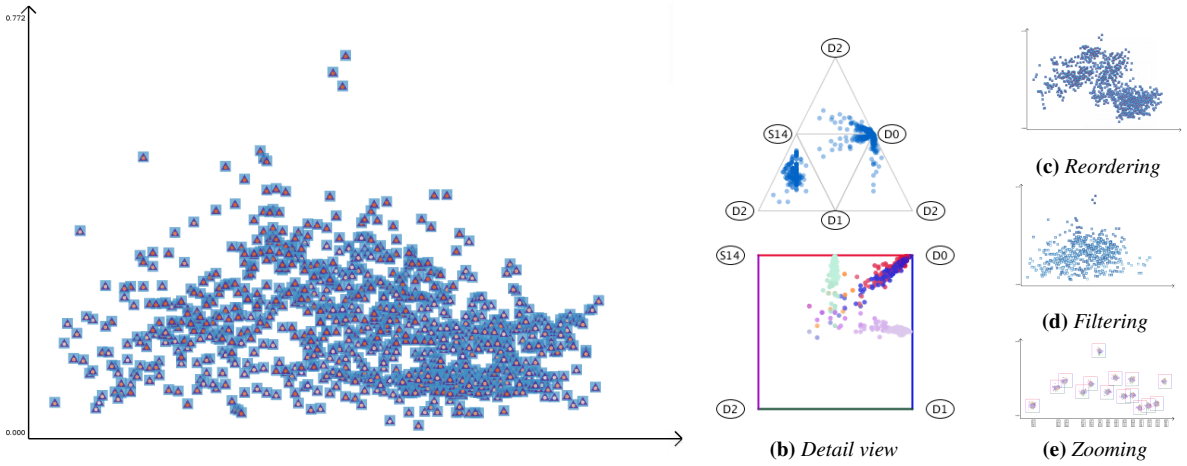
Visual Exploration of Multiway Dependencies in Multivariate Data

Hoa Nguyen*
University of Utah

Paul Rosen†
University of South Florida

Bei Wang‡
University of Utah



(a) *Overview* **(b)** *Detail view* **(c)** *Reordering* **(d)** *Filtering* **(e)** *Zooming*

Figure 1: Overview+detail of the marketing research dataset. Our approach for representing multiway dependencies in multivariate data begins with (a) an overview supported by a glyph representation of all pairwise, 3-way, and 4-way relationships for 4 variables. The overview can be (c) reordered, (d) filtered, (e) zoomed, and individual glyphs can be selected. When selected, the variables of the selected glyph will then populate the detail view (b).

Abstract

Analyzing dependencies among variables within multivariate data is an important and challenging problem, especially when the number of data points is large, the number of variables is high, or multiway dependencies are of interest. Several visualization methods have been proposed to aid in the exploration of such information through the direct visualization of the summary statistics. These methods are typically limited to the study of all possible pairwise relationship but in a manner that does not scale to large multidimensional data. In cases where 3-way relationships are investigated, only subsets of dimensions are considered. In this paper, we propose a novel technique for analyzing multiway dependencies through an overview+detail visualization. In this approach, the overview represents all pairwise, 3-, and 4-way dependencies in the data using glyphs that provide a global visual exploration interface for selecting candidate relationships. Exploration is supported through interactive filtering, sorting, zooming, and selection operations. Once selected, the detailed view helps in developing an inference by providing specific information about those selected variables. Various use cases demonstrate how our approach helps to explore multiway dependencies efficiently in large datasets.

Keywords: multivariate data visualization, variable dependency visualization, correlation visualization, statistical visualization

Concepts: •**Human-centered computing** → **Visualization techniques;**

*e-mail:hoanguyen@sci.utah.edu
†e-mail:prosen@usf.edu
‡e-mail:beiwang@sci.utah.edu

SA '16 Symposium on Visualization, December 05 - 08, 2016, Macao
ISBN: 978-1-4503-4547-7/16/12
DOI: http://dx.doi.org/10.1145/3002151.3002162

1 Introduction

Determining dependency is a task of utmost importance in many fields of science, engineering, and business. For example, in science, extensive parameter space searches can be reduced by understanding which output variables are dependent upon which input variables. In business, dependencies between two or more variables can help managers predict and improve their product positioning. However, this analysis is challenging when the number of potential relationships is large and/or multiway dependencies are of interest. With current techniques, it is infeasible to represent the detail of all possible multiway dependencies.

Several visualization methods have been proposed to aid in the exploration of such information through the direct visualization of summary statistics. For correlation, several approaches have been proposed, including static correlation visualization for large time-varying volume data [Chen et al. 2011], multifield-graphs [Sauber et al. 2006], etc. One major limitation of these approaches is that they represent only pairwise relationships in a single view. UnTangle Map [Cao et al. 2015] proposes a triangle mesh layout, based on a greedy algorithm, to represent sets of 3 variables. This method does not represent all possible relationships but tries to choose the most relevant ones. Unfortunately, the algorithm makes assumptions about which relationships are relevant and does not accommodate situations in which users need to understand all potential multiway relationships. Finally, to the best of our knowledge, no techniques have looked at 4-way relationships.

Therefore, we propose a new interactive statistical visualization tool to effectively perform exploration tasks considering all possible 2- (i.e., pairwise), 3-, and 4-way relationships in the data. For measuring dependencies, we use the coefficient of determination, or R^2, which is a common measure for the fit of a statistical model. Our approach uses an overview+detail style interface with a simple glyph representation guiding the exploration. To aid in exploring the data, we provide a robust set of interactive mechanisms, including selection, filtering, panning, zooming, and animation, to help find relationships of interest. We provide visual encodings of multiway dependencies in a detail view using a point-based representation modified and extended from UnTangle Maps. The resulting approach enables efficient sifting through many multivariate relationships and is capable of supporting large datasets.

In summary, we provide a new interactive visual exploration tool with which users can easily interact and effectively perform statistical dependency tasks. Our tool includes:

- An extension of UnTangle Maps to support 4-way dependency exploration.
- New glyph-based visual encodings for 2-, 3-, and 4-way dependencies, which support flexible investigation.
- An interactive overview with robust interactive mechanisms that represents large numbers of multiway dependencies and enables quick reduction to meaningful relationships.

