UnTangle Map: Visual Analysis of Probabilistic Multi-Label Data

Nan Cao, Yu-Ru Lin, and David Gotz

Abstract—Data with multiple probabilistic labels are common in many situations. For example, a movie may be associated with multiple genres with different levels of confidence. Despite their ubiquity, the problem of visualizing probabilistic labels has not been adequately addressed. Existing approaches often either discard the probabilistic information, or map the data to a low-dimensional subspace where their associations with original labels are obscured. In this paper, we propose a novel visual technique, *UnTangle Map*, for visualizing probabilistic multi-labels. In our proposed visualization, data items are placed inside a web of connected triangles, with labels assigned to the triangle vertices such that nearby labels are more relevant to each other. The positions of the data items are determined based on the probabilistic associations between items and labels. UnTangle Map provides both (a) an automatic label placement algorithm, and (b) adaptive interactions that allow users to control the label positioning for different information needs. Our work makes a unique contribution by providing an effective way to investigate the relationship between data items and their probabilistic labels, as well as the relationships among labels. Our user study suggests that the visualization effectively helps users discover emergent patterns and compare the nuances of probabilistic information in the data labels.

Index Terms—Visualization, multidimensional visualization, probability vector

1 Introduction

Probabilistic multi-label data is a common type of output from many different types of analysis models in the fields of data mining and machine learning. Such data consists of a set of data items, each described by a probability vector[1] that indicate the likelihood that the item has been categorized by various data labels. For instance, in statistical classification[2] or fuzzy clustering [1], an algorithm is used to label (with a category or cluster) a new data item. The algorithms are typically based on a training dataset containing items whose labels are known, or on a distance measure capturing the similarity between the input data items. Because such methods produce labels that are not mutually exclusive, the analysis result for each item is typically represented as an n-dimensional probability vector where n is the number of possible labels. The ith entry in the probability vector indicates the likelihood of the new data item belonging to the ith category or cluster.

This form of probabilistic multi-label analysis has been used across a wide spectrum of applications. For example, in movie classification tasks, an individual movie might be labeled as both an "action" movie and a "comedy", each with different levels of confidence. In a market segmentation analysis, an individual customer may be probabilistically assigned to multiple segments. In biochemistry, a protein sequence can be assigned to multiple structural categories. In document retrieval, a document may be relevant to multiple topics in varying degrees. In our own everyday life, we often associate with individual people, simultaneously, in multiple communities. For example, the same person can at once be a co-worker and a friend, or a business contact and an extended family member [2]. In all of these cases, the data items (e.g., movies, customers, etc.) may be associated with multiple labels (e.g., movie genres, customer segments, etc.) according to a set of probabilistic values that represent uncertainty levels for corresponding labels.

As these examples show, probabilistic multi-label data is common and widely available in many application domains. Yet despite this ubiquity, few visualization techniques have been designed specifically for such data. Existing work in this area generally follows one of two basic paradigms: (a) visualizing data through a set of independent coordinates, or (b) mapping data to a dimension-reduced plane for visualization. Scatterplot matrices (SPM) [3] and parallel coordinates [4] are commonly used representatives for the first approach. Techniques that follow the second approach, in which high dimensional data is projected to a low dimensional (2-3D) subspace for graphical presentation include multidimensional scaling (MDS) [5] and RadVis [6].

We argue that adopting either of these paradigms for probabilistic multi-label data introduces significant drawbacks. While the first approach of using a set of independent coordinates is useful for discovering correlations between labels, it is typically much more challenging to identify higher level trends or summaries among labels (e.g., which labels are the most dominant or isolated). Meanwhile, the

1. http://en.wikipedia.org/wiki/Probability_vector
2. http://en.wikipedia.org/wiki/Statistical_classification

- *N. Cao is with Graph Computing, IBM T.J. Watson Research Center, Yorktown Heights, NY. E-mail: nancao@us.ibm.com.*
- *Y.-R. Lin is with the School of Information Sciences, University of Pittsburgh Pittsburgh, PA. E-mail: yurulin@pitt.edu.*
- *D. Gotz is with the School of Information and Library Science, University of North Carolina Chapel Hill School, Mannign Hall, Chapel Hill, NC 27599. E-mail: gotz@unc.edu.*

Manuscript received 10 Dec. 2014; revised 14 Mar. 2015; accepted 8 Apr. 2015. Date of publication 20 Apr. 2015; date of current version 2 Jan. 2016. Recommended for acceptance by S. Miksch.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2015.2424878

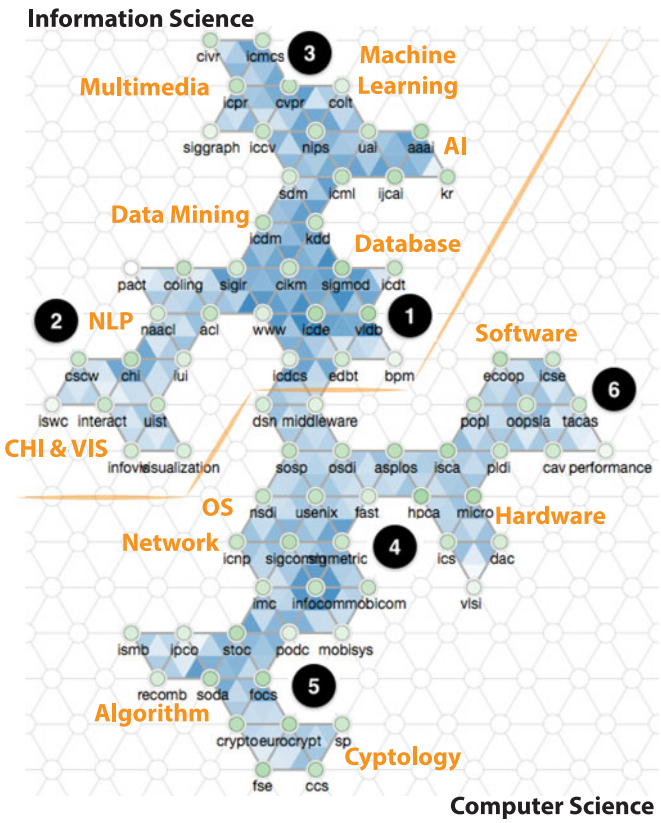

Fig. 1. Visualizing data with multiple probabilistic labels via UnTangle Map. The data shown here is from DBLP and consists of authors and conferences in computer science. We consider conferences as labels because authors are likely to publish in multiple conferences with different probabilities. In the visualization, probabilistic labels (conferences) are placed at triangle vertices and data items (authors) are scattered as points inside the triangles according to their probabilistic associations with the corresponding labels. The positioning of the conference labels are automatically determined based on the intrinsic correlation structure in the data. Interesting patterns revealed by UnTangle Map in this case include: the clusters (labeled by numbers) of different research communities dealing with various directions, and the clear separation between the traditional computer science at the right bottom and the modern information science at the top left.

second approach of using dimension-reduction may help convey proximity between items and labels (if the labels are also mapped onto the same plane), but relationships between items and labels are ambiguous due to loss of information that occurs as part of the reduction process.

In this paper, we introduce *UnTangle Map*, an extension of our previous work on this topic [7]. We describe an improved method for visualizing probabilistic multi-label data (see Fig. 1), which is more scalable, flexible, and intelligent. Following the previous design, we create a triangle mesh in which each triangle serves as a set of axes for a ternary plot.[3] Labels are placed on the triangle vertices and data items are placed inside each of the triangles. The data items are positioned according to the items' probabilistic associations with the triangle's three vertex labels. This design conveys a set of normalized distributions computed for each of the ternary plots, and at the same time allows for the discovery of higher-level patterns through the connections between neighboring ternary plots. Extending our earlier version [7], this article include three specific novel aspects: (1) the introduction of a more precise and controllable layout technique for placing label vertices based on different layout objectives (e.g., to better emphasize correlation or clustering structure); (2) an aggregation-based rendering technique to improve scalability to large numbers of data items; and (3) a more comprehensive evaluation that includes additional case studies on real-world datasets which further demonstrate the design's features, and new discussion dedicated to comparing different layout criteria and their corresponding results.

The key contributions of this paper include:

Visualization design. We identify the main challenges in visualizing data with probabilistic multi-labels and describe the visual design of UnTangle Map, which addresses those challenges. In particular, our novel design leverages the ideas of independent coordinates and subspace creation in order to support several visual query tasks in a probabilistic multi-label dataset where the labels are not mutually exclusive. Our expanded design includes changes to support more scalable visualization of large datasets.

Label arrangement methods. We describe a novel, automated, optimization-based layout algorithm for label arrangement. It adopts a data-driven approach to both assign labels to vertices, and to position the vertices across a two-dimensional trianglular mesh of slots so as to optimize specific layout quality measures. We present the layout measures, describe the overall algorithm, and demonstrate that our approach produces high quality layouts when applied to real-world datasets.

User evaluation. In addition to quantitative evaluation measures, we demonstrate the value of our approach to users by presenting the results from a formal user study. The study results highlight the ability of UnTangle Map to support a variety of visual analysis tasks. Moreover, our evaluation results are compared to two baseline visualization techniques to demonstrate the benefits that our approach provides. The remainder of this paper is organized as follows. We first discuss the problem scope in Section 2, followed by the literature review in Section 3. We present our design and rationale in Section 4, and the label placement algorithm in Section 5. We present the evaluation in Section 6 that includes results from quantitative experiments (Sections 6.1 and 6.2), case studies (Section 6.3) and a user study (Section 6.4). Finally, Section 7 concludes the paper and discusses possible future directions.

2 Problem Scope

Before drilling into technical details, in this section, we discuss our problem scope by formulating the problem, identifying the challenges, and clarifying the research goals.

Problem formulation. Here we describe the specific properties of probabilistic multi-label data and the key visual query tasks on such data.

