UnTangle: Visual Mining for Data with Uncertain Multi-labels via Triangle Map

Yu-Ru Lin*, Nan Cao†, David Gotz‡ and Lu Lu§
*University of Pittsburgh
†IBM TJ Watson Research Center
‡University of North Carolina at Chapel Hill
§Hong Kong University of Science and Technology

***Abstract*—Data with multiple uncertain labels are common in many situations. For examples, a movie may be associated with multiple genres with different levels of confidence, and a protein sequence may be probabilistically assigned to several structural subcategories. Despite their ubiquity, the problem of visualizing uncertain labels has not been adequately addressed. Existing approaches often either discard the uncertainty information, or map the data to a low-dimensional subspace where their associations with multiple labels are obscured. In this paper, we propose a novel visual mining technique, UnTangle, for visualizing uncertain 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 provides both (a) an automatic label placement algorithm, and (b) adaptive interaction mechanisms that allow users to control the label positioning for different visual queries. Our work makes a unique contribution by providing an effective way to investigate the relationship between data items and their uncertain 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 uncertainty information in the data labels.**

***Index Terms*—visual mining; multi-labels; probablistic labels; uncertainty data; ternary plot**

I. Introduction

Data with multiple uncertain labels are common to many applications. For example, in the movie classification, a movie may be labeled as both an "action" movie and a "comedy", each with different levels of confidence. In market segmentation, a customer may be probobilistically 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 everyday social life, people tend to participate simultaneously in multiple communities such as co-workers, friends, family and extended family members [1]. In all 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 the level of uncertainties for corresponding labels.

Despite the ubiquity of data with uncertain multi-labels, little work has been done in visualizing such data. Existing work generally falls in two paradigms: (a) visualizing data through a set of independent coordinates, or (b) mapping data to a dimension-reduced plane for visualization. Scatterplot matrices [2] and parallel coordinates [3] are representatives for the first paradigm. In the second paradigm, techniques such as multidimensional scaling [4] and RadVis [5], [6] are used to project high dimensional data onto a low dimensional (2D or

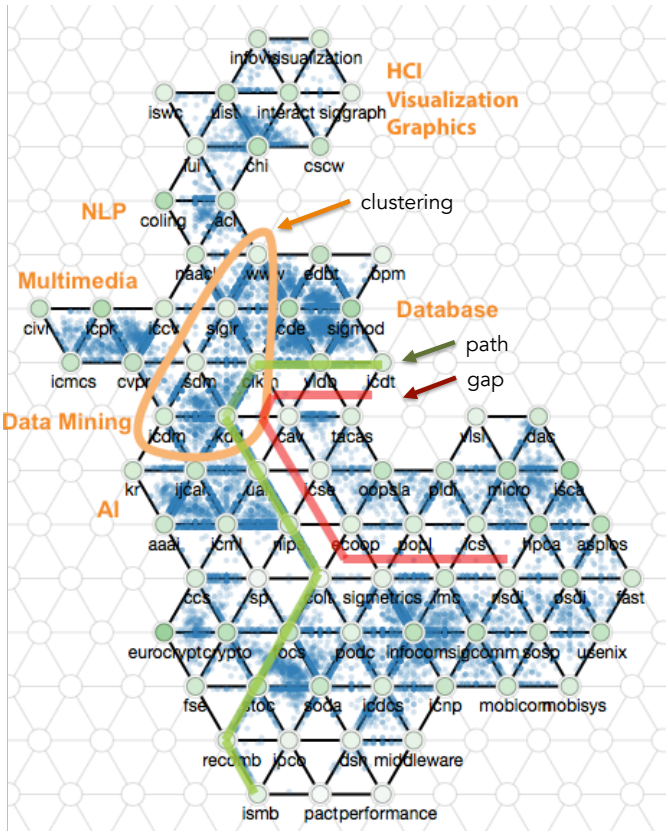

Fig. 1. Visualizing data with multiple uncertain labels via UnTangle. The data shown here is from DBLP and consists of authors and conferences in Computer Science. We consider conferences as uncertain labels because authors are likely to publish in multiple conferences with different probabilities. Here, uncertain labels (conferences) are placed at triangle vertices and data items (authors) are scattered as points inside each of 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 in this case include: the clustering (highlighted in orange) of research communities dealing with various aspects of data, the gap (in red) between the software engineering communities and data-centric communities at the top left, and a long path (in green) that connects conferences through authors in different areas according to their co-participation in these conferences.

3D) subspace.

We argue that for visualizing uncertain multi-labels, there are significant drawbacks with each of these paradigms. First, while a set of independent coordinates is useful for discovering the correlation between labels, it is not easy to show higher level summaries among labels (e.g., which labels are the most dominant or isolated). Second, while using a dimension-reduction technique may help convey proximity between items and labels (if the labels are also mapped onto the same plane), the relationship between an item with the set of labels is ambiguous due to loss of information that occurs as part of the reduction process.

