi, and Michael Sharpnack for their comments on this work. Finally this work was partially supported by the following grants: NIH R01CA141090, NSF-DC:SMALL-0916196 and NSF-IIS-1065107.

References

- [AEL*09] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., HOLGER T., MAGNOR M.: Quality-Based Visualization Matrices. In *Proceedings of Vision, Modeling, and Visualization (VMV)* (2009), pp. 341–350. 2
- [BHC*05] BAUGH L. R., HILL A. A., CLAGGETT J. M., HILL-HARFE K., WEN J. C., SLONIM D. K., BROWN E. L., HUNTER C. P.: The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development (Cambridge, England)* 132, 8 (Apr. 2005), 1843–54. 7
- [BN11] BRANDES U., NICK B.: Asymmetric relations in longitudinal social networks. *IEEE transactions on visualization and computer graphics* 17, 12 (Dec. 2011), 2283–90. 3
- [CM84] CLEVELAND W. S., MCGILL R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* 79, 387 (1984), pp. 531–554. 2
- [DWvW12] DINKLA K., WESTENBERG M. A., VAN WIJK J. J.: Compressed Adjacency Matrices: Untangling Gene Regulatory Networks. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2457–2466. 3
- [EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. *IEEE transactions on visualization and computer graphics* 14, 6 (Jan. 2008), 1141–8. 2
- [Fri02] FRIENDLY M.: Corrgrams: Exploratory Displays for Correlation Matrices. *The American Statistician* 56, 4 (Nov. 2002), 316–324. 2
- [HBO10] HEER J., BOSTOCK M., OGIEVETSKY V.: A tour through the visualization zoo. *Communications of the ACM* 53, 6 (June 2010), 59–67. 2
- [HHG12] HELLER R., HELLER Y., GORFINE M.: A consistent multivariate test of association based on ranks of distances. *Biometrika* (Jan. 2012). 3
- [HMM00] HERMAN I., MELANCON G., MARSHALL M.: Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics* 6, 1 (2000), 24–43. 3
- [Kei00] KEIM D.: Designing pixel-oriented visualization techniques: theory and applications. *IEEE Transactions on Visualization and Computer Graphics* 6, 1 (Jan. 2000), 59–78. 2, 5
- [LAE*12] LEHMANN D. J., ALBUQUERQUE G., EISEMANN M., MAGNOR M., THEISEL H.: Selecting Coherent and Relevant Plots in Large Scatterplot Matrices. *Computer Graphics Forum* 31, 6 (Sept. 2012), 1895–1908. 2
- [LMK07] LAM H., MUNZNER T., KINCAID R.: Overview use in multiple visual information resolution interfaces. *Visualization and Computer Graphics, IEEE Transactions on* 13, 6 (2007), 1278–1285. 2
- [MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming overdraw in scatter plots. *IEEE Trans. Vis. Comput. Graph.* 19, 9 (2013), 1526–1538. 2
- [PSC*12] PICCOLO S. R., SUN Y., CAMPBELL J. D., LENBURG M. E., BILD A. H., JOHNSON W. E.: A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* 100, 6 (Aug. 2012), 337–44. 7
- [RN13] REIMHERR M., NICOLAE D. L.: On Quantifying Dependence: A Framework for Developing Interpretable Measures. *Statistical Science* 28, 1 (Feb. 2013), 116–130. 2
- [RRF*11] RESHEF D. N., RESHEF Y. A., FINUCANE H. K., GROSSMAN S. R., MCVEAN G., TURNBAUGH P. J., LANDER E. S., MITZENMACHER M., SABETI P. C.: Detecting novel associations in large data sets. *Science (New York, N.Y.)* 334, 6062 (Dec. 2011), 1518–24. 3, 6
- [Sam69] SAMMON J.: A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers C-18*, 5 (May 1969), 401–409. 2
- [SBSB06] SCHNEIDMAN E., BERRY M. J., SEGEV R., BIALEK W.: Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440, 7087 (Apr. 2006), 1007–12. 3
- [SC12] STIGLER B., CHAMBERLIN H. M.: A regulatory network modeled from wild-type gene expression data guides functional predictions in *Caenorhabditis elegans* development. *BMC Systems Biology* 6, 1 (Jan. 2012), 77. 8
- [SDG*08] SAHOO D., DILL D. L., GENTLES A. J., TIBSHIRANI R., PLEVITIS S. K.: Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome biology* 9, 10 (Jan. 2008), R157. 2, 3
- [SLF*13] SNIDER B., LIBERALI P., FRECHIN M., STOEGER T., PELKMANS L.: Predicting functional gene interactions with the hierarchical interaction score. *Nature methods* 10 (Oct. 2013), 1089–1092. 3, 8
- [SRB07] SZÉKELY G. J., RIZZO M. L., BAKIROV N. K.: Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35, 6 (Dec. 2007), 2769–2794. 2, 3, 4
- [SSB*10] SAHOO D., SEITA J., BHATTACHARYA D., INLAY M. A., WEISSMAN I. L., PLEVITIS S. K., DILL D. L.: MiDRaG: a method of mining developmentally regulated genes using Boolean implications. *Proceedings of the National Academy of Sciences of the United States of America* 107, 13 (Mar. 2010), 5732–7. 3
- [ST11] SIMON N., TIBSHIRANI R.: Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et al., *Science* Dec 16, 2011, 2011. 3
- [TT85] TUKEY J. W., TUKEY P. A.: Computer Graphics and Exploratory Data Analysis: An Introduction. In *Proc. the Sixth Annual Conference and Exposition: Computer Graphics '85, Vol. III, Technical Sessions* (1985), Nat. Computer Graphics Association, pp. 773–785. 2
- [VMCJ10] VIAU C., MCGUFFIN M. J., CHIRICOTA Y., JURISICA I.: The FlowVizMenu and parallel scatterplot matrix: hybrid multidimensional visualizations for network exploration. *IEEE transactions on visualization and computer graphics* 16, 6 (Jan. 2010), 1100–8. 2
- [WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (Oct. 2005), 157–164. 2
- [War94] WARD M.: XmdvTool: integrating multiple methods for visualizing multivariate data. In *Proceedings Visualization '94* (Oct. 1994), IEEE Comput. Soc. Press, pp. 326–333. 2
- [WF09] WILKINSON L., FRIENDLY M.: The History of the Cluster Heat Map. *The American Statistician* 63, 2 (May 2009), 179–184. 2
- [YXZ*07] YAN S., XU D., ZHANG B., ZHANG H.-J., YANG Q., LIN S.: Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1 (Jan. 2007), 40–51. 2