We present below the visualization problem dealing with probabilistic multi-labels. Let $(x_i)_{i=1...n} \in X$ be the n data items in data set X. Let $(l_k)_{k=1...m} \in L$ be the m different labels in label set L. Each of the items is associated multiple labels with different level of uncertainties, which can be represented by a probabilistic vector $\vec{p_i} = \langle p_{i1}, p_{i2}, \ldots, p_{im} \rangle$ with

3. http://en.wikipedia.org/wiki/Ternary_plot

real value $p_{ik} \in [0, 1]$ for $i = 1 \ldots n$, $k = 1 \ldots m$. The probabilistic value p_{ik} usually represents the posterior probability of data item x_i for the label k.

Challenges. There are several key challenges for visualizing the aforementioned data:

- *Scalability.* The number of labels in the data may be large—datasets with dozens or hundreds of labels are typical (e.g., the genre labels in a movie dataset, or the topic labels in a document corpus). Existing methods for multivariate visualization, including scatterplot matrices and parallel coordinates, typically suffer from scalability issues.
- *Subspace ambiguity.* Multidimensional scaling and other projection-based techniques map data items to a low-dimensional subspace, which can distort the original relationships between data items and labels. The process results in information loss and introduces ambiguity.
- *Visual summary of probabilistic distributions.* Most existing tools lack the capacity to summarize the distribution of labels, e.g., to inform which labels are more or less populated among the data items.

Goals. We identify visual query tasks as our design goals in the context of the aforementioned problems and challenges. Generally, our work has been motivated by the necessity of supporting visual inquiry on the data with probabilistic multi-labels:

- *Q1. Item-label relationship.* How do data items associate with many different labels? How much, in a probabilistic sense, is an item associated with a specific label compared with other labels?
- *Q2. Label summary.* Which labels are most (or least) populated among the data items?
- *Q3. Two-way label interaction.* How are common items shared between two labels? Which labels share items most intensely?
- *Q4. Three-way label interaction.* For data items strongly associated with two labels, are there additional label (s) that are also strongly associated?
- *Q5. Multi-way label interaction.* For a set of labels, which is the most dominant (having the strongest association with the data items) and which is the most isolated (having the weakest association with the data items)?

Proper support for these tasks requires overcoming the above-mentioned challenges. For example, a solution to Q1 needs to address both the scalability and subspace ambiguity issues, while a solution to Q2 corresponds to the visual summary challenge. Furthermore, Q3-Q5 relate to the challenge of visualizing the interactions among labels. In particular, Q3 relates to interactions between pairs of labels (two-way), Q4 relates to ternary interaction (three-way), and Q5 relates to interactions among many labels (multi-way). Our goal is to provide a visual technique that can support all of these visual query tasks.

3 Related Work

In prior work, probabilistic multi-label data have most commonly been visualized using either multidimensional or graph-based techniques. In particular, versions of these visualization methods have been applied to data produced by fuzzy clustering, topic modeling, and classification. In this section, we provide a brief review of related work in these areas. We then identify a set of challenges to be addressed in our visualization design.

3.1 Visualizing Labels as Multidimensional Data

One approach to visualizing probabilistic multi-label data is to use methods designed for multidimensional or multivariate data (mdmv) [8]. In this approach, each label in the data corresponds to a dimension, and data items are associated with each of these dimensions of labels through a probabilistic value. In this sense, label data can be viewed as multiple dimensions of numeric variables. It can therefore be visualized with multidimensional methods. Existing techniques generally fall into two visual paradigms: (a) independent coordinates or (b) a dimension-reduced plane.

Representative techniques in the independent coordinates category include scatterplot matrices [3] and parallel coordinates [4]. Scatterplot matrices [3] represent data items in all pairwise permutations of dimensions such that the relationships between any two specific dimensions can be discovered and compared. However, since the number of matrices grows quadratically with number of dimensions (labels), this visualization does not scale well as the number of labels grows. Although interaction techniques such as Rolling the Dice [9] may be used to help users explore the data, discovering relationships among many labels remains challenging. Like scatterplot matrices, parallel coordinates [4] and many of its variants (e.g., [10], [11]) are only effective when the number of dimensions is small [12]. Moreover, clutter reduction is needed for data with many dimensions [13]. Besides scalability, a major issue with such independent coordinate representations is that they do not facilitate higher-level visual comparison among labels, such as identifying the most dominant or isolated labels according to the distribution of data items.

Given the challenges of scale, the second paradigm uses dimension reduction to map data into a lower-dimensional space for visualization. Multidimensional scaling [5] is one of the most popular techniques in this category. MDS seeks to preserve high dimensional distances in a low (2D or 3D) dimensional space. Principal Component Analysis [14] and various linear transformation methods [15] project data by maximizing the variance of data items based on different constraints. Self-Organizing Maps [16] use a 2D lattice to portray the distribution of data element in the high dimensional space via a learning process. A modified Sammon Mapping [17] preserves the distance between data elements and cluster center in a low dimensional space. RadVis [6] projects the multidimensional data into barycentric coordinates [18] based on a force-directed layout model [19]. t-SNE [20] creates a single map that reveals structure at many different scales. Compared to the independent coordinate representations, these methods are more scalable for high dimensional data. However, when projecting data items and labels to a lower dimensional space, proximity among items and labels are distorted and information is lost. This means that the visualized information may no longer be accurate and can become ambiguous.

3.2 Visualizing Labels as Graphs

The relational patterns inside probabilistic multi-label data can also be illustrated with graph-based visualizations [21], especially by using node-link diagrams [22] based on distance-embedding layout algorithms (e.g., force-directed layout) [23]. In this case, nodes in the graph represent labels or data items, while edges and their weights represent the relationships between data items (e.g, item similarity), between labels (e.g., dimension correlation), or between labels and items (e.g., the probabilities of the items being categorized by the labels). In the visualization, the distances between pairs of nodes are in reverse proportion to the strength of the corresponding relationships (i.e., two closely positioned nodes indicate a strong relationship between them). This design usually has significant readability issues when the graph is dense [22]. This is a critical limitation for probabilistic multi-label data, where a complete graph is required to capture the relationships between all pairs of data items and labels. Although visual clutter can be reduced by hiding or filtering the edges, the visual ambiguity that is introduced makes data interpretation difficult.

3.3 Visualization of Statistical Analysis Results

The statistical results produced by classification, fuzzy clustering [1], and topic modeling [24] can considered forms of probabilistic multi-label data. Visualizations designed of these results are therefore closely related to our work.

Fuzzy clustering. Fuzzy clustering methods assign data items to one or more clusters with a degree of uncertainty (hence the term "fuzzy"). Rousseeuw proposed Silhouettes [25], a method that attempts to interpret fuzzy clusters in a one-dimensional diagram. Each data element is represented as a small dot and packed inside its most likely cluster. Wiswedel et al. [26] extend this design with interactive functions that allow users to select and discard the elements in each cluster to fine-tune the clustering results. Techniques were also proposed for representing overlapping clusters as lines or bubbles [27], [28], [29], [30]. However, they all more or less suffer from line crossings or set overlaps when data are dense. Recently, Cao et al. [31] and Streit et al. [32] introduced techniques for visualizing fuzzy clusters by grouping data items based on similarity of their probability vectors. These designs represent fuzzy clusters via clear cuts which could be misleading.

There has also been work that represents fuzzy clusters in a projection space, where contour or lines are used to depict soft cluster boundaries [17], [33], [34]. Simonetto et al. [35] and others [36], [37] developed methods to generate Euler-like diagrams for visualizing overlapping clusters. ContexTour [34] uses a contour map to represent the density distribution of data items, showing a smooth and fuzzy margin between two adjacent clusters. These designs use projection for dimension reduction, a method with limitations that have already been discussed.

Topic models. A branch of work closely related to fuzzy clustering is topic modeling applied to text data [38], [39]. Using techniques such as Latent Dirichlet Allocation [38], text documents can be automatically associated with one or more topics for search or organizational purposes. Recent advances in topic visualization have focused either on topic transition [40], or on viewing topics across different information facets [41], [42]. For most of these techniques, the probabilistic topic assignment is first converted into a hard assignment for simplicity, and hence they are not suitable for visualizing probabilistic multi-label data.

In text visualization, it is common to treat documents as high dimensional data based on the bag-of-words vector space representation. Dimension reduction techniques can be used to visualize keywords or documents on a 2D plane, with related items reflected through the spatial clustering of keywords (or documents) [43], [44], [45]. For example, Iwata et al. proposed the probabilistic latent semantic visualization model (PLSV) [46] to generate a more interpretable distribution of documents by considering various visualization criteria. However, as discussed before, such dimension reduced representations suffer from visual distortion and potential loss of information.

Classification. A wide variety of analysis models have been developed to categorize data items based on a set of predefined labels. The results from these classification methods are most commonly visualized using projection-based approaches [47], [48]. To reduce visual clutter, Rheingans and Desjardins [49] aggregate the data items in a projection view by using a heatmap that visualizes the probability of a given class for each value combination of two features. Seifert and Lex [50] place the labels on a circle within which data items are positioned using barycentric coordinates (similar to RadVis [6]). These approaches are all based on projection with the aforementioned limitations.

Recently, Alsallakh et al. [51] proposed a novel visualization method in which labels are displayed as ring sectors containing histograms representing the classification probability of all data items. The labels are also connected by chords whose thickness represents the classification confusion between them. However, when the number of labels is large, reading the detailed data distributions and classification confusion statistics between labels becomes difficult.

4 Visualization Design

The design of UnTangle Map seeks to overcome the challenges discussed in Section 2. Here, we discuss the design details and rationales. We then illustrate how the design can generate meaningful visual patterns, and present a set of intersection functions that support additional analysis capabilities.

4.1 Design Rationale

In order to support the visual query tasks outlined above, the key idea of our approach is to visualize item-label relationships, label summaries, and label interactions through a set of intelligently connecting ternary plots.