In this paper, we propose a novel visual mining technique, *UnTangle*, specifically designed for visualizing data with uncertain multi-labels (see Fig. 1). In this visualization, we generate a web of connected triangles called ternary plots [7] and place labels on the triangle vertices. Data items are placed inside each of the triangle according to the items' probabilistic associations with the labels corresponding to the vertices. UnTangle can automatically arrange the vertices such that the relationships among labels and the patterns of items across relevant labels can be easily identified. We also provide a set of interaction functions that allow users to interactively control the label positioning, to link items across labels, or to focus on a particular set of labels. Our proposed technique displays uncertain label information through a set of marginal 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.

The key contributions in this paper include: **(1) Novel visualization design:** We identify the main challenges in visualizing data with uncertain multi-labels and propose a novel visual design, UnTangle, that addresses those challenges. In particular, our new design leverages the ideas of independent coordinates and subspace creation in order to support several visual query tasks in an uncertain label dataset. **(2) Automatic label placement:** We propose a layout scheme that arranges labels' placement either in a data-driven manner or in a user-driven manner. In a data-driven manner, the layout of labels is generated automatically based on the intrinsic correlation structure of the data through an efficient (linear-time) algorithm. We show that our algorithm produces high quality layouts on empirical datasets. In a user-driven manner, users can customize the labels' placements through a set of interactive functions. Throughout the interaction process, the system provides recommendations for choosing subsequent labels and their corresponding placements, thereby using data-driven computation to inform users' decisions. **(3) Extensive evaluation:** We demonstrate the features of our approach through case studies on two empirical uncertain multi-label datasets. Further, we conduct an extensive user evaluation on UnTangle's effectiveness in different visual query tasks. We compare UnTangle with two other baseline tools and show that our new design has more satisfactory results overall.

The rest of this paper is organized as follows. We first discuss related work in Section II, followed by the problem formulation in Section III. We present our design and rationale in Section IV, and the label placement algorithm in Section V. We present the evaluation in Section VI that includes algorithm evaluation (Section VI-A), case studies (Section VI-B) and user study (Section VI-C). Finally, Section VII concludes the paper and discusses possible future directions.

II. Related Work

In this section we review work related to the visualization of uncertain labels. This includes visualization aimed at multidimensional data, fuzzy clusters, and topic modeling results.

Multidimensional Data. One approach to visualizing uncertain labels is to use methods designed for multidimensional or multivariate data (mdmv) [8]. 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 [2] and parallel coordinates [3]. Scatterplot matrices [2] represent data items in all pairwise permutations of dimensions such that the relationships between any two specific dimensions can be discovered and compared.

Visualizing uncertain label data is more complex than visualizing categorical data [9]. A categorical variable takes m possible values from a set of category labels $\{l_1, l_2, \ldots, l_m\}$ and each data item is unambiguously associated with one label. Techniques such as dimensional stacking [9] has been used to visualize the relationship of this categorical dimension with other dimensions. When visualizing data with labels, each dimension corresponds to a label, 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. This allows such data, for example, to be visualized through scatterplot matrices.

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 [10] may be used to help users explore the data, discovering relationships among many labels is still not easy. Like scatterplot matrices, parallel coordinates [3] and many of its variants, e.g., [11], [12] are only effective when the number of dimensions is small [13], and clutter reduction is needed for data with many dimensions [14]. 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.

The second paradigm uses dimension reduction to map data into a lower dimensional space for visualization. Multidimensional scaling (MDS) [4] 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. Like MDS, many other techniques for projecting high-dimensional data to a low-dimensional space (often a 2D plane) have been used.

Principal Component Analysis [15] and various linear transformation methods [16] project data by maximizing the variance of data items based on different constraints. Self Organization Maps [17] uses a 2D lattice to portray the distribution of data items in a high-dimensional space via a learning process. A modified Sammon Mapping [18] preserves the distance between data items and cluster centers in a low-dimensional space. RadVis [5], [6] projects the multidimensional data into a barycenter coordinates [19], [20]. 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 and becomes ambiguous (e.g., data items may be placed next to an unrelated label).

Fuzzy Clustering. The visualization of results from fuzzy clustering [21] is closely related to our work. Fuzzy clustering methods assign data items to one or more clusters with a degree of uncertainty (hence the term "fuzzy"). Rousseeuw proposed Silhouettes [22], 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.* [23] extended this design with interactive functions that allow users to select and discard the items in each cluster to find-tune the clustering results. Klawonn *et al.* [24] packed data items inside a cluster area, but instead of a 1D axis, clusters are represented as circles on a 2D plane such that the relationship across different clusters can be reflected as proximity in space. There has also been work that represents fuzzy clusters in a projection space, where contour or lines are used to depict soft cluster boundaries [25], [18], [26]. Simonetto *et al.* [27] and others [28], [29] developed methods to generate Euler-like diagrams for visualizing overlapping clusters. ContexTour [26] uses a contour map to represent the density distribution of data items, showing a smooth and fuzzy margin between two adjacent clusters.