2 Related Work

Multivariate relationship analysis is an important visual analysis task [Chan et al. 2010]. To gain insight from the complex multivariate data [Keim et al. 2006], a number of analysis approaches have been proposed, such as sampling [Thompson 1992; Chen et al. 2011], clustering [Beham et al. 2014], reducing the number of variables [Jeong et al. 2009], or the introduction of object and dimensional correlation during projection from multidimensional space to 3D [Teoh and Ma 2005]. Many visualization techniques have been proposed to improve correlation identification, but these techniques are not optimally designed for large or high-dimensional data.

SPLOM & Related Techniques. A Scatterplot Matrix (SPLOM) [Hartigan 1975; Huang et al. 2012] shows the relationships of all pairs of variables by organizing a grid of 2D scatterplots. However, each scatterplot must render every data point. This problem can be mitigated by approaches such as Corrgrams [Friendly 2002], which display a matrix of correlation glyphs. Nevertheless, as the number of variables increases, the number of plots grows quadratically, making it difficult to present all of data. The Correlation Coordinates Plots and Snowflake Visualization improve upon this layout [Nguyen and Rosen 2016]. Navigation can also help search larger spaces [Elmqvist et al. 2008]. Another method, based on flow-field analysis and applied to scatterplots, uses sensitivity coefficients to highlight local variation of one variable with respect to another [Chan et al. 2010]. Finally, multivariate data can be projected from their attribute space to 2D, such that points with similar attributes are located close to each other [Janicke et al. 2008].

Parallel Coordinates & Related Techniques. Parallel Coordinates Plots (PCPs) [Inselberg 1985] are another well-known visualization technique for exploring multivariate datasets in a pairwise manner. However, user's ability to infer relationships is often overestimated [Harrison et al. 2014; Kay and Heer 2016]. Various modifications to PCPs, such as using color, opacity, smooth curves, frequency, density, or animation [Heinrich and Weiskopf 2013; Viau et al. 2010; Yuan et al. 2009], have been shown to improve relationship identification over the standard implementation.

Other Multivariate Data Techniques. Many methods use correlation coefficients to calculate relationships among variables in data. Gosink et al. [Gosink et al. 2007] present a method that increases the utility of query-driven techniques by visually conveying statistical information about the trends that exist between variables in a query. In this method, correlation fields, created between pairs of variables, are used with the cumulative distribution functions of variables expressed in a user's query. Qu et al. [Qu et al. 2007] used the correlation coefficient to calculate the strengths between different data variables in weather data analysis and visualization. Glatter et al. [Glatter et al. 2008] used two-bit correlation to study temporal patterns in large multivariate data. Sukharev et al. [Sukharev et al. 2009] proposed a method based on analyzing pairwise correlation in time-varying multivariate data by using point-wise correlation coefficients and canonical correlation analysis. Another pairwise correlation visualization approach used local anisotropic correlation structures in the vicinity of uncertain isosurfaces and used glyphs to visualize these dependencies [Pfaffelmoser et al. 2013]. Jen introduced a design for exploring correlations between two scalar fields [Jen et al. 2004]. Some methods have used data mining techniques to gain insight. Gu and Wang presented three hierarchical clustering methods based on quality threshold, k-means, and random walks to investigate the correlations with varying levels of detail [Gu and Wang 2010].

Large Data Techniques. Several approaches deal with large and complex correlation fields. The Multifield-Graph is used to give an overview of how multiple fields correlate and to show the strength of their correlation [Sauber et al. 2006]. The core of their approach is the computation of correlation fields, which are scalar fields containing the local correlations of subsets of the multiple fields. [Chen et al. 2011] also introduced a sampling scheme to summarize the correlation connection in time-varying multivariate datasets. This scheme consists of three steps: selecting important samples from the volume, prioritizing distance computation for sample pairs, and approximating volume-based correlation. This sample-based approach enables users to obtain an approximate correlation coefficient in a cost-effective manner, making it scalable for large datasets. Furthermore, [Nagaraj et al. 2011] proposed a multifield comparison measure for scalar fields that helps in studying relations between them. The comparison measure is insensitive to noise in the scalar fields and to noise in their gradients. Additionally, [Liu and Shen 2016] proposed a novel association analysis method that guides visual exploration of scalar-level associations in the multivariate context. They model directional interactions between scalars of different variables as information flows to explore the scalars of interest with confident associations in the multivariate spatial domain, and provide guidelines for visual exploration.

UnTangle Map. UnTangle Map [Cao et al. 2015] is an effective way to investigate the relationships between data items and their probabilistic labels, as well as the relationships among labels. The design extends the traditional ternary plot, useful for pairwise and 3-way relationship finding, into an interactive mesh of triangles in order to effectively show item-label relationships, and to enable the scattering patterns of items to aggregate into a visual summary of the underlying labels. However, this design has some limitations. First, it does not represent 4-way relationships, nor it is obvious how to extend the approach to higher dimensional relationships. The mesh is laid out in a greedy manner that requires interaction when all relationships have been explored. Furthermore, with large numbers of dimensions, either the plots need to shrink or a smaller

percentage of relationships will be shown.

All the approaches reviewed here assist in investigating dependencies in multivariate data. However, these techniques have limitations: first, they are limited in the dimensionality of relationship (to either pairwise or 3-way relationships); second, they are limited by the number of data points they can visualize; and finally, they are limited by the number of relationships they can display simultaneously. The goal with our approach is to address these limitations.

3 Multivariate Dependency Modeling

3.1 Pairwise Statistical Correlation

Pairwise relationships can be measured by a wide variety of techniques. Here, we focus on the Pearson Correlation Coefficient [Benesty et al. 2008; Ke et al. 2008; Magnello and Vanloon 2009; Wang and Zheng 2013] and Spearman Rank Correlation Coefficient [Hogg and Craig 1995; Hogg and Craig 1998], although our technique can generalize to other measures.