A ternary plot, as illustrated in Fig. 2a, is a barycentric plot of three variables, with each variable corresponding to a vertex on an equilateral triangle. Typically, the three variables sum to 1.0 or 100 percent, and the position of any given point on the triangle indicates the ratios of three variables. UnTangle Map builds upon basic ternary plots to visualize data items with probabilistic labels. To show items associated with three labels, we assign the labels to each of the vertices of a triangle, and plot a data item on the ternary

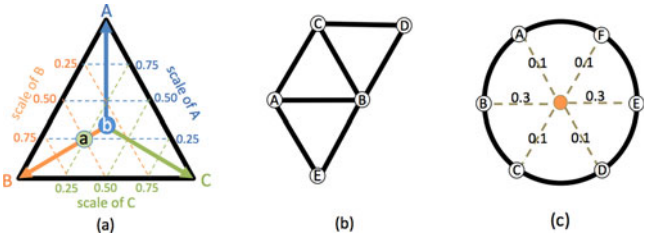

Fig. 2. (a) A ternary plot and the 3D barycentric coordinate system. (b) A ternary plot mesh. (a) Ambiguity is unavoidable when the number of labels (dimensions) is larger than 3.

plot at a position whose distance to each label encodes the item's association, represented as a probabilistic value, with the label. For example, as shown in Fig. 2a, there are three labels A, B, and C, plotted on the vertices. The item a is associated with A, B, and C with probabilities 0.25, 0.5, and 0.25, respectively. As a has stronger association with B, it is positioned at a point on the perpendicular direction of edge AC and proportionally close to B. Another data item b is located at the center of the ternary plot which means it is associated with the three labels with equal probabilities of $1/3$. Sometimes, different data items with same proportions to the data labels may overlapped together (e.g. $P_1(0.33, 0.33, 0.33)$ and $P_2(0.01, 0.01, 0.01)$). We differentiate these cases by adjusting their size and opacity based on the magnitude of their probability vectors, where larger magnitude is reflected by higher opacity and larger size.

When labels are more than three, we combine multiple ternary plots where the set of vertices correspond to the set of labels. The data items are repetitively plotted inside each ternary plot based on their normalized probabilities associated with the three corresponding labels. Those items that are irrelevant to all the labels of a ternary plot are removed. The result is a mesh of connected triangles as shown in Fig. 2b. Inside each individual triangle, the ternary plot serves a subspace for unambiguously displaying the item-label relationship. The connected ternary plots form a triangle mesh that allows patterns to aggregate into visual summaries of the labels. Furthermore, different label interactions are captured by the patterns around the vertices and edges that connect triangles.

The triangular design is based on the goal of avoiding ambiguity. In particular, the three-dimensional barycentric coordinate system in a ternary plot makes the position of each item, representing its relative probabilistic associations with the three corresponding labels, unambiguous in the two-dimensional plane. In contrast, when barycentric coordinate systems contain more than three vertices (an n-dimensional shape with $n > 3$) on a 2D plane, ambiguity is unavoidable. For example, Fig. 2c shows a data item from a 6-dimensional space projected to 2D, after which the 2D-distances from the vertices (labels) no longer uniquely represent the item's true values.

4.2 Visual Patterns

The basic design presented above produces a variety of meaningful visual patterns that support the various tasks outlined in Section 2.

A first set of patterns, which are observed within a single ternary plot, allow for the interpretation of item-label relationships (Q1). As shown in Fig. 3, we identity four distinct archetypes that can help interpret the arrangement of probabilistic data points within a ternary plot. (a) In a *non-dominant pattern* the data items are distributed in the middle of the ternary plot with equal distances to the three label vertices, and none of the labels are overly associated with the items. (b) In a *uni-dominant pattern*, the data items are concentrated at a corner where the closest label has a dominant relationship with the items. (c) In a *bi-dominant pattern*, the data items are located along an edge where the two closest labels both have strong associations with the items. (d) Finally, in a *balanced flow pattern*, two labels (A and C) have equally strong associations with data items regardless of the strength of the third label (B). The data items in a balanced flow pattern are distributed along an axis perpendicular to edge connecting the two strong labels (AC) towards the third vertex (B). Note that the uni-dominant pattern also helps support Q2, while the bi-dominant pattern helps address Q3.

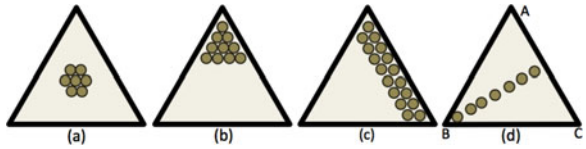

Fig. 3. Typical patterns for item-label relationship: (a) non-dominant, (b) uni-dominant, (c) bi-dominant, and (d) balanced flow patterns.

Variants of the four archetypes defined above can also be highly informative. For example, Fig. 4a shows data items distributed around the corners of a triangle, suggesting each of the labels has a dominant relationship with a portion of the data items. Fig. 4b shows data items distributed along the edges, suggesting that each of the pairs of labels shares a portion of items in common without a strong third-label association. Fig. 4c shows a linear pattern parallel to the edge AB, suggesting that the items have a relatively constant association with the label C.

A second set of patterns can be defined when considering pairs of neighboring ternary plots, which allow users to interpret higher-level label interactions (Q4). As shown in Fig. 5a, when two connected ternary plots share a vertex (A), users can visually compare the relationship between A and the other connected labels. For example, Fig. 5a suggests the associations with label B and C are stronger with respect to A when compared with D and E. When two triangles share an edge as shown in Fig. 5b, the connected ternary plots allows a user to compare the relationship between two labels (e.g., A or D) given a common baseline (BC). For example, the figure suggests that, given that data items are associated with B and C, the association with A is stronger than with D.

A third set of typical patterns can be seen when viewing arrangements of multiple (more than 2) adjacent ternary

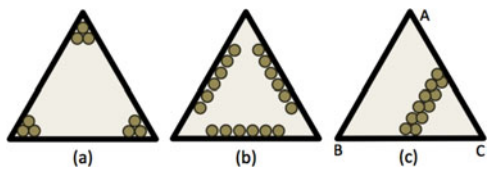

Fig. 4. Other patterns for item-label relationship: (a) three-corner, (b) three-edge, and (c) constant patterns.

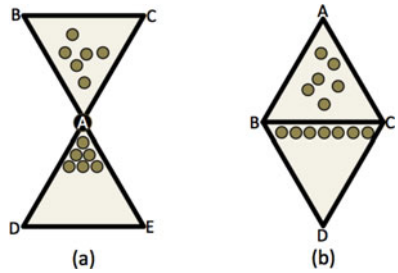

Fig. 5. Typical patterns for higher-level label interactions: (a) shared vertex and (b) shared edge patterns.

plots. Such a configuration allows for the interpretation of multi-way label interactions (Q5) as well as global label summaries (Q2). As shown in Fig. 6, there are three different archetypes in this category. First, (a) in a *global dominant pattern*, the label vertex at the center appears to be uni-dominant across all connected ternary plots, meaning that the corresponding label has the strongest association with the data items among all other present labels. Second, (b) in a *complimentary pattern*, the non-dominant pattern appears in all connected ternary plots, meaning that the data items have relatively balanced associations across all of the present labels. Finally, (c) in an *isolated pattern*, the bi-dominant patterns appear in all connected ternary plots, with the label at the center having the weakest association compared with all other present labels—in other words, the center label is isolated from the rest of present labels.

To further assist user interpretation, UnTangle Map automatically scores each vertex to determine how isolated or dominant it is with respect to its neighbors. That score is then used to color-code the corresponding vertices. By default, red is used to indicate an isolated label while green is used to indicate a globally-dominant label. White is used for vertices that fall in between those extremes. A gradient is used to interpolate between the red, white, and green color stops.

The patterns described here are able to convey many meaningful insights about the data being visualized. However, there are some limitations in our design. First, as we will discuss in Section 6.4, linear relationships between two labels are not as easily captured in a ternary plot when compared to a scatterplot. Second, our design is focused on the task of visualizing the distribution of data items with probabilistic labels, and therefore does not consider the visualization of other types of variables (such as numerical or categorical variables). These two limitations show that UnTangle Map's approach can compliment existing methods that more directly support these tasks. Third, because our design relies on a grid of connected equilateral triangles, each of the vertices (labels) has at most six direct neighbors. This can potentially limit a user's ability to explore very high-order label interactions. To overcome this limitation, UnTangle Map provides user interaction capabilities that allow for the interactive customization of label placements. This interactive feature is described in next section.

4.3 Interactions

UnTangle Map provides a set of interactions that further support the process of information seeking and data interpretation.

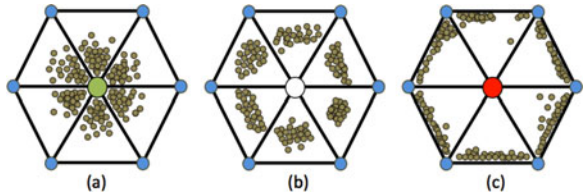

Fig. 6. Typical patterns for multi-way label interactions: (a) global dominant, (b) complimentary, and (c) isolated patterns.

Smart layout. The positioning of labeled vertices can be generated either in a data-driven manner or in a user-driven manner. When a dataset is first loaded in the UnTangle Map visualization, the system automatically generates an initial layout, arranging labels on a triangle mesh according to the internal distributions of the data items (see Section 5). This primary view is augmented with an inset window that shows an overview of all available data labels. By interacting with these views, users can add, delete, or reconfigure labels in the primary view. First, users can add a new label vertex to the primary view by dragging the label from the inset window to any empty slot in the triangle mesh. Labels can be added more than once to the visualization, meaning that multiple vertices may correspond to the same label. Similarly, users can drag a label vertex already present in the primary view from its current position to any of the available empty slots to change its location. Vertices can be removed by dragging them off the primary view space.

While the manual placement of labels provides users with the greatest flexibility, automated algorithms are used to help guide the user to a more effective visualization. When users begin to drag a label, UnTangle Map highlights an empty slot in red that corresponds the best position to place the dragged label based on a data-driven, correlation-based computation. Similarly, when users click on an empty slot, the label that best fits (in a data-driven, correlation-based manner) the slot is highlighted in the inset window. The algorithm used to drive these recommendations is described in Section 5.

Switch of correlation measure. By default, Spearman's correlation coefficient is used as the basis for the algorithms within UnTangle Map. However, users are able to select from three different correlation coefficient functions (Pearson's, Spearman's, and Kendall's) in the toolbar to control how the underlying statistics are computed by the system.

Brush. UnTangle Map supports two types of brushing operations. First, users can brush the inset window to select a set of focused labels into the primary view. Second, inside each ternary plot, users can brush the individual data items to highlight the same set of items in other ternary plots.