Topic Models. A branch of work closely related to fuzzy clustering is topic modeling applied to text data [30], [31]. Using techniques such as Latent Dirichlet Allocation [30], text documents can be automatically associated with one or more topics for search and organization purpose. Recent advances in topic visualization have focused either on topic transition [32], [33], [26], or on viewing topics across different information facets [26], [34], [35]. In many of these techniques, the probabilistic topic assignment is first converted into a hard assignment for simplicity, and hence they are not suitable for visualizing uncertain labels.

In text visualization, it is common to treat documents as high dimensional data based on the bag-of-word 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) [36], [37], [38]. For example, Iwata *et al.* proposed the probabilistic latent semantic visualization model (PLSV) [39] 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.

Based on the literature, we identify several key challenges for visualizing data with uncertain labels: **(1) Scalability:** The number of uncertain labels 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). Most popular multivariate visualization tools, including scatterplot matrices and parallel coordinates, suffer from the scalability issue. **(2) Subspace ambiguity:** Multidimensional scaling or similar techniques map data items to a low-dimensional subspace, which can distort the original relationships between data items and labels, result in information loss and introduce ambiguity. **(3) 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.

As will be described later, the design of UnTangle seeks to overcome these challenges.

III. Problem Formulation

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

We present below the visualization problem dealing with uncertain multi-labels. Let $(x_i)_{i=1\ldots n} \in X$ be the n data items in data set X. Let $(l_k)_{k=1\ldots m} \in L$ be the m different labels in label set L. Each of the items is associated with 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 a 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. Without loss of generality, we assume $\sum_k p_{ik} = 1$.

We identify visual query tasks in the problem context defined previously. Along with the challenges described in Section II, our work has been motivated by the necessity of supporting visual inquiry tasks on the data with uncertain multi-labels. The tasks include:

Q1. Item-label relationship: How do data items associate with many different labels? How strong, in a probabilistic sense, is a data 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 frequently?

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.

IV. Visualization Design

We describe the visual design for UnTangle and the rationale behind our design. We illustrate how the design can generate meaningful visual patterns that achieve the query tasks, and present a set of intersection functions that further support our design.

A. 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 connecting ternary plots.

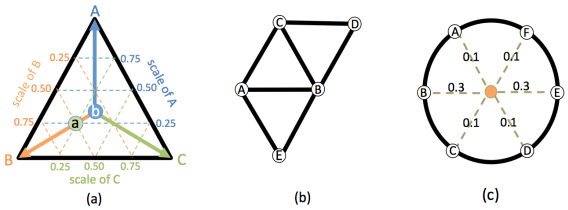

Fig. 2. (a) A ternary plot and the 3D barycentric coordinate system. (b) A ternary plot mesh. (a) Ambiguity is unavoidable when the number of labels (dimensions) is larger than 3.

A ternary plot, as illustrated in Fig. 2(a), 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%, and the position of any given point on the triangle indicates the ratios of three variables. UnTangle builds upon basic ternary plots to visualize data items with uncertain 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 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 illustrated in Fig. 2(a), there are three labels A, B, and C, plotted on the vertices, and 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 $1/3$.

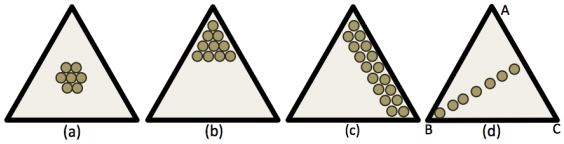

Fig. 3. Typical patterns for item-label relationship: (a) non-dominant, (b) uni-dominant, (c) bi-dominant, and (d) balanced flow patterns.

When labels are more than three, we combine multiple ternary plots where the set of vertices correspond to the set of labels, and within each ternary plots, data items are positioned according to the marginal distribution with the three labels corresponding to the plot's three vertices. The result is a mesh of connected triangles as shown in Fig. 2(b). Inside each individual triangle, the ternary plot provides 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 visual patterns (described below) around the vertices and edges that connect triangles.

This design is based on the consideration of avoiding ambiguity. Particularly, the three dimensional barycentric coordinate system in a ternary plot makes the position of each item, representing its probabilistic associations with the three corresponding labels, unambiguous in the two-dimensional plane. Note that when the barycentric coordinate system is laid in more than three vertices (an n-dimensional with $n > 3$) on a 2D plane, ambiguity is unavoidable. For example, Fig. 2(c) 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.

B. Visual Patterns

We illustrate how our design goal can be achieved through the range of visual patterns that emerge from UnTangle.