The most common correlation measure is the Pearson Correlation Coefficient (PCC). The PCC, $\rho(x,y)$, measures the linear relationship between two variables x and y with means $\bar{x}$ and $\bar{y}$ and standard deviations σ_x and σ_y. It is defined as:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad (3.1.1)$$

PCC makes two important assumptions about the data. First, it assumes a linear relationship. However, finding nonlinear relationships can be important [Chen et al. 2010]. Second, data must be approximately normally distributed with no significant outliers.

The Spearman Rank Correlation Coefficient (SRCC) is the nonparametric version of the PCC that measures the strength of association between two ranked variables. The rank or order Rx of the data points x is calculated for each variable independently. Then, the PCC of the ranked variables is calculated as $PCC(Rx,Ry)$.

3.2 Multivariate Dependency and the Coefficient of Determination

The coefficient of determination, or R^2, is used to measure how well the data fit a model [Allison 1998; Keith 2006]. Our usage of the measure is under the context of multiple correlation, which is a measure of how well a given (dependent) variable can be predicted using a linear combination of other (independent) variables. The value of R^2 ranges between 0 and 1, where a higher value indicates better predictability of the dependent variable. A value of 1 indicates that the independent variables can perfectly predict the dependent variable, and a value of 0 indicates that no linear combination of the independent variables is a better predictor than the fixed mean of the dependent variable.

Multiple correlation requires the selection of a set of independent variables, $x_1, x_2, ..., x_N$, and a single dependent variable, y. R^2 can then be computed using the following equation:

$$R^2 = \mathbf{c}^T R_{xx}^{-1} \mathbf{c} \qquad (3.2.1)$$

The correlation matrix R_{xx} represents the inter-correlations between independent variables. The vector $\mathbf{c}$ contains the pairwise correlation $r_{x_i y}$ between the independent variables x_i and the dependent variable y. They take the form:

$$R_{xx} = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} & \cdots & r_{x_1x_N} \\ & \ddots & & \vdots \\ r_{x_2x_1} & & & \\ \vdots & & \ddots & \\ r_{x_Nx_1} & \cdots & & r_{x_Nx_N} \end{pmatrix}, \mathbf{c} = \begin{pmatrix} r_{x_1y} \\ r_{x_2y} \\ \vdots \\ r_{x_ny} \end{pmatrix} \qquad (3.2.2)$$

If all the independent variables are uncorrelated, the matrix R_{xx} is the identity matrix and R^2 simply equals $\mathbf{c}^T\mathbf{c}$, the sum of the squared correlations with the dependent variable. If the predictor variables are correlated among themselves, R_{xx}^{-1} will account for this.

4 Visual Design of Multiway Dependencies

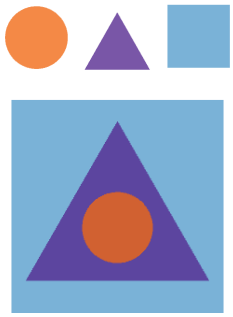
The visualization of dependencies for multivariate data is challenging due to the sheer number of potential relationships. For a given dataset of n variables, the number of dependency relationships is $\binom{n}{2}$, $3 * \binom{n}{3}$, and $4 * \binom{n}{4}$ for 2-, 3-, and 4-way relationships, respectively. For a dataset of 20 variables, for example, there are 190 2-way, 3420 3-way, and 19,380 4-way relationships. A static display showing detailed information about all potential variable relationships for such multiway dependencies may be overwhelming, confusing, and difficult to make any judgment upon. Therefore, we develop a multiscale overview+detail design that initially visualizes summaries of all relationships but provides a variety of interactions to filter and investigate the details surrounding interesting variable combinations.

4.1 Overview Design

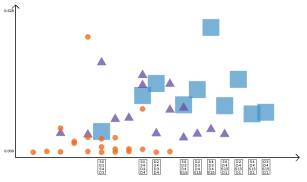
The premise of our design is quite simple: display as many summaries of relationships as possible, while providing the ability to sort, filter, and investigate their corresponding details.

4.1.1 Individual Dependency Glyphs

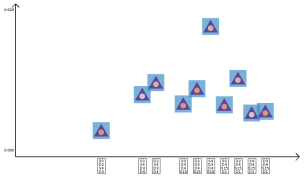
Our first design goal is to represent all possible dependencies on a single interface. First, the three dependency types can be represented as different glyph shapes and colors. A 2-way relationship is represented with an orange circular glyph (top left of Figure 2a). A 3-way dependency is represented as a purple triangle (top middle of Figure 2a). A 4-way relationship is represented by a light blue squares (top right of Figure 2a). These simple visual encodings can be embedded in the overview such that both the x- and y-axis



(a) *Visual encodings, circle, triangle, square, for 2-, 3-, and 4-way dependencies, respectively. Glyphs are shown individually (top) and composited (bottom).*



(b) *Separate glyph overview.*



(c) *Composite glyph overview.*

Figure 2: Glyphs used to represent multiway dependencies in the overviews.

are controlled independently, such as in Figure 2b. For each axis, the user may select metrics, including dimension sorting through ordered permutations, R^2, and time (for time series data).

4.1.2 Multiple Dependency Glyphs

Given the large number of potential relationships, we are interested in designing a glyph to reduce the visual clutter.

Consider a set of four variables. Among these variables, for any two, the dependency is symmetric (i.e., either variable may be the dependent variable). Therefore, there are $\binom{4}{2}$ or 6, possible 2-way relationships. For possible 3-way dependencies, each of the four variables can be dependent to $\binom{3}{2}$ or 3, combinations of the other variables. That means a total of 12 3-way dependencies. Finally, each variable can be dependent to all others in a 4-way dependency, for a total of four 4-way dependencies. Among four variables, 22 dependencies exist, which summarize all possible relationships.

To visually represent all potential relationships among these variables, we can composite the glyphs from Figure 2a top into the glyphs seen in Figure 2a bottom. Now, this glyph summarizes 22 dependencies—the circle summarizes 6 2-way dependencies, the triangle summarizes 12 3-way dependencies, and the square summarizes 4 4-way dependencies. Additional information is provided by modifying the color of the glyphs based upon the average R^2 score of the corresponding dependencies. The solid color represents $R^2 = 1$ and white represents $R^2 = 0$.