Zoom and Pan. When there are many labels, the triangle mesh can grow large, making the size of each triangle small. Users can zoom in to a focused ternary plot by double-clicking it. Users can also pan the entire mesh to navigate through the full grid of triangular plots even when tightly zoomed.

5 Layout Design and Implementation

The layout process for UnTangle Map consists of two major steps: (1) the layout of labels by connecting them in a

triangle grid to generate a web of three dimensional subspaces, and (2) the plotting of data items inside each triangle to generate ternary plots.

5.1 Layout Data Labels

The procedure for creating the data label layout can be broken into two steps: (1) creating a triangle grid whose vertices are empty slots that will be used to place labels, and (2) allocating labels to the slots via the optimization of an objective function.

5.1.1 Creating Triangular Grid

We begin by creating a grid of equilateral triangles, $G_s = \langle S, E_s \rangle$, based on triangular tiling.[4] The vertices of the grid, S, are empty slots used for placing labels. E_s denotes the set of edges on the triangles. In addition to creating the desired equilateral triangles, such a grid provides efficient spatial indexing so that the grid coordinates of each vertex (i.e., slot) can be easily used for allocating labels (either in a data-driven or user-driven manner).

Theoretically, the grid can be infinitely large to support the allocation of an unlimited number of labels. In practice, we create a grid on a virtual plane that is several times larger than the grid visible on the display (viewport), and only a portion of the grid is shown on the viewport at a given time. This virtual plane can be navigated through the zoom and pan interaction functions as described in Section 4.3. The size of the virtual plane is empirically determined, and we found that a grid that is five times of the viewport is more than sufficient for practical use.

5.1.2 Allocating Labels to the Label Slots

We seek to assign labels to positions on the grid such that nearby labels are more relevant to each other in terms of shared data items. To achieve this, we introduce an efficient layout optimization algorithm and a corresponding layout objective function.

Layout objective. We design an objective function such that correlated labels are connected or clustered on the grid. Let the layout of m labels $L = \{l_1, \ldots, l_m\}$ on a triangle grid be denoted as $G_L = \langle V, E \rangle$, where $V = \{v_1, \ldots, v_m\}$ is the set of label vertices located on the grid slots $S = \{s(v_1), \ldots, s(v_m)\}$. To simplify the notation, we write s_a as $s(v_a)$, the slot of label vertex v_a. E is the set of edges such that edge $e = (v_i, v_j)$ exists if s_i and s_j are connected on the grid. Let T be the set of ternary plots in G_L. A ternary plot, $t \in T$, is a triangle on the triangle grid whose vertices represent three different labels respectively. Our goal is to produce a label layout such that labels with high correlations are connected through edges, or clustered through a mesh of connected triangles. This goal is achieved by maximizing the objective $\mathscr{F}$:

$$\mathscr{F} = \alpha \frac{1}{|E|} \sum_{(v_i, v_j) \in E} c_{ij} + (1 - \alpha) \frac{1}{|T|} \sum_{t \in T} c_t,$$

$$G_L = \text{argmax}_{V,E}\, \mathscr{F},$$

where, $\alpha \in [0, 1]$ is a parameter that balances between two optimization terms. The first term ensures the connections of related label vertices in G_L by maximizing the average value of the pair-wised correlation of label vertices. The second term preserves the cluster structure via maximizing the average correlation of all the ternary plots in G_L.

Here, c_{ij} indicates the correlation between two connected label vertices v_i and v_j and $c_t = (c_{ij} + c_{jk} + c_{ki})/3$ indicates the correlation of a triangle (a ternary plot) t whose vertices are v_i, v_j, and v_k, respectively. Specifically, c_{ij} can be computed as the correlation of two probabilistic vectors $\vec{p_i}$ and $\vec{p_j}$. The kth element in a probabilistic vector $\vec{p_i}$ corresponds to the ith data item's association with the label k in terms of the probabilistic value. The correlation can be computed by using Pearson correlation coefficient, or nonparametric measures such as Spearsman's rank correlation coefficient or Kendall's rank correlation coefficient. Nonparametric measures are used when the normality assumption does not hold in the data, which is common in a probabilistic multi-label dataset. We use Spearman's rank correlation as default, and provide other correlation types as user-selectable options.

Implementation. Optimizing the above layout objective is an NP hard problem. We propose to find a solution via a greedy algorithm as summarized in Algorithm 1. This algorithm allocates labels to slots in an iterative manner. Specifically, the algorithm starts by placing the first label at the center slot in the triangle grid and selecting and placing the next label by maximizing the value of the objective function. The next label to be placed at each iteration can be selected based on different strategies, such as random selection or selecting the label that is most correlated to other labels on the grid. In our implementation, we use dynamic programming to efficiently enumerate all of the potential choices for label-slot assignment to best maximize the layout function. In Algorithm 1, $utility(G'_L)$ denotes the utility value of an instance layout G'_L, which is computed based on the objective function $\mathscr{F}$.

Algorithm 1. UnTangle Map Layout

Data: The label set L; The triangle slot grid $G_s = \langle S, E_s \rangle$
Result: The solution triangle mesh $G_L = \langle V, E \rangle$
begin:
  $u_{best} \leftarrow 0$; $G^*_L \leftarrow \emptyset$;
  for $l_i \in L$ **do**
    place l_i at the center slot s_0 on G_s;
    $L' \leftarrow L - \{l_i\}$; $G'_L \leftarrow \emptyset$;
    while $L' \neq \emptyset$ **do**
      $u_{max} \leftarrow 0$; $(l^*, s^*) \leftarrow (null, null)$;
      $S' \leftarrow get\_valid\_slots(G_s, G'_L)$;
      $R \leftarrow L' \times S'$;
      for $(l_j, s_k) \in R$ **do**
        place label l_j at slot s_k;
        update G'_L;
        $u \leftarrow utility(G'_L)$;
        if $u_{max} < u$ **then**
          $u_{max} \leftarrow u$; $(l^*, s^*) \leftarrow (l_j, s_k)$;
        remove l_j from s_k;
      place l^* at s^*;
      $L' \leftarrow L' - \{l^*\}$;
      update G'_L;
    $u \leftarrow utility(G'_L)$;
    if $u_{best} < u$ **then**
      $u_{best} \leftarrow u$; $G^*_L = G'_L$;
  return G^*_L;

4. http://en.wikipedia.org/wiki/Triangular_tiling

At each iteration, the algorithm searches in the space of $R = L' \times S'$ to find a label-slot assignment $(l_i, s_j) \in R$ that maximizes the layout objective. Here, L' and S' respectively indicate the sets of all the free labels and slots, and R indicates the set of all the potential label-slot assignments. In our implementation, we accelerate this search process by reducing the search space based on the layout constraint. In other words, we only investigate the free slots that are able to enclose boundary edges and generate new triangles as shown in Fig. 8.

Finally, the algorithm terminates when all of the labels have been placed in the triangle grid.

5.2 Plotting Data Items

Once the ternary plots are generated based on the above layout algorithm, we plot the data items inside each ternary plot based on the barycentric coordinate system. More precisely, given a position $\mathbf{v}$ inside a ternary plot, its (Cartesian) coordinates can be computed through the coordinates of the three triangle vertices:

$$\mathbf{v} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \lambda_3 \mathbf{v}_3,$$

where $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ are triangle vertices whose coordinates are known. $(\lambda_1, \lambda_2, \lambda_3)$ are the barycentric coordinates of the point $\mathbf{v}$, subjected to the constraint $\sum_{i \in \{1,2,3\}} \lambda_i = 1$. Here, λ_i are given by the associations of an item with the three given labels (l_1, l_2, and l_3) respectively, in terms of their probabilistic values, and $\mathbf{v}_i$ is the corresponding label position. When there are more than three labels in a dataset, the data items' distributions with any three given labels, l_1, l_2, and l_3 are given by the normalized probabilities. For example, the normalized probabilities of an item i associating with the three given labels can be computed by $u_{ik} = p_{ik}/(p_{i1} + p_{i2} + p_{i3})$ for $k \in \{1, 2, 3\}$.

Fig. 7 shows an example visualization of a DBLP dataset by using the aforementioned algorithm. In this figure, each ternary plot consists of 3,000 data items. This scale can slow the rendering process and interfere with the high-speed demands of user interaction. Overplotting from visual clutter can also negatively impact legibility. This challenges emerge as data size increases.

To address these challenges, we compute a ternary heatmap in which data items inside a ternary plot are hierarchically aggregated into triangular data bins, producing a multi-granularity representation that supports multiple levels-of-detail in the visualization. Specifically, we approach this goal by recursively splitting a triangle into four sub-triangles in a top-down approach and counting the number of data items inside each triangle as a weight. The weights are globally normalized and represented as the opacity of the triangle's fill color. Fig. 9 shows an example of rendering UnTangle Map in different levels of granularity. The medium level of granularity balances between rendering performance and data details and better capturing visual patterns when compared with rough level of granularity. Fig. 1 shows another example of applying the ternary heatmap. Compared with Fig. 7, the spatial density of item distribution can still be clearly captured in Fig. 1, but the heatmap allows faster rendering and interaction. We present and discuss the scalability of this approach in Section 6.

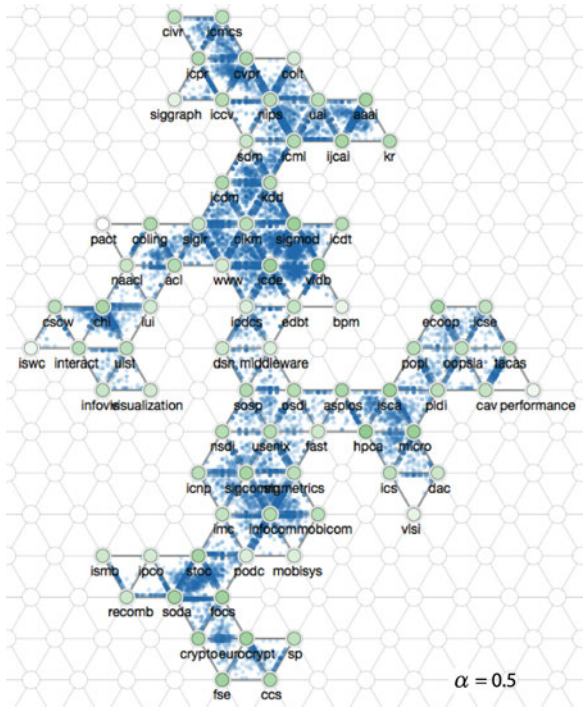

Fig. 7. The layout result of the DBLP dataset generated based on Algorithm 1 and the item plotting method. Here, the two optimization terms in the layout objective are weighted equally. The rendering can be further enhenced by using ternary heatmaps as shown in Fig. 1.