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 uncertain data points within a ternary plot: (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) uni-dominant pattern: the data items are concentrated at a corner where the closest label has a dominant relationship with the items; (c) bi-dominant pattern: the data items are located along an edge where the two closest labels both have strong associations with the items; (d) balanced flow pattern: two labels (A and C) have equally strong associations with data items regardless of the strength of the third label (B). In a balanced flow pattern, the data items 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.

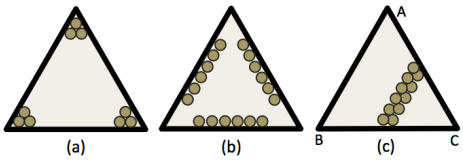

Fig. 4. Other patterns for item-label relationship: (a) three-corner, (b) three-edge, and (c) constant patterns.

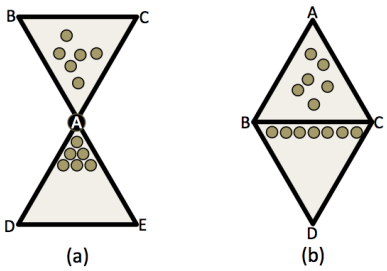

Fig. 5. Typical patterns for higher-level label interactions: (a) shared vertex and (b) shared edge patterns.

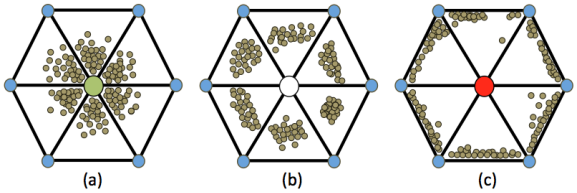

Fig. 6. Typical patterns for multi-way label interactions: (a) global-dominant, (b) complimentary, and (c) isolated patterns.

Variants of the four archetypes defined above can also be highly informative. For example, Fig. 4(a) 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. 4(b) 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. 4(c) 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 allows users to interpret higher-level label interactions (Q4). As shown in Fig. 5(a), 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. 5(a) 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. 5(b), the connected ternary plots allows a user to compare the relationship between two labels (e.g., A or D) given a common baseline (BC). In Fig. 5(b), given the 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 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: (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; (b) 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; (c) 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 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 interesting low- and high-level structures from the data. However, there are some limitations in our design. First, as we will discuss in Section VI-C, linear relationships between two labels are not 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 uncertain labels, and therefore does not consider the visualization of other types of variables (such as numerical or categorical variables). The two limitations can be overcome when integrating with existing tools. 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 provides user interaction capabilities that allow for the the interactive customization of label placements. This interactive feature is described in next section.

C. Interactions

UnTangle provides a set of interactions that further support the process of visual query and data interpretation.

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 visualization, the system automatically generates a initial layout, arranging labels on a triangle mesh according to the internal distribution of the data items. This primary view is augmented with an inset window that shows an overview of all available data labels. 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 remove 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 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 V.

Switch of correlation measure. By default, Spearman's correlation coefficient is used as the basis for the algorithms within UnTangle. 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 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.

V. Label and Item Placement

In this section, we describe our method for positioning the data items and labels in the UnTangle visualization.

A. Displaying Items on Ternary Plots

The data items are displayed on a ternary plot based on the barycentric coordinate system. 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 computed as marginal distributions over the three corresponding labels. Concretely, given a dataset with m labels $\{l_1, \ldots, l_m\}$, a data item's associations with any three labels l_1, l_2, and l_3 are given by the following renormalization:

$$p_{ik}^{(l_1 l_2 l_3)} = p_{ik} / \sum_{k' \in \{1,2,3\}} p_{ik'},$$

where p_{ik} is the probability of the i-th data item for the label $k \in \{1,2,3\}$, and the denominator represents the join association of the data item with respect to the three labels. In Untangle, we use opacity to encode the information about this join association for each data point.

B. Generating the Layout of Labels

The label layout is generated in two steps: (1) creating a triangle grid, and (2) allocating labels to the grid slots.

Creating triangular grid. We begin by creating a grid of equilateral triangles based on triangular tiling [40]. Such a grid provides efficient spatial indexing so that the grid coordinates can be easily used for allocating labels (either in a data-driven or user-driven manner).

Theoretically, this approach would support an infinitely large grid to support the allocation of an unlimited number of labels. In practice, we create a grid on a virtual plane that is sized five times larger than the viewport. We then only show a portion of the grid on the viewport at a given time. This virtual plane can be navigated through the zoom and pan interaction functions as described in Section IV-C. Our experience shows that, in practice, this approach provides more than enough visualization space for our user population.