The glyphs are positioned in such a way that both the x- and y-axes can be controlled independently. The user may select metrics, including ordered permutations of dimensions, R^2_{min}, R^2_{max}, R^2_{avg} and time. R^2_{min}, R^2_{max}, and R^2_{avg} are calculated by performing the associated operations on all dependencies represented by the glyphs.

4.1.3 Interactions

Ordering. Recall the user may select different metrics for sorting, including ordered permutations of dimension, R^2_{min}, R^2_{max}, R^2_{avg} and time. Switches among the sorting metrics are handled through animation to maintain context.

Individual vs. Composite Glyphs. Users have the option to switch between the composite glyphs (Figure 2c), and the individual relationship glyphs (Figure 2b). They may also limit the relationships of interest (e.g. only 2- and 3-way relationships). Similar to the sorting operation, when switching configurations, glyphs are animated to maintain context.

Filtering. We provide an upper and a lower threshold for filtering the R^2 scores of glyphs, such that users can reduce the volume of data to be visualized and find meaningful relationship glyphs. Users can raise/decrease the threshold when they want to identify glyphs that represent stronger/weaker dependencies, respectively.

Navigation. Users can zoom and translate the projection space to navigate and explore variable dependencies. The size of the glyphs changes based upon the number of visible glyphs. When more than $1M$ glyphs are displayed on the screen, each is replaced by a point (Figure 8a). When the number of glyphs is small enough, a 4-way detail view (explained in the following section) is shown (Figure 1e). Otherwise, they appear as composite glyphs.

Selection. We provide three selection mechanisms. The first mechanism allows the user to select which variables to include or exclude from the analysis. The second allows users to create a selection box or lasso around a region of interest, and the associated zoom and translate operations are updated. The final mechanism selects an individual glyph. Once selected, its corresponding set of relationships is highlighted in the detail view.

4.2 Detail View Design

When a glyph is selected in the overview, the detail view is updated to provide details of the four variables represented by the glyph.

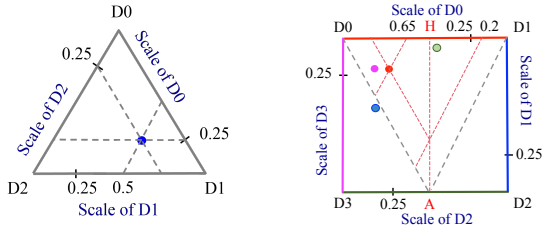
4.2.1 Visual Encoding Design

Our approach uses an extended version of UnTangle Maps [Cao et al. 2015] to represent the relationships. Untangle Map represents 2- and 3-way dependencies using a ternary plot as shown in Figure 3a. This is a barycentric plot of three variables with each at a vertex, and the three variables $D0$, $D1$, $D2$ are vertices of the triangle. When an item is associated with the three variables with varying probabilities, such probability information presents the detailed relationships among the three variables. In particular, if an item contains values for the three variables D_i as v_i respectively (for $i \in \{0,1,2\}$), the probability of this item being D_i is $v_i/(v_0+v_1+v_2)$. For example, the item i (blue point) is associated with $D0$, $D1$, and $D2$ with probabilities 0.25, 0.5, and 0.25. The position of the point closer to $D1$ indicates this higher probability.

The standard UnTangle Map display does not directly provide information of 4-way dependencies. Understanding 4-way dependencies in UnTangle Maps requires mentally stitching the information of 4 plots together. Instead of identifying the relationship in four different plots, we propose a new representation that builds on the previous 3-way design by replicating, rotating, and overlapping the ternary plots to reveal the 4-way relationship.

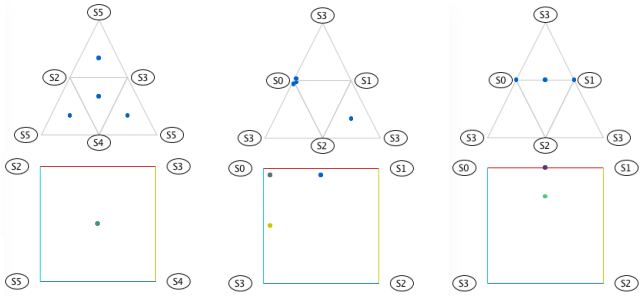
In this design, the relationship is broken up into four triangle plots contained within a square. Each edge of the square $(D0,D1)$, $(D1,D2)$, $(D2,D3)$, and $(D3,D0)$ has been colored red, blue, green, and purple, respectively. Each data point is broken up into four components of the corresponding color. Each component (i.e. a colored point) is placed in the square using a ternary plot made up of the edge vertices and the midpoint of the opposite edge (e.g. edge $(D0,D1)$ with point A in Figure 3b). A colored point is placed in this plot using the two vertex variables and the sum of the probabilities of the other two variables.

For example, consider that a data point in Figure 3b has a probability of $D0$, $D1$, $D2$, and $D3$ as 0.65, 0.2, 0.1, and 0.05, respectively. The ternary plot for $(D0,D1)$ is highlighted with the red and gray dashed lines. The red point (corresponding to $(D0,D1)$) is placed based upon the probability 0.65, 0.2, and 0.15 $(0.1+0.05)$. $(D1,D2)$ (blue point) is placed with probability 0.2, 0.1, 0.7. $(D2,D3)$ (green) is placed with probability 0.1, 0.05, 0.85. Finally, $(D3,D0)$ (purple) is placed with probability 0.05, 0.65, 0.3.



(a) *2- & 3-way mapping* **(b)** *4-way mapping*

Figure 3: Mapping for 2-, 3-, and 4-way dependencies.



(a) *Non-dominant* **(b)** *Uni-dominant S0* **(c)** *Bi-dominant S0/S1*

Figure 4: Visual patterns for non-, uni-, and bi-dominant relationships.

4.2.2 Visual Patterns

In both the standard and our extended version of UnTangle Map, the important visual pattern is proximity to a vertex or an edge. Proximity to a vertex indicates dominance of a single variable. Proximity to an edge indicates dominance of two variables. Some of these visual patterns can be seen in the examples in Figure 4. When no variable shows dominance (Figure 4a), all points are centered in the triangle or square. When a single variable is dominant (Figure 4b), the points focus around vertex S0. In bi-dominant relationship (Figure 4c), the points focus between two vertices, S0 and S1.