5.3 Discussion

Both the objective function and the greedy layout algorithm are key determinants in resulting visualizations produced by our methods. This section provides more discussion regarding these two critical components of our design.

Choices of objective. We investigate several design choices that can be used to emphasize different correlation patterns within the data. First, because the number of total edges and triangles can vary, we can choose in the objective function to optimize the averaged (Fig. 7) instead of total correlations (Fig. 10) in both optimization terms. This design helps to reduce the number of total edges and triangles in G_L, providing a clearer and more informative view. Second, in the layout objective function, we can choose to balance between two optimization terms that capturing both pair-wised correlations and the structure of clusters at the same time. Fig. 11 illustrates the effect of these two terms on the layout results generated respectively based on $\alpha = 1$ and $\alpha = 0$. When choosing to balance these terms with an equal importance (i.e., $\alpha = 0.5$), features of both terms are partially captured as shown in Fig. 7.

Layout performance. To achieve a better maximization of the layout objective and avoid locally optimal solutions, we employ the stochastic hill climbing technique.[5] More specifically, instead of always selecting the best label-slot assignments during the update stage, the stochastic hill climbing method probabilistically chooses alternative slots (e.g., the second or third best). We illustrate this strategy through the following experiment.

5. http://en.wikipedia.org/wiki/Stochastic_hill_climbing

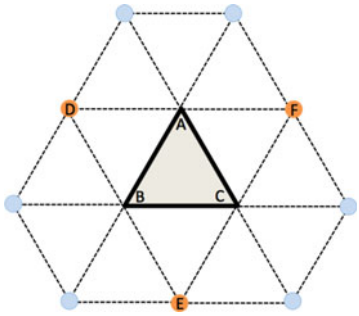

Fig. 8. Surrounding the ternary plot ABC, there are three valid free slots D, E, and F that are able to enclose boundary edges of the ternary plot.

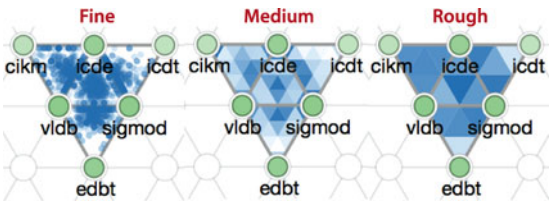

Fig. 9. Generating the ternary heatmap for supporting different levels of details. The level of granularity from left to right: fine, medium and rough.

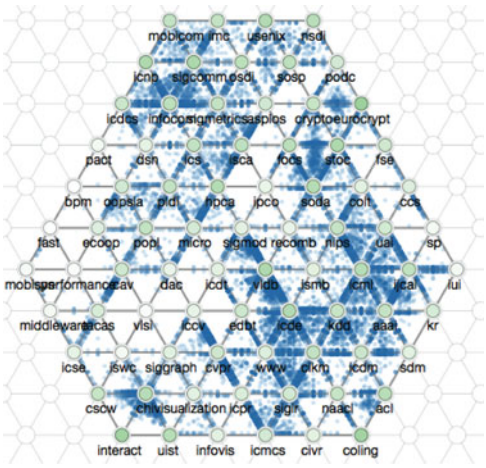

Fig. 10. The layout result by maximizing the total correlation of labels.

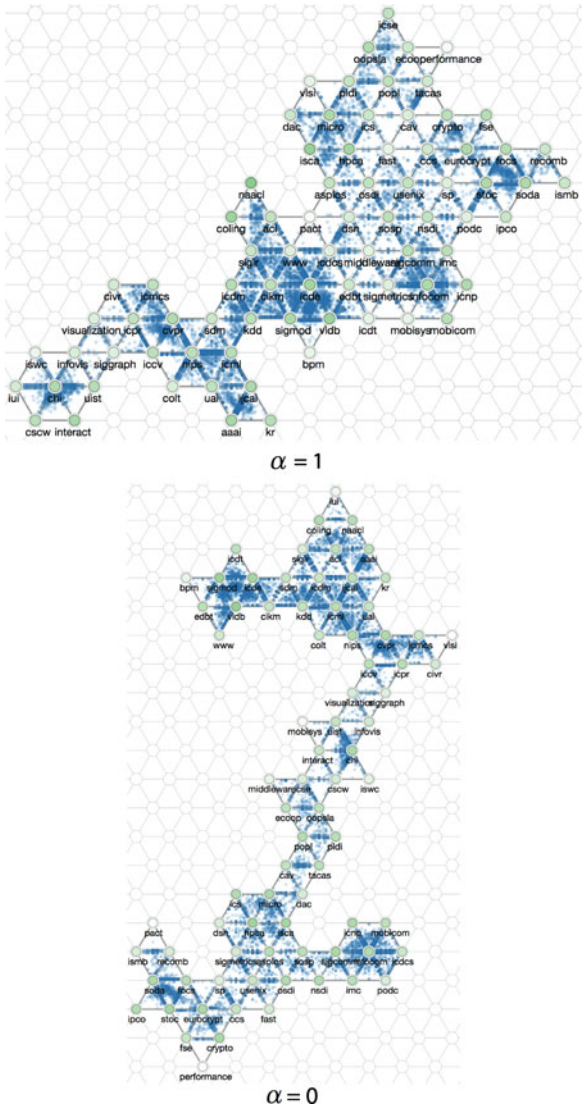

Fig. 11. Balancing between two optimization terms by α. When $\alpha = 1$, the pair-wised correlations of labels are best preserved; when $\alpha = 0$, the structure of label clusters is captured (the figure illustrates two primary clusters of the input data).

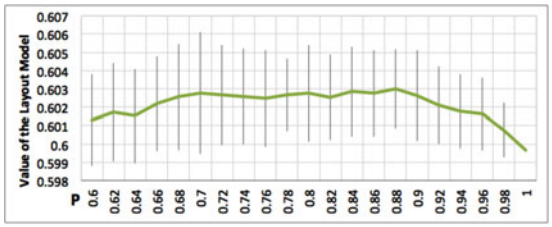

Fig. 12. Layout improvement based on stochastic hill climbing.

Let P be the probability of selecting the best label-slot assignment during the update stage. Given P, the probability for selecting the second best or the third best are defined equivalently as $(1 - P)/2$. To examine the impact of different P values, we tested our layout algorithm for values of P in the range $[0.6, 1]$ with an increasing step of 0.02. For each given P, we repeated the layout procedure for 100 times, and recorded the final utility values of the layouts in each trail. The mean value and standard deviation corresponding to each P are shown in Fig. 12. The results show empirically that stochastic hill climbing most improves the layout performance for $0.7 < P < 0.9$.

We note, however, that while this stochastic approach improves the layout results, the choice of the best P can depend somewhat on the dataset. Moreover, experimentally searching through the possible values for P to identifying the optimal setting for a given dataset can be computationally expensive. For this reason, the above strategy is suggested for use only as an offline procedure when producing more precise results is necessary.

Extension of usage. Beyond probabilistic multi-label data, UnTangle Map can also be applied to visualize multidimensional data with non-negative values. The constraint ensures that an item's multidimensional values are additive and can be meaningfully normalized across different dimensions during the process of generating barycenter layout as discussed in Section 5.2. Many approaches can be used to produce multidimensional data representation, such as non-negative matrix factorization [52], or convert real numbers to a desirable range, such as min-max scaling.

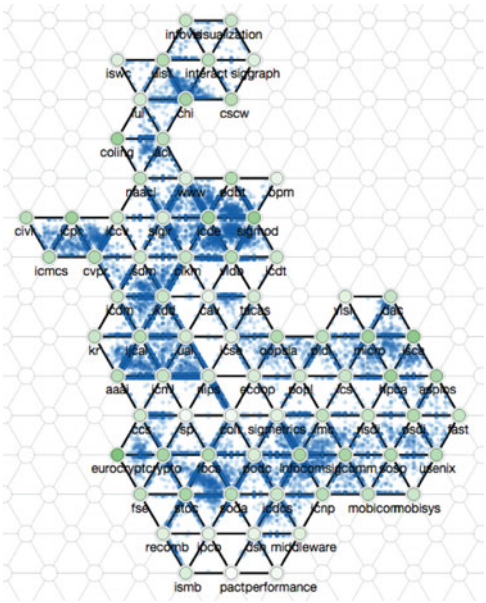

Fig. 13. UnTangle Map visualization of the DBLP dataset generated based on the layout algorithm described in [7].

However, whether the attribute values across different dimensions are additive is usually determined by the semantics of the data.

6 Evaluation

In this section, we present several different types of evaluation results: (1) experiments that compare different layout algorithms within UnTangle Map, and experiments that compare UnTangle Map with alternative baseline methods; (2) an evaluation of the scalability of our approach; (3) two case studies conducted by an expert using real-world multi-label datasets; and (4) a controlled user study allowing us to quantitatively test performance with UnTangle Map against Parallel Coordinates Plots (PCP) and scatter plot matrices.

6.1 Experiments

We conducted two experiments for comparing the layout techniques proposed in this paper to the layout algorithm introduced in our previous paper [7], as well as to two other visualization designs: SPM and PCP. All of our experiments are conducted using a dataset extracted from the computer science bibliography database DBLP.[6] The dataset includes two types of elements: a set of *authors* and the *conferences* in which they published papers. In this use case, we are interested in exploring how authors publish in related conferences and how conferences share common participants. Hence, we consider the conference names as labels applied to the authors. We compute the confidence in a given label for each author by looking at how often that author has published at the corresponding conference.