Allocating labels to the grid 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. Let the layout of m labels $L = \{l_1, \ldots, l_m\}$ be $G_L = \langle V, E \rangle$ where $V = \{v_1, \ldots, v_m\}$ is the set of label vertices located on the grid slots $\{s(v_1), \ldots, s(v_m)\}$. 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. Our objective is to find G_L such that the correlation (sharing of items) among nearby labels are maximized, i.e.,

$$G_L = \text{argmax}_{V,E} \sum_{(v_i,v_j)\in E} c_{ij},$$

where c_{ij} is the correlation between two labels l_i and l_j. Specifically, c_{ij} can be computed as the correlation of two probabilistic vectors $\vec{p}_i$ and $\vec{p}_j$. The l-th element in a probabilistic vector $\vec{p}_i$ corresponds to the l-th data item's association with the label i 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 typical in an uncertain label dataset. We use Spearman's rank correlation as default, and provide other types of correlation functions as user-selectable options.

We use a linear-time greedy algorithm to find G_L based on the following search heuristic. Given a list of slots S_A, $|S_A| \geq 2$ on which labels are allocated, the heuristic returns the unallocated slot s_i for label l_i such that:

$$s_i = \text{argmax } (c_{ij} + c_{ik}),$$
$$\text{subject to } s_i \notin S_A, s_j \in S_A, s_k \in S_A,$$
$$s_i \in \Gamma(s_j, s_k), (v_j, v_k) \in E,$$

where $\Gamma(s_a, s_b)$ for a given edge (s_a, s_b) on the grid is the two neighboring slots of the edge. This heuristic searches for an unallocated slot that is adjacent to an existing edges, which tends to choose locations that close a triangle on the grid. When $S_A = \emptyset$, the slot at the center of the grid is returned. When $|S_A| = 1$, only one slot on the grid is allocated, and an arbitrary neighbor of the allocated slot is returned.

When the label layout is generated automatically, the search starts with $S_A = \emptyset$ and stops when all labels have been allocated. The labels are allocated in order such that the m labels $\{l_1, \ldots, l_m\}$ are sorted based on their one-dimensional projected coordinates obtained via multidimensional scaling with the following objective:

$$\max_{z_1,\ldots,z_m} \sum_{i<j} \delta_{ij} ||z_i - z_j||^2,$$

where $\delta_{ij} = 1 - c_{ij}$ represents the lack-of-correlation between labels l_i and l_j, z_a is the one-dimensional coordinate of label l_a. Labels with higher correlations are closer on the sorting list when labels are sorted based on their one-dimensional projected coordinates.

When user interaction is involved, the recommended slot for a label l_i is given by the heuristic search with S_A consisting of the list of allocated slots as they are currently configured.

VI. Evaluation

In this section, we first evaluate the performance of our label placement method. We then present case studies and user study both quantitatively and qualitatively.

A. Label Placement Evaluation

Given the heuristic nature of our proposed algorithm, it is important to evaluate its performance with real-world datasets. We begin this evaluation by defining two performance measures. First, we define the overall correlation between neighbors, M_1, as:

$$M_1 = \frac{\sum_{i<j} I_{ij} c_{ij}}{|I_{ij}|},$$

where c_{ij} is the correlation defined previously. I_{ij} is a binary indicator which returns $I_{ij} = 1$ if labels l_i and l_j are connected on the grid, or zero otherwise.

The overall lack-of-correlation among non-neighbors, denoted as M_2, is defined as:

$$M_2 = \frac{\sum_{i<j} \delta_{ij}/D_{ij}}{\sum_{ij} 1/D_{ij}},$$

where δ_{ij} is the lack-of-correlation measure defined previously, and D_{ij} is the shortest distance between labels l_i and l_j on the grid. The inverse of D_{ij} gives higher weight to the pair (i, j) if l_i and l_j are closer on the grid.

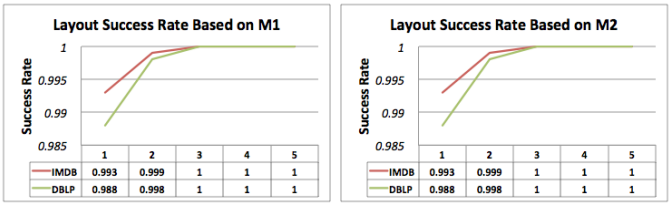

Fig. 7. The success rate for our layout based on the M_1 and M_2 scores.

Based on the definitions of the two measures, a better layout is expected to have higher values in terms of both M_1 and M_2. To evaluate an automatically generated layout produced by our algorithm, we compare it to a randomly perturbed version of itself. The layout is perturbed by randomly selecting n pairs of allocated slots and swapping the labels. We say the layout *succeeds* in such an experiment if the performance value, in terms of M_1 and M_2, of the original layout is higher than that of the perturbed layout. Performing this experiment multiple times allows us to calculate an average success rate that estimates layout performance.

Using this methodology, we tested the layout algorithm's performance with two real-world datasets—DBLP and IMDB—which are described in more detail in Section VI-C. For each dataset, we conducted 1000 experimental iterations and the results are shown in Fig. 7. Success rates are reported for both M_1 and M_2, and with a range of pair swaps, $n = 1, \ldots, 5$. In all cases, the success rate was over 98%, which means, in practice, our greedy algorithm works remarkably well.