5 Evaluation

To evaluate our approach, we apply our method to four datasets: a product marketing dataset with 47 variables, a particle physics dataset with 66 variables, the National Health and Aging Trends Study (NHATS) dataset with 60 variables, and the Hurricane Isabel dataset with 13 variables over 48 time steps.

5.1 Performance

We build our software using Processing. We have run our experiments on a variety of desktop and laptop systems running Linux, MAC OSX, and Windows.

The visualization rendering itself is interactive. Assume that the dataset has n variables and each variable has k data points. For the overview, our visualization represents $\binom{n}{2}+3*\binom{n}{3}+4*\binom{n}{4}$ multiway dependencies through $\binom{n}{4}$ glyphs. We have tested our approach up to $n=624$, and the system has maintained its interactivity. Rendering the detail view is dependent upon the number of data points. Each point needs to be rendered 8 times: 4 times for the 3-way UnTangle map and 4 times for our 4-way UnTangle map extension. Therefore, the total number of points rendered is $8k$.

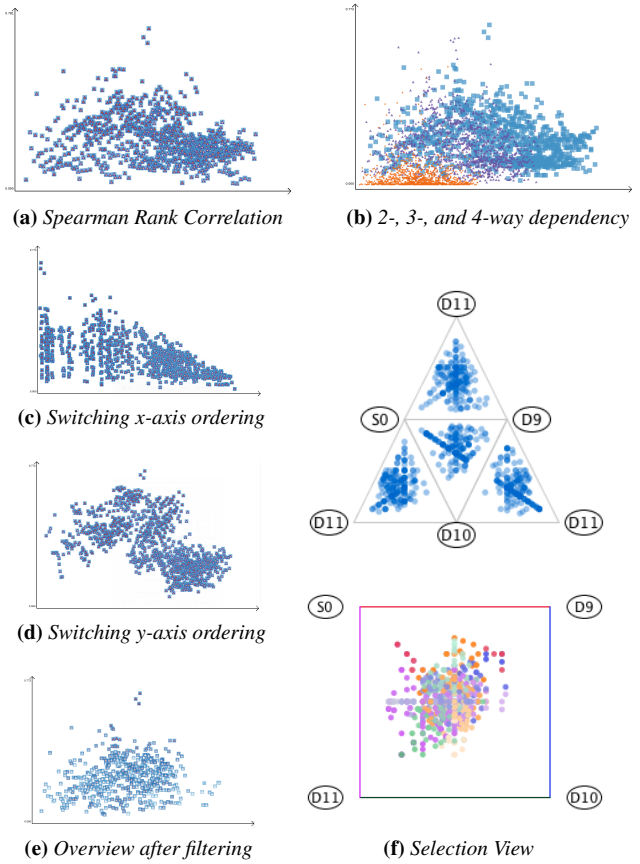
The main computational challenge is the precomputation needed for determining dependencies, in particular, the pairwise correlation coefficients. Computing Pearson Correlation Coefficients and Spearman Rank Correlation Coefficients takes $O(n^2k)$ and $O(n^2klog(k))$, respectively. Computing the Coefficient of Determination for Multiple Correlation, R^2 for 2-, 3-, and 4-way dependency takes dependency takes $O(n^2)$, $O(n^3)$, and $O(n^4)$, respectively. Therefore, this approach has an aggregate computing time of $O(n^2k+n^4)$ or $O(n^2klogk+n^4)$. In general, $k \gg n$, leading to the pairwise computation being the bottleneck. Fortunately, much of the computation is embarrassingly parallel, and is parallelized in our implementation.

5.2 Marketing Research Case Study

Marketing research data, often collected via surveys, is used to identify groups of individuals who might best be served by a particular product design. In this case, we use the Pacific Brands/Berlei Bras case study data, which is commonly used in business school marketing courses. Marketing researchers divide their questions into two types. First, segmentation variables, such as age, sex, income, etc., are used to differentiate groups of people (i.e. independent variables). Second are discriminant (i.e. dependent) variables, which are qualitative, such as feelings about color, texture, etc.

This dataset contains 21 segmentation variables and 26 discrimination variables, that is, a total of 47 variables with 1,081 2-way dependencies, 48,645 3-way dependencies, and 713,460 4-way dependencies. This requires 178,365 glyphs to represent all multiway dependencies.

First, after loading the data into the system, the overview is shown (Figure 1) with the Pearson Correlation Coefficient. Optionally, the Spearman Rank Correlation Coefficient can be selected (Figure 5a) when a non-parametric view of dependency is more appropriate. To understand 2-, 3-, and 4-way dependencies separately, these options are selected and animations are used to highlight their transitions into new positions (Figure 5b). The order of points can be modified along the x-axis (Figure 5c), y-axis (Figure 5d), or both (Figure 1c). In these cases, the x-axis is switched to R^2_{max} and the y-axis is switched to R^2_{avg}, with animation connecting the transitions. It is clear from many of these views that most dependencies are weak. A filter on $R^2 \in [0.6, 1.0]$ significantly reduces the number of relationships to explore (Figure 1d).



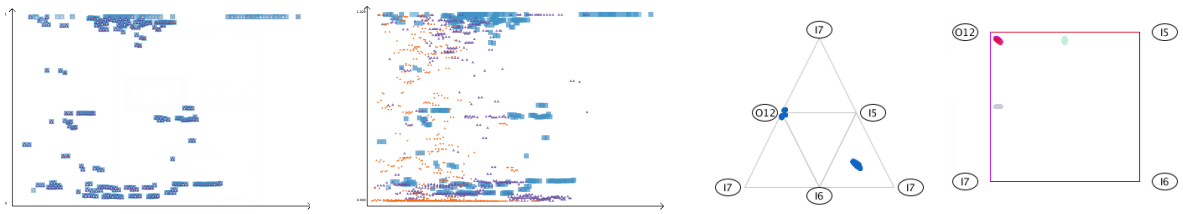
(a) *Spearman Rank Correlation* **(b)** *2-, 3-, and 4-way dependency*

(c) *Switching x-axis ordering*

(d) *Switching y-axis ordering*

(e) *Overview after filtering* **(f)** *Selection View*

Figure 5: A variety of overviews and one detail view of the marketing research dataset.