6.1.1 Comparison of Different Layout Algorithms

Fig. 13 shows the layout generated based on our original layout algorithm introduced in [7]. Particularly, this algorithm always places a label l_i at a slot next to an existing edge $e \in E$ in G_L, greedily maximizing $(c_{ij} + c_{ik})$ at each layout iteration. In contrast, results from our new algorithm are shown in Figs. 1 and 7. While both algorithms place relevant conferences close to one another, we find that new current algorithm generates a more precise representation of the data in which research communities are more clearly separated and reasonably connected.

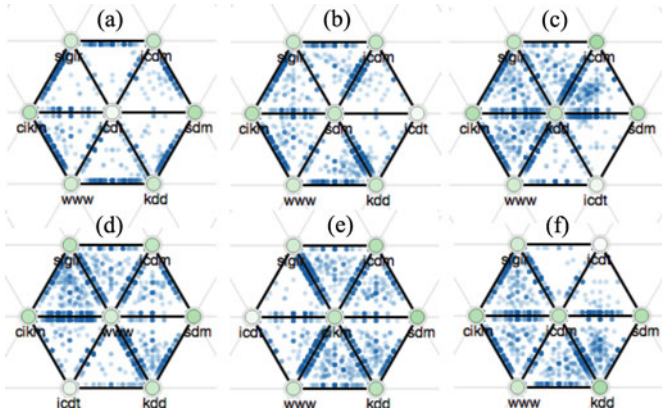

Fig. 14. Author distribution among data mining related conferences.

In particular, Fig. 1 captures six research communities (clusters) labeled by the numbers (1)-(6) respectively inside two general research areas related to (a) information science and (b) more traditional computer science. Within these clusters, we see about 13 distinct research directions. The clustering emerges due to the fact that conferences in similar areas tend to share the same groups of authors. Interestingly, the traditional database/data mining research community (1) is split into two branches focusing on interactions with two other specializations. One branch (2) consists of the conferences in the field of human computer interaction (HIC), visualization (Vis), and natural language processing (NLP), all of which leverage human factors for data analysis. Another branch (3) consists of the conferences in the field of machine learning, artificial intelligence, and multimedia, which focus on the data analysis techniques powered by statistical/algorithmic learning. This visual structure clearly illustrates two major directions being actively pursued within the research community. A similar pattern can also be found in computer science, in which three distinct communities focusing on systems (4), algorithms (5), and engineering (6) emerged. We note that these communities were not clearly separated in Fig. 13.

6.1.2 Comparison with PCP, SPM, and PCA

We use the example shown in Figs. 14, 15, and 1 to illustrate the advantages of UnTangle Map over PCP, SPM and PCA. Fig. 15a shows three of the seven data mining related conferences plotted using SPM. Each dot represents an author, and the x- and y-positions on the scatterplot indicate the probabilities of an author publishing in conferences x and y, respectively. Since SPM is efficient for discovering pairwise patterns, it is possible to capture which two conferences have stronger associations. For example, by looking at the row for the SDM conference, one can identify that SDM has a strong association with KDD and ICDM, as we have also shown in Fig. 14b by using UnTangle Map. However, based

6. http://dblp.uni-trier.de/db/

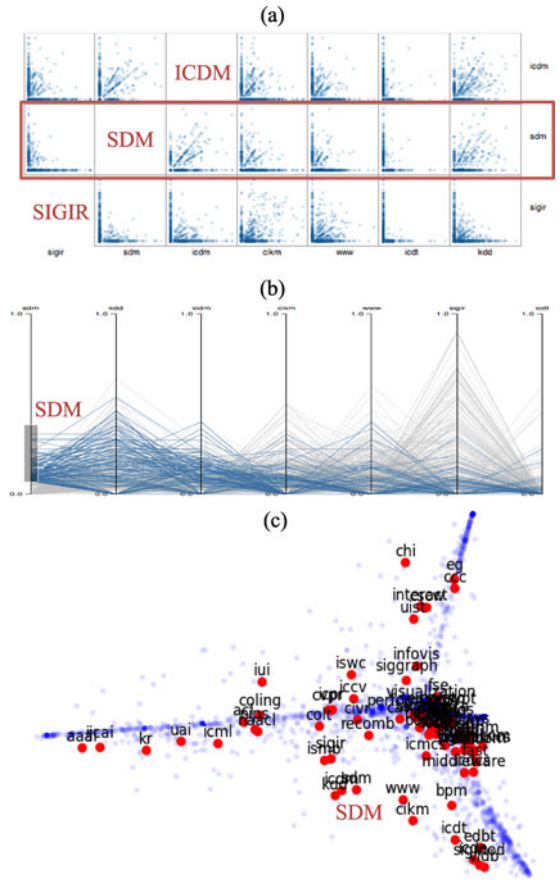

Fig. 15. Visualizing author distribution through (a) scatterplot matrix, (b) parallel coordinates plot, and (c) PCA projection.

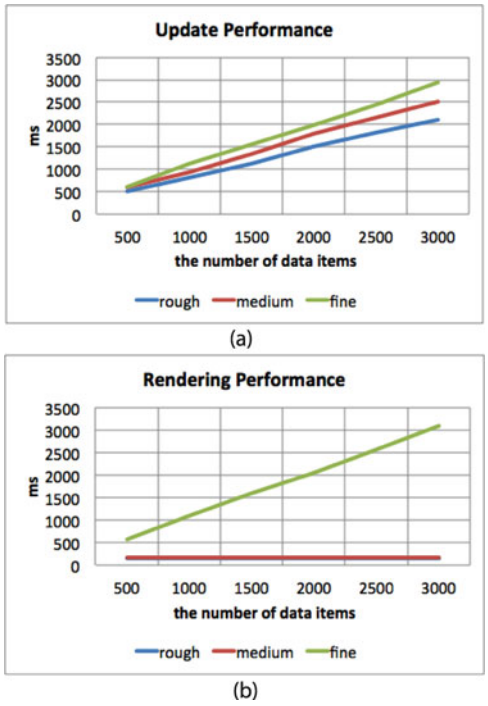

Fig. 16. Performance testing of UnTangle heatmap. (a) the performance of heatmap updating (regenerate data bins and recompute the data density distribution on top of these bins) (b) the performance of rendering.

on Fig. 15a, it is difficult to understand, overall, which conference has greater associations with other conferences. The dominance of the KDD conference among this set cannot be easily revealed in SPM because the information, spread across many different axes, is not easy to visually aggregate to identify dominant labels.

Fig. 15b shows these same conferences using PCP. Each author is plotted as a line segment crossing the axes which correspond to the probability of the author publishing at individual conferences. PCP is not as effective when there are large numbers of either data items or coordinates. Yet, with proper filtering, it is possible to discover strong associations. For example, in Fig. 15b, one can find that SDM shares many co-participants with KDD and ICDM. However, the zero probabilities of the authors in other conferences also form strong patterns in PCP that hinders the discovery of more useful information.

Fig. 15c shows the PCA projection of the entire DBLP dataset. When compared with Fig. 1, the PCA view is too cluttered for pattern detection and the position of each point is also unclear and fails to precisely (or quantitatively) reveal how exactly one researcher is related to different conferences.

As shown in Fig. 14, UnTangle Map is able to resolve these issues. On one hand, the ternary meshes allow data items to scatter over the probability value space; on the other hand, the meshes connected by labels (similar to axes or coordinates) allow patterns to be visually aggregated and form a visual summary of the labels.

6.2 Scalability Improvement via Ternary Heatmap

We evaluate the performance of updating and rendering the UnTangle Map at three different levels of granularity (rough, medium, and fine) based on the ternary heatmap rendering approach. Here, as shown in Fig. 9, "rough" indicates each ternary plots is portrayed by four sub-triangular bins, "medium" indicates 12 sub-triangular bins, and "fine" indicates the raw data items without using heatmap. Our experiments are based on the DBLP dataset described above, and we select subsets of data items with sizes ranging from $[500, 3{,}000]$ with an increasing step of 500. Fig. 16a reports the performance (in rendering time) of updating a visualization at each of the three levels of granularity. The results show that the time grows linearly with the number of data items. Fig. 16b reports the rendering time only. This shows that rendering time remains constant as the number data items increase for the heatmap-based approaches (rough and medium).

6.3 Case Studies

This section presents two use cases that demonstrate the ways in which UnTangle Map can help identify patterns in real-world datasets. These cases are manually generated by an expert user by using the interactions supported in UnTangle Map.

6.3.1 *Use Case: DBLP Data*

When drilling in to a specific set of conferences in the DBLP data through interaction, we can explore the co-participants among data-mining conferences. Fig. 14 shows six data mining conferences along with a database conference (ICDT) that has some ties to the data mining community. When ICDT is placed in the middle, as seen in Fig. 14a,

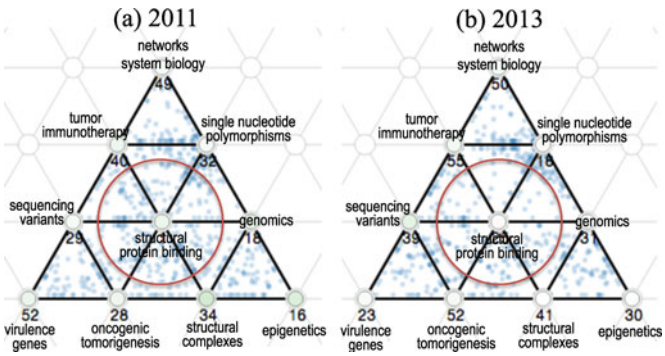

Fig. 17. University distribution among NIH-funded grant topics in year 2011 and 2013.