B. Case Studies

1) Use Case: DBLP Data: Our first example uses data extracted from DBLP[1], a Computer Science bibliography database. Our dataset includes two types of elements: the *authors* and the *conferences* in which they published papers. Here, we are interested in exploring how authors publish in related conferences and how conferences share common participants. Hence, we consider the conference names as uncertain 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.

When the DBLP dataset is first loaded into UnTangle, the system automatically determines the positioning of the full set of conference labels based on the intrinsic correlation structure in the data. As shown in Fig. 1, the automatically generated layout gives an overview of the dataset where relevant conferences are placed close to one another. Fig. 1(a) shows clusters of conferences that reflect several research communities in computer science, including HCI/Visualization/Graphics, NLP,

[1] http://dblp.uni-trier.de/db/

Multimedia, Data Mining, Database, and AI. The clustering emerges due to the fact that conferences in similar areas tend to share the same groups of authors. Interestingly, the top-left region consists of communities dealing with various aspects of data, and these communities appear to be connected through several Data Mining conferences. Fig. 1(b) highlights a gap between the top-left clusters and the region on the right, which consists of conferences in the related to software engineering. The gap suggests that the software engineering conferences rarely share authors with the more data-centric communities at the top-left. Fig. 1(c) shows a long path connected by edges between conferences that have many authors in common, such as CIKM and KDD, KDD and UAI. This long path provides insights about how authors bridge different areas due to their partially overlapping interests.

Next, drilling in to a specific set of conferences through interaction, we can explore the co-participants among data-mining conferences. Fig. 8 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. 8(a), 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. 8 (b)), 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. 8 (c)). Another two conferences, WWW and CIKM, also share a lot of authors with other conferences, but have fewer authors in common with SDM (Fig. 8 (d,e)). The ICDM-centered mesh also exhibits evenly distributed patterns (Fig. 8 (f)), but the dots around the center are sparser than those in the KDD-centered mesh, suggesting ICDM is less dominant than KDD – there are certain number of authors who primarily published in KDD, but fewer who only published in ICDM.

This exploration suggests how UnTangle can be used to explore the interaction among conferences based on the distribution of co-participating authors. More use case study results can be found in supplementary materials available online [41].

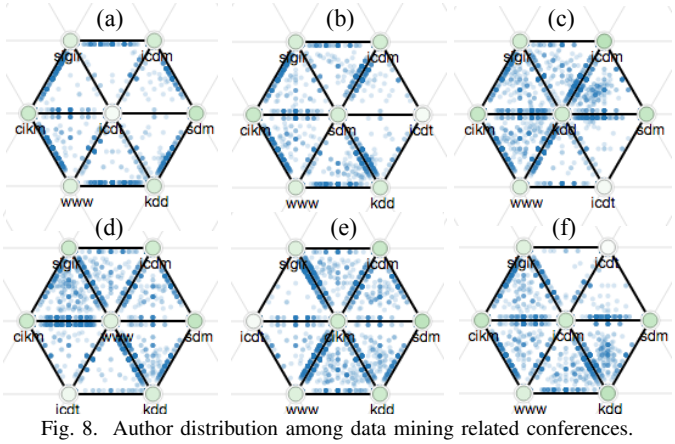

Fig. 8. Author distribution among data mining related conferences.

2) Comparison with PCP and SPM: We use the example shown in Fig. 8 to illustrate the advantages of UnTangle over SPM and PCP. Fig. 9(a) 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 SPC is efficient for discovering piarwise 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. 8(b) by using UnTangle plot. However, based on Fig. 9(a), 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 across many different axes do not visually aggregate to help identify dominant labels.

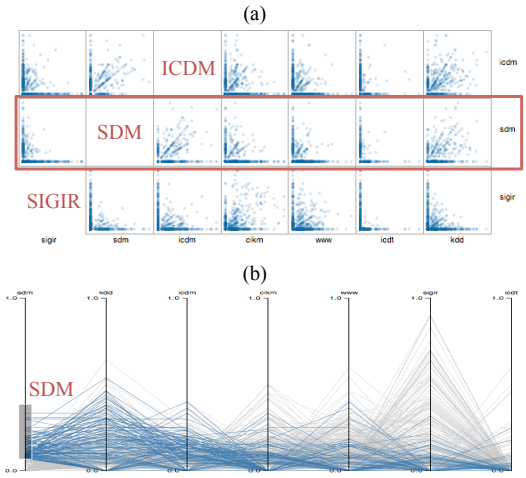

Fig. 9. Visualizing author distribution among data mining related conferences through (a) scatterplot matrix, and (b) parallel coordinates plot.