(a) *Composite glyph overview* **(b)** *2-, 3-, & 4-way glyphs overview* **(c)** *Detail view*

Figure 6: Two overviews and one detail view of the physics data.

After some navigation and exploration, a smaller number of glyphs occupy the screen to highlight their corresponding relationships (Figure 1e), which helps to quickly identify the strengths and directions among the relationships.

In the detail view in Figure 5f, the selected glyph contains variables $S0, D9, D10$, and $D11$:

- *S0*: I am very conscious of bras as fashion objects.
- *D9*: I like to shop in the same lingerie stores as my friends.
- *D10*: I use other people as a source of information for purchase decisions.
- *D11*: I use magazines or newspapers as a source of information for purchase decisions.

Figure 5f shows that some data points move toward *S0* but most of the data points are in the middle. There is no point towards *D10*, which indicates that *S0* is weakly dominant in the 4-way relationship. Figure 5f also shows that there are no points around *D10* in 3-way and 4-way dependencies. *D10* is less dependent upon other variables (*S0*, *D9*, *D11*). This shows that to design bras fashion objects, information of friends shopping destination, and magazines or newspapers are a good source of information, since *S0*, *D9*, *D11* are highly correlated. This previously unknown combination of opinions helps to quickly identify groups of individuals who are best served by a particular product design. The result might lead marketers to choose a particular design or advertising campaign.

5.3 Particle Physics Case Study

The physics dataset represents a parameter space search in simulations that model subatomic particles under the supersymmetric extension of the Standard Model. The data has 25 input and 41 output variables with $4,000$ items for each variable, which leads to $2.8M$ 4-way dependencies, $137k$ 3-way dependencies, and $2,145$ pairwise dependencies, for a total of $3M$ dependencies. We require $720k$ glyphs to represent all these relationships.

Determining dependency can be valuable in reducing the size of a parameter search space by linking input and output variables together. Many glyphs visible near the top of the overview coordinates in Figure 6a show that the variables of the physics data have strong dependencies. Users can confirm that this is a combination of 2-, 3-, and 4-way dependencies by separating the glyphs in Figure 6b. The overview of composite glyphs and 2, 3, and 4-way separated glyphs help us understand that this data has many dominant and strong relationships, since many glyphs are on the top of the plot.

These variables are input and output variables of the simulation. The expert would like to understand which inputs are correlated with which outputs. The expert is also interested in which inputs most strongly reflect linear correlation with a given output. With our tool, the expert can easily interact with various dependencies and perform the analysis tasks efficiently.

The expert can quickly select an interesting input/output variable, and the layout will automatically show variables that are correlated to the selected variable. For example, the detail view of the selected glyph in Figure 6c enables the expert to quickly identify the dependencies from selected variables (including input *I5*, *I6*, *I7*, and output *O12*). This shows that variable *O12* is highly correlated with others, and it is dominant.

Using UnTangle Map alone to answer the above questions would have required adding many dimensions to the layout and exploring one by one which inputs and outputs are correlated. However, by using our proposed visualization approach, the expert can quickly select the interesting input/output in the data, filter the layout and show only correlated dimensions.

5.4 National Health and Aging Trends Study (NHATS)

The National Health and Aging Trends Study (NHATS) includes data collection research being conducted by Johns Hopkins Bloomberg School of Public Health. The goal is to "foster research that will guide efforts to reduce disability, maximize health and independent functioning, and enhance quality of life at older ages". NHATS collects detailed information on activities and quality of life for a sample of Medicare beneficiaries over 65.

We explore a subset of the NHATS data that has 60 variables with $38k$ items. This data has $2M$ 4-way dependencies, $100k$ 3-way dependencies, and $1,770$ pairwise dependencies, for a total of $2.15M$ dependencies. We require $487k$ glyphs to represent these relationships.

Figure 7 shows an example analysis of the NHATS data. Figure 7a and 7b show the overview with composite and split glyphs for the data. It is immediately apparent that many relationships have low R^2 values, while a few have high max R^2. Using the lasso tool (Figure 7c) filters data down to a subset (Figure 7d).

After exploration, a specific relationship is investigated. Figure 7e shows the detail view of variables *d45*, *d46*, *d47*, *d48*. The 4 possible cases of the question "Is [Caretaker Name] paid by you (*d45*), your/his/her family, by a government program (*d46*), or by your/his/her insurance (*d47*) or other (*d48*)?". The centrality of these points in the square shows that these four variables have a non-dominant (uncorrelated) relationship. This makes sense, as the four options should be mutually exclusive cases of payment.

5.5 Hurricane Data Case Study

Finally, we explore the IEEE Visualization 2004 Hurricane Isabel contest dataset. It consists of 48 time steps, measuring 13 variables with a spatial resolution of $500 \times 500 \times 100$ ($25M$ points per time step). Combining all variables over all time steps leads to an exploration of 624 total variables (i.e. 13 variables x 48 time steps). This data has $25B$ 4-way dependencies, $250M$ 3-way dependencies and $194k$ 2-way dependencies, requiring $6B$ glyphs.

Figure 8 shows an example analysis. Figure 8a shows an overview of all dependency features of the 624 variables. The many points at the bottom of the chart show weak dependencies, yet patterns of strong dependencies are still visible. For example, repeated pattern between *QRAIN* and *QSNOW* variables is seen in the middle of chart. To investigate further, the view is filtered by selecting the *QRAIN*/*QSNOW* variables (Figure 8b). Further zooming onto *QRAIN*/*QSNOW* in Figure 8c shows more detailed glyphs that can be individually inspected.

The relationships can also be sorted by time horizontally, by variable name vertically, and filtered by R^2 (Figure 8d). Figure 8e shows the relationships sorted by time horizontally and R^2 average vertically. Noticing inconsistency in the *CLOUD* variable in Figure 8d warrants further investigation. Using a selection box the *CLOUD* glyphs are isolated in Figure 8f. In this view, a number of conclusions can be drawn. For example, this confirms that the relationship between *CLOUD* and *Pressure* (*P*) are not consistent over time. Similarly, the relationship between *CLOUD* with *QGRAUP* is not consistent from time steps 10 to 20.