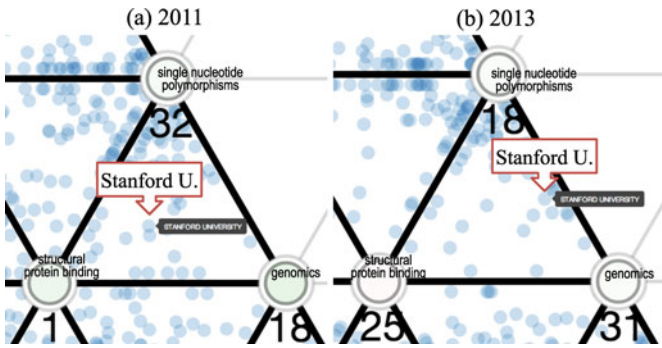

Fig. 18. An example of topic shift: Stanford University was positioned closer to the topic "structural protein binding" in 2011 than in 2013.

most author dots are found away from the center. This indicates that ICDT is relatively isolated compared to the data mining conferences. When centered on SDM (Fig. 14b), strong linear patterns appear along the edges connecting SDM with ICDM and KDD. This indicates that SDM frequently shares common participants with those two conferences. The evenly distributed dots on the KDD-centered mesh suggest that many authors who published in other data mining conferences also published in KDD (Fig. 14c). Another two conferences, WWW and CIKM, also share a lot of authors with other conferences, but have fewer authors in common with SDM (Figs. 14d and 14e)). The ICDM-centered mesh also exhibits evenly distributed patterns (Fig. 14f), but the dots around the center are sparser than those in the KDD-centered mesh. This suggests that ICDM is less dominant than KDD—there are a number of authors who have primarily published in KDD, but fewer who have only published in ICDM.

This exploration suggests how UnTangle Map can be used to explore the interaction among conferences based on the distribution of co-participating authors.

6.3.2 Use Case: NIH Data

In the second example, we use data downloaded from the NIH Map Viewer.[7] The data consist of information about grants awarded by the National Institutes of Health, including the awarded universities and topics associated with the grants (the topics are identified using Latent Dirichlet Allocation [38]). Here we focus on the relationship of topics and universities. We plot the universities as dots on the meshes of topics, based on the normalized topic proportions given by the topic modeling. In this case, we consider topics as probabilistic labels.

Figs. 17a and 17b show the distribution of universities among topics related to cancer, genetics and system biology related research, based on the grants awarded in the years of 2011 and 2013, respectively. By comparing the two plots, we observe that the topic "structural protein binding" was more dominant 2011 because the dots surrounding this topic appear to be sparser on the 2013 mesh. One of the universities we observe having such topical shift in awarded grants is Stanford University, which was placed closer to "structural protein binding" in 2011 but farther away in 2013 (Figs. 18a and 18b). Other topics are relatively stable, with some universities appearing to have increasingly strong interests in the topic "single nucleotide polymorphisms" (Figs. 17a and 17b).

6.4 User Study

To evaluate both the benefits and limitations of UnTangle Map, we conducted a formal user study that compared user performance on five distinct tasks using UnTangle Map and two commonly used baseline visualization techniques: scatter-plot matrices and parallel coordinate plots (PCP). In this section, we review the methodology we employed in our study and present a discussion of our key findings.

6.4.1 Study Setup

We conducted a formal user study to evaluate how well the UnTangle Map method supported five specific visual comprehension tasks. We recruited 10 people to participate in a within-subjects study comparing three distinct visualization techniques: UnTangle Map, SPM, and PCP. The ages of the participants ranged from 26 to 40, all were college educated, and four of 10 were female.

As is typical of a within-subjects study, each of the 10 participants was asked to perform each of the five tasks multiple times, once for each of the three visualization techniques being tested (UnTangle Map, SPM, PCP). Each of the three visualization types were provisioned with the same set of user interaction capabilities for label selection, axis reordering, and interactive brushing. For each task, we selected a single dataset for analysis (using one of the real data sets described in Section 6.3). We used the same dataset with all three visualization types for a given task to ensure a fair comparison. However, to avoid learning effects and to prevent users from applying background knowledge to solve the tasks, we replaced semantically meaningful label

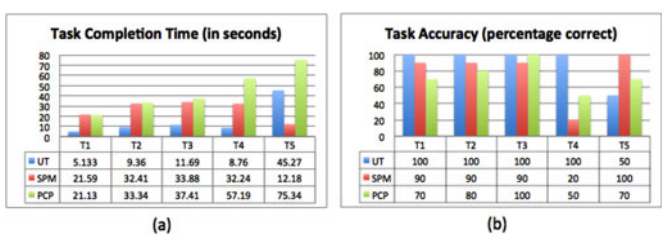

Fig. 19. Results for each of the five user study tasks (T1-T5) using UnTangle Map (UT), SPM, and PCP: (a) average response time measured in seconds, and (b) average response accuracy.

7. https://app.nihmaps.org/

TABLE 1
The Five Comprehension Tasks Performed by Subjects in Our Evaluation

Task	Aim	Description
T1	Isolated label	Which label, overall, is the weakest component in the probability vectors?
T2	Conditional probability, 1 prior	Given A, which has a stronger probability: B or C?
T3	Conditional probability, 2 priors	Given A and B, which has a stronger probability: C or D?
T4	Dominant label	Which label, overall, is the strongest component in the probability vectors?
T5	Pairwise correlation	Which label most strongly reflects linear correlation with a given label A?

names (e.g., conference names) with neutral identifiers (e.g., "I23") that were randomly re-assigned between treatments. This approach ensured that, for each of the three visualization types for a given task, users were answering the same question using the same data, but were unable to learn the correct answers.

Each of the 10 study sessions followed the same procedure. Subjects were first introduced to the study and shown an example of a probabilistic multi-label dataset. Next, participants were given brief lessons for each of the three visualization tools. This included lessons on basic interaction with the tools and techniques for visually identifying patterns. Then users were given equal time to practice with each of the three visualization types.

Data were then collected for the five official study tasks. Each task was repeated three times, once for each of the tested visualization tools. Speed and accuracy were recorded for each task. If a user gave up on a task, the time was listed as 120 seconds, a time roughly equal to the maximum time spent by a user on any single task in our experiments. This occurred three times out of a total of 150 individually performed and measured tasks. A post-study questionnaire was completed at the conclusion of each session to gather subjective feedback from the study participants.

6.4.2 Study Tasks and Results

Every participant in the user study was asked to perform five different comprehension tasks. Importantly, the chosen study tasks, summarized in Table 1, were *not* selected to be a comprehensive representation of all types of questions that analysts might ask when analyzing probabilistic multi-label dataset. Rather, the five tasks capture a subset of common tasks for which we hypothesized that UnTangle Map would be particularly well (T1-4) or poorly (T5) suited. In this way, the study was designed to identify strengths and weaknesses of the proposed approach, helping to frame where the method can be used to compliment capabilities provided by other existing techniques. All the statistics reported below are based on the paired t-test (for within-subject study) and they verified all our hypothesizes (see Fig. 19).

T1: Isolated label identification. In this task, users were asked to identify the label that was most isolated from the rest of the dataset. That is, the label for which the most data points had the lowest probabilities. For example, in the DBLP dataset where labels represent conferences, the isolated conference would be the one at which the set of authors were least likely to publish. This task was accurately performed with all three visualization tools included in the study. Accuracy rates were all 70 percent or above with no statistically significant difference. However, more meaningful differences were found in task completion time. Users performed significantly faster ($p < 0.05$) with UnTangle Map than with either SPM or PCP, both of which exhibited similar timings. This tells us that while all three tools support T1, UnTangle Map required the least mental processing to arrive at the correct answer.

T2: Conditional probability with one prior. In this task, users were asked to identify which of two labels had, overall, a stronger probability given a prior relationship to a third label. For example, in a dataset of paper authors where labels represent conferences, users might want to know at which of two different conferences are authors most likely to publish given that we know that they already published in a third conference. We hypothesized that this type of task was especially well suited for UnTangle Map given the triangular representation of the axes, and the speed measurements provided statistically significant ($p < 0.05$) evidence when compared against either SPM or PCP. Similar to the results of T1, while users indeed performed the tasks faster with UnTangle Map, accuracy rates did not show any significant variation as the task, in general, was correctly performed across all three tools.

T3: Conditional probability with two priors. Like task T2, this task focused on conditional probability. However, this time users were asked to consider problems with two priors (e.g., authors known to publish at two conferences). Users with UnTangle Map could answer this question by examining two neighboring triangles that share a side defined the two priors. This capability resulted in statistically faster task completion times ($p < 0.05$) for UnTangle Map when compared to either PCP or SPM. Once again, there was no statistical difference in terms of accuracy. For both T2 and T3, the accuracy measurements were somewhat unexpected. We had hypothesized that task accuracy for conditional probability tasks would be higher with UnTangle Map. However, the results did not show any statistically meaningful differences. The accuracy gap was diminished in part, we believe, by much longer times (approximately triple) spent answering questions when using either PCP or SPM. We speculate that in many practical settings, where time does not allow users to meticulously investigate a specific question, PCP and SPM would indeed be more prone to errors.

T4: Dominant label identification. In this task, users were asked to identify which label was most strongly represented, in that it dominated the probabilities compared to other labels. For example, in the DBLP dataset where labels represent conferences, the dominant label would be the one at which authors, in general, are most likely to publish. Interestingly, results for this task showed a strong benefit for UnTangle Map in terms of both speed ($p < 0.05$

compared to PCP) and accuracy ($p < 0.05$ compared to both PCP and SPM). We believe that the better performance is due to the spatial arrangement produced by the UnTangle Map layout algorithm, in which triangles are positioned radially around a point representing an individual label. Users were able to accomplish task T4 with UnTangle Map by visually searching for clusters of points gathered around a single vertex. In comparison, PCP and SPM required users to work to mentally integrate information located in various regions of the visualization before arriving at an answer.

T5: Pairwise correlation. In this task, users were asked to identify the pair of labels with the strongest linear correlation. This task was chosen to test our hypothesis that pairwise correlation was a feature for which UnTangle was especially poorly suited. The results of our study confirmed this limitation. SPM was better than UnTangle ($p < 0.05$) in terms of accuracy, and better than PCP in terms of speed ($p < 0.05$). As one would expect, SPM was clearly the right tool for identifying and comparing pairwise correlations.