Fig. 9(b) shows these same conferences using PCP. Each author is plotted as a line segment crossing the axes which coorespond to probability of the author publishing at individual conference labels. PCP is not effective when there are too many data items and too many coordinates. Yet, with proper filtering, it is possible to discover strong associations. For example, in Fig. 9(b), one can find that SDM shared 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.

As shown in Fig. 8, UnTangle 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.

C. User Study

To evaluate both the benefits and limitations of our approach for visualizing uncertain label datasets, we conducted

TABLE I
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?

a formal user study that compared user performance on five distinct tasks using UnTangle and two commonly used baseline visualization techniques: scatterplot matrices (SPM) and parallel coordinate plots (PCP). In this section, we review the methodology in our study and discuss our key findings.

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

As is typical of a within-subjects study, each of the ten participants was asked to perform each of the five tasks multiple times, once for each of the three visualization techniques being tested (UnTangle, SPM, PCP). Each of the three visualization types were given in a counterbalanced order and 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 VI-B). 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 names (e.g., conference names) with neutral identifiers (e.g., "I23") that were randomly re-assigned between treatments. This 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 ten study sessions followed the same procedure. Subjects were first introduced to the study and shown an example of an uncertain label dataset. Next, participants were given brief lessons for each of the three visualization tools. 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.

2) Study Tasks and Results: Every participant in the user study was asked to perform five different comprehension tasks,

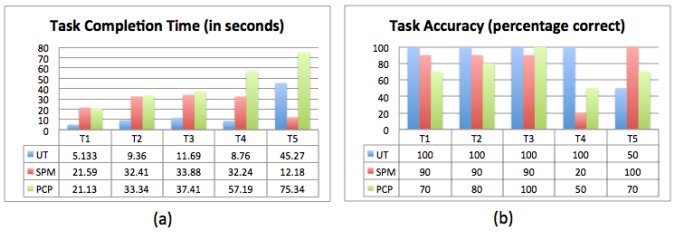

Fig. 10. Results for each of the five user study tasks (T1-T5) using UnTangle (UT), SPM, and PCP: (a) average response time measured in seconds, and (b) average response accuracy.

summarized in Table I. The five tasks were chosen to capture a subset of common tasks for which we hypothesized that UnTangle would be particularly well (or poorly) suited. They were *not* selected to be a comprehensive representation of all types of questions that analysts might ask when analyzing uncertain label dataset. 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 statistical significance reported below are based on the paired t-test (for within-subject study).

Fig. 10 shows the study results. Here, we briefly summarize the results and more detailed discussion can be found in supplementary materials available online [41]. In tasks T1–T3, users performed significantly faster ($p < 0.05$) with UnTangle than with either SPM or PCP, both of which exhibited similar timings. This indicates that while all three tools support these tasks, UnTangle required the least mental processing to arrive at the correct answer. In task T4, the results showed a strong benefit for UnTangle in terms of both speed ($p < 0.05$ compared to PCP) and accuracy ($p < 0.05$ compared to both PCP and SPM). In task T5, the results confirmed that UnTangle was not suitable for displaying pairwise correlations. As one would expect, SPM was clearly the right tool for identifying and comparing pairwise correlations.

VII. CONCLUSION

In this paper, we presented a novel visual mining technique, UnTangle, for visualizing data with uncertain multi-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 labels. We presented the design through a number of archetypical visual patterns and their interpretations. We also demonstrated, using two

real-world uncertain 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 technique outperforms two widely-used baseline tools in several visual query tasks tested with uncertain label data. As part of future work, we will explore the combination of UnTangle with other visualization techniques (such as scatter plots, bar charts and line graphs) in order to facilitate the exploration of uncertain labels in combination with other types of variables (e.g., numerical and categorical).

REFERENCES

[1] S. L. Feld, "The focused organization of social ties," *American journal of sociology*, p. 10151035, 1981.

[2] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield, "Scatterplot matrix techniques for large n," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 424–436, 1987.

[3] A. Inselberg and B. Dimsdale, "Parallel coordinates for visualizing multi-dimensional geometry," in *International Conference on Computer graphics*. Springer-Verlag New York, Inc., 1987, pp. 25–44.

[4] I. Borg, *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[5] L. Nováková and O. Štepánková, "Multidimensional clusters in radviz," in *Proceedings of WSEAS International Conference on Simulation, Modelling and Optimization*, 2006, pp. 470–475.

[6] J. Sharko and G. Grinstein, "Visualizing fuzzy clusters using radviz," in *InfoVis*. IEEE, 2009, pp. 307–316.

[7] "Ternary plot — Wikipedia, the free encyclopedia," [Online; accessed 28-March-2014]. [Online]. Available: http://en.wikipedia.org/wiki/Ternary_plot

[8] P. C. Wong and R. D. Bergeron, "30 years of multidimensional multivariate visualization." in *Scientific Visualization*, 1994, p. 333.