With over 25B 2-, 3-, and 4-way dependencies, the Hurricane Isabel data is large and impractical to explore completely. Our approach, enables quickly reducing the variables of interest. Without our approach the relationship between *CLOUD* and *Pressure* might not be isolated for analysis, but it is clear through our visualization that they are not consistent over time.

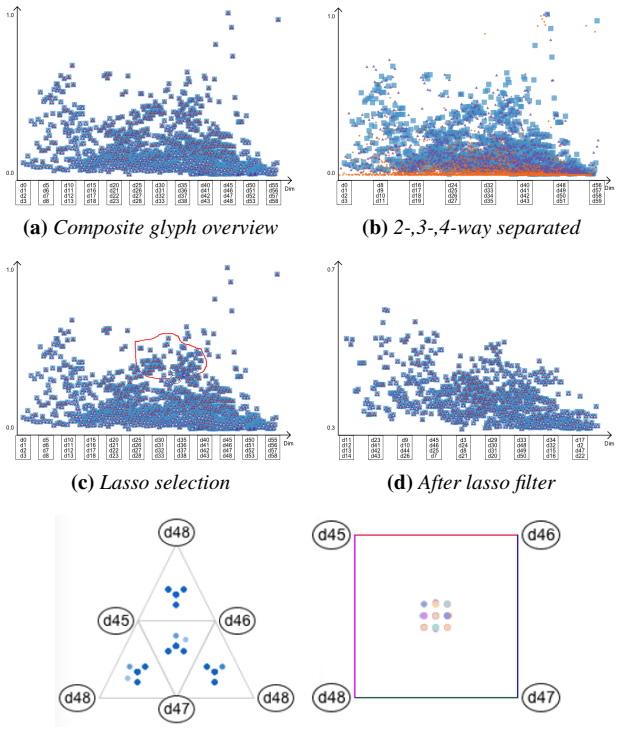
6 Discussion and Conclusions

We have proposed a method that visualizes multiway dependencies from multivariate data. Previous work has focused on 2-way or 3-way correlations. UnTangle Map can represent only 2-way or 3-way dependencies. We propose a new glyph-based visualization for high-dimensional data that includes an extension to UnTangle for 4-way dependencies. The combination of these designs and filtering/selection interactions provides a powerful visual exploration mechanism that is intuitive and effective.

Our approach is scalable to both the number of variables and the size of data, as demonstrated by the Hurricane Isabel dataset, which contains hundreds of variables and millions of points. Few other approaches have attempted to analyze this number of variable dependencies. The practical limit of our approach probably lies in the range of 500-1000 variables.

We chose to limit our approach to 4-way dependencies for a number of reasons. First, the number of 5-way relationships is huge, e.g., $\binom{624}{5} = 775$ billion. Second, as the number of independent variables grows, there is a naturally increasing coefficient of determination (i.e. more input variables are more likely to explain an output variable). Nevertheless, most of our visual encodings could be extended to 5-way dependencies.

Finally, our approach uses the R^2 coefficient of determination for multiple correlation with the Pearson Correlation Coefficient and Spearman Rank Correlation Coefficient. Many other statistical models could be used in place of the coefficient of determination, depending on the requirements of the analysis.

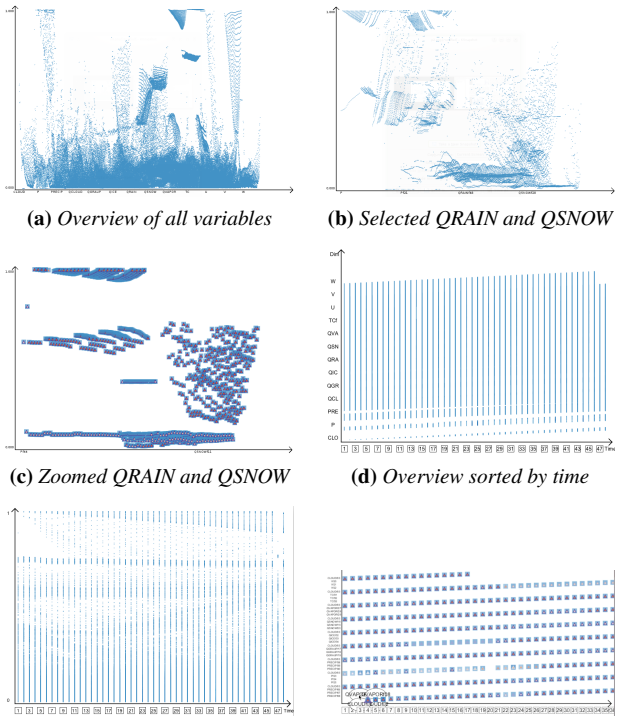


(a) *Composite glyph overview* **(b)** *2-,3-,4-way separated*

(c) *Lasso selection* **(d)** *After lasso filter*

(e) *Detail view of data*

Figure 7: A variety of overviews and a single detail view of the NHATS data.



(a) *Overview of all variables* **(b)** *Selected QRAIN and QSNOW*

(c) *Zoomed QRAIN and QSNOW* **(d)** *Overview sorted by time*

(e) *Overview sorted by time & R^2* **(f)** *Selection of CLOUD*

Figure 8: Hurricane Isabel data by variable series (a-c) and timeline series (d-f).

Acknowledgments

This work was supported in part by the National Science Foundation (III-1513616 and ACI-1443046).

References

ALLISON, P. D. 1998. *Multiple Regression: A Primer*. Sage Publications.

BEHAM, M., HERZNER, W., GRÖLLER, M. E., AND KEHRER, J. 2014. Cupid: Cluster-based exploration of geometry generators with parallel coordinates and radial trees. *IEEE TVCG 20*, 12, 1693–1702.

BENESTY, J., CHEN, J., AND HUANG, Y. 2008. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Trans. on ASLP 16*, 4, 757–765.