6.4.3 Qualitative Feedback

The formal tasks provided quantitative data to compare UnTangle with both PCP and SPM techniques. To compliment these measurements, we asked each study participant to complete a short post-study questionnaire in which we asked for qualitative feedback about the different visualization tools. In general, feedback regarding UnTangle was positive. In terms of ease of interpretation, users gave a usability score of 6.2 out of 7. Similarly, users found the tool were useful, responding with an average 6.1 out of 7. When comparing to both SPM and PCP, users felt strongly that UnTangle conveyed certain insights that were harder to see in other tools (6.5 of 7 for SPM, 6.6 of 7 for PCP).

The features most frequently identified as valuable for UnTangle were (a) an overview of correlations at a glance, (b) the ability to show relationships along more than two labels. Participants were nearly unanimous in identifying the most significant limitation of UnTangle: pairwise linear correlations were hard to detect. This is directly reflected in the quantitative study results as well, as seen in the T5 results.

Another limitation identified in the questionnaire was the limit of six neighboring labels in UnTangle visualizations due to the regular grid used for vertex layout. Fortunately, this issue can be alleviated as users become more familiar with the interaction functions and learn to dynamically manipulate the label positions.

7 Conclusion

In this paper, we presented a novel design, UnTangle Map, for visualizing data with probabilistic labels. Our design extends the traditional ternary plot 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. In addition to the basic design, we described a number of archetype patterns and their interpretations. We also demonstrated, using three real-world probabilistic multi-label datasets, how our design provides a synoptic view of the data and, at the same time, helps identify meaningful relationships between items and labels. User evaluation results were presented, indicating our design outperforms two widely-used baseline tools in several information-seeking tasks performed with probabilistic multi-label data. Nevertheless, our design has limitations, especially in identifying pairwise linear relationship between labels. As part of our future work, we plan to extend UnTangle Map's capability of addressing more related information seeking and comparison tasks, and a more comprehensive user study to evaluate these extended capability. We will also explore the combination of UnTangle Map with other visualization techniques (such as scatter plots, bar charts and line graphs) in order to facilitate the exploration of probabilistic labels in combination with other types of variables (e.g., numerical and categorical).

Acknowledgments

This material is partly supported by the US Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication program under Contract Number W911NF-12-C-0028.

References

[1] M.-S. Yang, "A survey of fuzzy clustering," *Math. Comput. Model.*, vol. 18, no. 11, pp. 1–16, 1993.
[2] S. L. Feld, "The focused organization of social ties," *Am. J. Sociol.*, vol. 86, pp. 1015–1035, 1981.
[3] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield, "Scatterplot matrix techniques for large N," *J. Am. Statistical Assoc.*, vol. 82, no. 398, pp. 424–436, 1987.
[4] A. Inselberg and B. Dimsdale, "Parallel coordinates for visualizing multi-dimensional geometry," in *Proc. Int. Conf. Comput. Graph.*, 1987, pp. 25–44.
[5] I. Borg, *Modern Multidimensional Scaling: Theory and Applications*. New York, NY, USA: Springer, 2005.
[6] J. Sharko and G. Grinstein, "Visualizing fuzzy clusters using RadViz," in *Proc. Int. Conf. Inform. Vis.*, 2009, pp. 307–316.
[7] Y.-R. Lin, N. Cao, D. Gotz, and L. Lu, "UnTangle: Visual mining for data with uncertain multi-labels via triangle map," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 340–349.
[8] P. C. Wong and R. D. Bergeron, "30 years of multidimensional multivariate visualization," in *Proc. Sci. Vis.*, 1994, p. 333.
[9] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1539–1148, Nov./Dec. 2008.
[10] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen, "Visual clustering in parallel coordinates," *Comput. Graph. Forum*, vol. 27, no. 3, pp. 1047–1054, 2008.
[11] M. R. Berthold and L. O. Hall, "Visualizing fuzzy points in parallel coordinates," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 3, pp. 369–374, Jun. 2003.
[12] D. Holten and J. J. Van Wijk, "Evaluation of cluster identification performance for different PCP variants," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 793–802, 2010.
[13] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *Proc. IEEE Symp. InfoVis*, 2004, pp. 89–96.
[14] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Comput. Stats.*, vol. 2, no. 4, pp. 433–459, 2010.
[15] Y. Koren and L. Carmel, "Visualization of labeled data using linear transformations," in *Proc. IEEE Symp. Inf. Vis.*, 2003, pp. 121–128.
[16] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.
[17] A. Kovács and J. Abonyi, "Vizualization of fuzzy clustering results by modified Sammon mapping," in *Proc. CINTI*, 2002, pp. 177–188.
[18] C. J. Bradley, *The Algebra of Geometry: Cartesian, Areal and Projective Co-ordinates*. Buckingham, England: Highperception Limited, 2007.

[19] S. G. Kobourov, "Spring embedders and force directed graph drawing algorithms," *arXiv preprint arXiv:1201.3011*, 2012.
[20] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learning Res.*, vol. 9, no. 2579–2605, p. 85, 2008.
[21] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Trans. Vis. Comput. Graph.*, vol. 6, no. 1, pp. 24–43, Jan.–Mar. 2000.
[22] M. Ghoniem, J. Fekete, and P. Castagliola, "A comparison of the readability of graphs using node-link and matrix-based representations," in *Proc. IEEE Symp. Inform. Vis.*, 2004, pp. 17–24.
[23] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1998.
[24] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.
[25] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[26] M. R. Berthold, B. Wiswedel, and D. E. Patterson, "Interactive exploration of fuzzy clusters using neighborgrams," *Fuzzy Sets Syst.*, vol. 149, no. 1, pp. 21–37, 2005.
[27] U. Brandes, S. Cornelsen, B. Pampel, and A. Sallaberry, "Path-based supports for hypergraphs," *J. Discrete Algorithms*, vol. 14, pp. 248–261, 2012.
[28] B. Alper, N. H. Riche, G. Ramos, and M. Czerwinski, "Design study of linesets, a novel set visualization technique," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2259–2267, Dec. 2011.
[29] C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Trans. Visualization Comput. Graph.*, vol. 15, no. 6, pp. 1009–1016, Nov./Dec. 2009.
[30] P. Xu, F. Du, N. Cao, C. Shi, H. Zhou, and H. Qu, "Visual analysis of set relations in a graph," *Comput. Graph. Forum*, vol. 32, no. 3pt1, pp. 61–70, 2013.
[31] N. Cao, D. Gotz, J. Sun, and H. Qu, "DICON: Interactive visual analysis of multidimensional clusters," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2581–2590, Dec. 2011.
[32] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, and S. Hochreiter, "Furby: Fuzzy force-directed bicluster visualization," *BMC Bioinformatics*, vol. 15, no. Suppl 6, p. S4, 2014.
[33] R. Hammah and J. Curran, "Fuzzy cluster algorithm for the automatic identification of joint sets," *Int. J. Rock Mech. Mining Sci.*, vol. 35, no. 7, pp. 889–905, 1998.
[34] Y.-R. Lin, J. Sun, N. Cao, and S. Liu, "ContexTour: Contextual contour visual analysis on dynamic multi-relational clustering," in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 418–429.
[35] P. Simonetto, D. Auber, and D. Archambault, "Fully automatic visualisation of overlapping sets," *Comput. Graph. Forum*, vol. 28, no. 3, pp. 967–974, 2009.
[36] N. H. Riche and T. Dwyer, "Untangling euler diagrams," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1090–1099, Nov./Dec. 2010.
[37] G. Stapleton, P. Rodgers, J. Howse, and L. Zhang, "Inductively generating euler diagrams," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 1, pp. 88–100, Jan. 2011.
[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learning Res.*, vol. 3, pp. 993–1022, 2003.
[39] S. T. Dumais, "Latent semantic analysis," *Annu. Rev. Inform. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.
[40] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "TextFlow: Towards better understanding of evolving topics in text," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2412–2421, Dec. 2011.
[41] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted visualization for rich text corpora," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1172–1181, Nov./Dec. 2010.
[42] N. Cao, D. Gotz, J. Sun, Y.-R. Lin, and H. Qu, "SolarMap: Multifaceted visual analytics for topic exploration," in *Proc. IEEE Int. Conf. Data Mining*, 2011, pp. 101–110.
[43] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann, "The infosky visual explorer: Exploiting hierarchical structure and document similarities," *Inform. Vis.*, vol. 1, no. 3/4, pp. 166–181, 2002.
[44] Y. Chen, L. Wang, M. Dong, and J. Hua, "Exemplar-based visualization of large document corpus (InfoVis2009-1115)," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1161–1168, Nov./Dec. 2009.
[45] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Proc. Inf. Vis.*, 1995, pp. 51–58.
[46] T. Iwata, T. Yamada, and N. Ueda, "Probabilistic latent semantic visualization: Topic model for visualizing documents," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 363–371.
[47] X. Geng, D.-C. Zhan, and Z.-H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 35, no. 6, pp. 1098–1107, Dec. 2005.
[48] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Non-linear dimensionality reduction techniques for classification and visualization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 645–651.
[49] P. Rheingans and M. Desjardins, "Visualizing high-dimensional predicitive model quality," in *Proc. Int. Conf. Vis.*, 2000, pp. 493–496.
[50] C. Seifert and E. Lex, "A novel visualization approach for data-mining-related classification," in *Proc. Int. Conf. Inform. Vis.*, 2009, pp. 490–495.
[51] B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch, and A. Rauber, "Visual methods for analyzing probabilistic classification data," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1703–1712, Dec. 31, 2014.
[52] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

Nan Cao is a research staff member at IBM T.J. Watson Research Center. His research interests include data visualization, visual analysis, and data science. He creates novel visualizations for representing complex (i.e., big, dynamic, multivariate, heterogeneous, and multirelational) graph data in the domains of social science and medical informatics.

Yu-Ru Lin is an assistant professor in the School of Information Sciences, University of Pittsburgh. Her research interests include human mobility, social and political network dynamics, and computational social science. She has developed computational approaches for mining and visualizing large-scale, time-varying, heterogeneous, multirelational, and semistructured data.

David Gotz is an associate professor in the School of Information and Library Science, University of North Carolina at Chapel Hill. He is an assistant director in the Carolina Health Informatics Program, and an associate member in the UNC Lineberger Cancer Center. His research interests include data visualization, visual analytics, data science and analysis, and medical informatics.