[9] J.-F. Im, M. J. McGuffin, and R. Leung, "GPLOM: the generalized plot matrix for visualizing multidimensional multivariate data," *IEEE TVCG*, vol. 19, no. 12, pp. 2606–2614, 2013.

[10] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE TVCG*, vol. 14, no. 6, pp. 1539–1148, 2008.

[11] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen, "Visual clustering in parallel coordinates," *Computer Graphics Forum*, vol. 27, no. 3, pp. 1047–1054, 2008.

[12] M. R. Berthold and L. O. Hall, "Visualizing fuzzy points in parallel coordinates," *IEEE Trans. on Fuzzy Systems*, vol. 11, no. 3, pp. 369–374, 2003.

[13] D. Holten and J. J. Van Wijk, "Evaluation of cluster identification performance for different pcp variants," *Computer Graphics Forum*, vol. 29, no. 3, pp. 793–802, 2010.

[14] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *InfoVis*. IEEE, 2004, p. 8996.

[15] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Comp Stats*, vol. 2, no. 4, pp. 433–459, 2010.

[16] Y. Koren and L. Carmel, "Visualization of labeled data using linear transformations," in *InfoVis*. IEEE, 2003, pp. 121–128.

[17] T. Kohonen, "The self-organizing map," *Proc. of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[18] A. Kovács and J. Abonyi, "Vizualization of fuzzy clustering results by modified sammon mapping," in *CINTI*, 2002, pp. 177–188.

[19] C. J. Bradley, *The algebra of geometry: Cartesian, areal and projective co-ordinates*. Highperception Limited, 2007.

[20] S. G. Kobourov, "Spring embedders and force directed graph drawing algorithms," *arXiv preprint arXiv:1201.3011*, 2012.

[21] M.-S. Yang, "A survey of fuzzy clustering," *Mathematical and Computer modelling*, vol. 18, no. 11, pp. 1–16, 1993.

[22] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[23] M. R. Berthold, B. Wiswedel, and D. E. Patterson, "Interactive exploration of fuzzy clusters using neighborgrams," *Fuzzy Sets and Systems*, vol. 149, no. 1, pp. 21–37, 2005.

[24] F. Klawonn, V. Chekhtman, and E. Janz, "Visual inspection of fuzzy clustering results," in *Advances in Soft Computing*, 2003, pp. 65–76.

[25] R. Hammah and J. Curran, "Fuzzy cluster algorithm for the automatic identification of joint sets," *International Journal of Rock Mechanics and Mining Sciences*, vol. 35, no. 7, pp. 889–905, 1998.

[26] Y.-R. Lin, J. Sun, N. Cao, and S. Liu, "Contextour: Contextual contour visual analysis on dynamic multi-relational clustering," in *SDM*, 2010.

[27] P. Simonetto, D. Auber, and D. Archambault, "Fully automatic visualisation of overlapping sets," *Computer Graphics Forum*, vol. 28, no. 3, pp. 967–974, 2009.

[28] N. H. Riche and T. Dwyer, "Untangling euler diagrams," *IEEE TVCG*, vol. 16, no. 6, pp. 1090–1099, 2010.

[29] G. Stapleton, P. Rodgers, J. Howse, and L. Zhang, "Inductively generating euler diagrams," *IEEE TVCG*, vol. 17, no. 1, pp. 88–100, 2011.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[31] S. T. Dumais, "Latent semantic analysis," *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.

[32] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM TIST*, vol. 3, no. 2, p. 25, 2012.

[33] 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 TVCG*, vol. 17, no. 12, pp. 2412–2421, 2011.

[34] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "Facetatlas: Multifaceted visualization for rich text corpora," *IEEE TVCG*, vol. 16, no. 6, pp. 1172–1181, 2010.

[35] N. Cao, D. Gotz, J. Sun, Y.-R. Lin, and H. Qu, "Solarmap: Multifaceted visual analytics for topic exploration," in *ICDM*. IEEE, 2011, pp. 101–110.

[36] 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," *Information Visualization*, vol. 1, no. 3-4, pp. 166–181, 2002.

[37] Y. Chen, L. Wang, M. Dong, and J. Hua, "Exemplar-based visualization of large document corpus (infovis2009-1115)," *IEEE TVCG*, vol. 15, no. 6, pp. 1161–1168, 2009.

[38] 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 *InfoVis*. IEEE, 1995, pp. 51–58.

[39] T. Iwata, T. Yamada, and N. Ueda, "Probabilistic latent semantic visualization: topic model for visualizing documents," in *ACM SIGKDD*. ACM, 2008, pp. 363–371.

[40] "Triangular tiling — Wikipedia, the free encyclopedia," [Online; accessed 28-March-2014]. [Online]. Available: http://en.wikipedia.org/wiki/Triangular_tiling

[41] "Online supplementary materials for UnTangle." [Online]. Available: http://goo.gl/TpIDOu