CAO, N., LIN, Y.-R., AND GOTZ, D. 2015. Untangle map: Visual analysis of probabilistic multi-label data. *IEEE TVCG PP*, 99, 1–1.

CHAN, Y.-H., CORREA, C. D., AND MA, K.-L. 2010. Flow-based scatterplots for sensitivity analysis. In *IEEE VAST*, 43–50.

CHEN, Y. A., ALMEIDA, J. S., RICHARDS, A. J., MULLER, P., CARROLL, R. J., AND ROHRER, B. 2010. A nonparametric approach to detect nonlinear correlation in gene expression. *Journal of CGS 19*, 3, 552–568.

CHEN, C.-K., WANG, C., MA, K.-L., AND WITTENBERG, A. T. 2011. Static correlation visualization for large time-varying volume data. *PacificVis*, 27–34.

ELMQVIST, N., DRAGICEVIC, P., AND FEKETE, J.-D. 2008. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE TVCG 14*, 6, 1141–1148.

FRIENDLY, M. 2002. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician 1*.

GLATTER, M., HUANG, J., AHERN, S., DANIEL, J., AND LU, A. 2008. Visualizing temporal patterns in large multivariate data using modified globbing. *IEEE TVCG 14*, 6, 1467–1474.

GOSINK, L., ANDERSON, J. C., BETHEL, E. W., AND JOY, K. I. 2007. Variable Interactions in Query-Driven Visualization. *IEEE TVCG 13*, 6, 1400–1407.

GU, Y., AND WANG, C. 2010. A study of hierarchical correlation clustering for scientific volume data. In *Advances in Visual Computing*, 437–446.

HARRISON, L., YANG, F., FRANCONERI, S., AND CHANG, R. 2014. Ranking visualizations of correlation using weber's law. *IEEE TVCG 20*, 12, 1943–1952.

HARTIGAN, J. A. 1975. Printer graphics for clustering. *Journal of Statistical Computation and Simulation 4*, 3.

HEINRICH, J., AND WEISKOPF, D. 2013. State of the art of parallel coordinates. In *Eurographics STAR*, 95–116.

HOGG, R. V., AND CRAIG, A. T. 1995. *Introduction to Mathematical Statistics, 5th ed.* Macmillan.

HOGG, R. V., AND CRAIG, A. T. 1998. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall.

HUANG, T.-H., HUANG, M. L., AND ZHANG, K. 2012. An interactive scatter plot metrics visualization for decision trend analysis. In *Int'l Conf. on Mach. Learning & App.*, 258–264.

INSELBERG, A. 1985. The plane with parallel coordinates. *Visual Computer 1*, 2, 69–91.

JANICKE, H., BOTTINGER, M., AND SCHEUERMANN, G. 2008. Brushing of attribute clouds for the visualization of multivariate data. *IEEE TVCG 14*, 6, 1459–1466.

JEN, D., PARENTE, P., ROBBINS, J., WEIGLE, C., TAYLOR, R., BURETTE, A., AND WEINBERG, R. 2004. Imagesurfer: a tool for visualizing correlations between two volume scalar fields. In *IEEE VIS*, 529–536.

JEONG, D. H., ZIEMKIEWICZ, C., FISHER, B., RIBARSKY, W., AND CHANG, R. 2009. ipca: An interactive system for pca-based visual analytics. In *EuroVis*, 767–774.

KAY, M., AND HEER, J. 2016. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE TVCG 22*, 1, 469–478.

KE, Y., CHENG, J., AND NG, W. 2008. Efficient correlation search from graph databases. *IEEE Trans. Know. & Data Eng. 20*, 12, 1601–1615.

KEIM, D. A., MANSMANN, F., SCHNEIDEWIND, J., AND ZIEGLER, H. 2006. Challenges in visual data analysis. In *IEEE InfoVis*, 9–16.

KEITH, T. 2006. *Multiple Regression and Beyond*. Pearson Education.

LIU, X., AND SHEN, H. W. 2016. Association analysis for visual exploration of multivariate scientific data sets. *IEEE TVCG 22*, 1 (Jan), 955–964.

MAGNELLO, E., AND VANLOON, B. 2009. *Introducing Statistics: A Graphic Guide*. Icon Books Ltd.

NAGARAJ, S., NATARAJAN, V., AND NANJUNDIAH, R. S. 2011. A gradient-based comparison measure for visual analysis of multifield data. In *EuroVis 2011*, 1101–1110.

NGUYEN, H., AND ROSEN, P. 2016. Improved identification of data correlations through correlation coordinate plots. In *Int'l Conf. on Info. Vis. Theory & App.*

PFAFFELMOSER, TOBIAS, AND WESTERMANN. 2013. Correlation visualization for structural uncertainty analysis. *International Journal for Uncertainty Quantification 3*, 2.

QU, H., CHAN, W.-Y., XU, A., CHUNG, K.-L., LAU, K.-H., AND GUO, P. 2007. Visual analysis of the air pollution problem in hong kong. *IEEE TVCG 13*, 6, 1408–1415.

SAUBER, N., THEISEL, H., AND SEIDEL, H. 2006. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE TVCG 12*, 5, 917–924.

SUKHAREV, J., WANG, C., MA, K., AND WITTENBERG, A. T. 2009. Correlation study of time-varying multivariate climate data sets. In *PacificVis*, 161–168.

TEOH, S., AND MA, K.-L. 2005. Hifocon: Object and dimensional coherence and correlation in multidimensional visualization. In *Advances in Visual Computing*, vol. 3804, 235–242.

THOMPSON, S. 1992. *Sampling*. John Wiley, Sons, Inc.

VIAU, C., MCGUFFIN, M. J., CHIRICOTA, Y., AND JURISICA, I. 2010. The flowvizmenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration. *IEEE TVCG 16*, 6, 1100–1108.

WANG, J., AND ZHENG, N. 2013. A novel fractal image compression scheme with block classification and sorting based on pearson's correlation coefficient. *IEEE Trans. on Image Proc. 22*, 9.

YUAN, X., GUO, P., XIAO, H., ZHOU, H., AND QU, H. 2009. Scattering points in parallel coordinates. *IEEE TVCG 15*, 6, 1001–1008.