# CS-GY 9223 Project Proposal - Interactive UnTangle Map

Hang Dong
New York University
New York, USA
hd1191@myu.edu

## Abstract

*Probabilistic multi-label data are used in various fields such as statistical classification, fuzzy clustering, and topic modeling. UnTangle Map[1] is a web-based visualization of multi-label data that offers smart label layout on triangular tilling, and demonstarte item-label relationship using barycentric coordinates. In this paper, we present implementation details on reconstructing UnTangle Map, and the modifications we made on visualization and interaction, namely, multiple view, layer toggling, interactive zooming, interactive user editing, label tooltip, and data item highlighting. We propose a set of analytical task on multi-dimensional data and analyze how our modifications can improve the system's performance in those tasks. Our code* [1] *is written in JavaScript with D3.js*[2].

## 1. Introduction

Multi-dimensional data are ubiquitous in real-life data and in machine learning. Probabilistic multi-label data is a special type of multi-dimensional data where data items are represented as probability vectors, in which each dimension corresponds to the possibility of a label. It can also be generalized to any multi-dimensional data with non-negative values that can be meaningfully normalized.

Analysis of multi-label data can be done in various levels, including label-item correlation for single item or group of items, and label-label correlation for pairs, triples, or multiple labels. Depending on the use case, the analysis may help people select important features or evaluate model output.

Visualization and dimensionality reduction play an important role in multi-dimensional data analysis. UnTangle Map[1] is a novel multi-label data visualization. It present a label layout algorithm that generate a network of labels on triangular tilling, and plot data items as ternary plot. In this paper, we will be focusing on reconstructing this system, and to improve its performance with multiple view and enriched interaction.

[1] https://github.com/mahouoji/UnTangleMap
[2] https://d3js.org/

## 2. Related Work

In common approaches, multi-label data can be losslessly visualized using parallel coordinates[3] or scatter plot matrix[2], or be projected to 2D plane with Dimensionality Reduction techniques such as PCA and t-SNE. There are visualization tools for specific usage, for example EnsembleMatrix[6] for multiple classifiers and LDAvis[5] for topic modeling.

More related to our paper, UnTangle Map[1] addresses the problem of visualizing both label correlations and item-label relationship by generating label network on triangular grid based on correlation and plotting data items using ternary plot. A ternary plot visualize three variables that sum to a constant within a equilateral triangle using barycentric coordinates. UnTangle Map helps to demonstrate visual patterns of labels and items, for example, label clusters, path and gaps between clusters and item distribution across multiple labels. Compared to traditional visualizations, UnTangle Map is more scalable both on label and on data items.

## 3. Method

UnTangle Map [1] build the layout in two steps: first compute the label layout on triangular grids based on their correlation, and then plot data items within each triangle as a ternary plot.

### 3.1. Layout data labels

#### 3.1.1 Triangular grid

We build the triangle grid, or triangular tilling, upon a Discrete Triangular Coordinate System described in [4]. As shown in Fig.1, triangle vertices are addressed by zero-sum triplets. Triangles can be represented by the center, with a *positive* $\triangle$ *triangle* as $+1$ sum triplet and a *negative* $\triangledown$ *tri-*

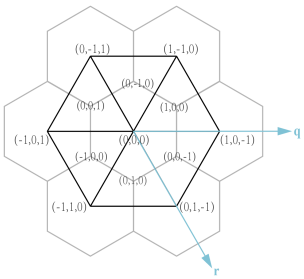



Figure 1: the discrete triangular coordinate system

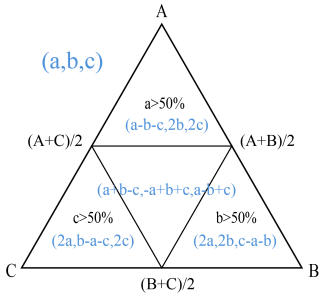



Figure 2: update ternary heatmap in a top-down approach

*angle* as $-1$ sum triplet. This coordinate system can be extended with the Barycentric Coordinate System to address the entire 2D plane and can be converted to and from the Cartesian Coordinate System in SVG.

#### 3.1.2 Layout algorithm

A good label layout where correlated labels are close to each other is achieved by maximizing the objective function $\mathcal{F}$:

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

The first term stands for the average pair-wise correlation of labels where $c_{ij}$ is the correlation coefficient of label $i$ and $j$. The correlation coefficient can be computed in multiple functions, such as Spearsman's rank correlation, Kendall's rank correlation coefficient, and Pearson correlation coefficient. The second term stands for the average correlation among all triangle faces with $c_t = \frac{c_{ij}+c_{jk}+c_{ik}}{3}$ for label $i$, $j$, $k$ that forms a triangle. Maximizing the second term encourages clustering of labels. Although finding the optimal is NP-hard, decent solutions can be find through a greedy approach aided by stochastic hill climbing and mixed-initiative interactions.

We implement the layout algorithm according to the pseudo code in [1]. Labels are added to or removed from the graph in an incremental way. They are placed on candidate slots where new boundary can be created. When a label is updated, its neighboring faces are added to or removed from the graph and the neighbor status of its six neighboring vertices is updated. A candidate slot is an empty vertex having two adjacent neighbors in the graph (except for the first and second label added to the graph). The triangular coordinate system allow neighbouring vertices and faces to be easily computed and indexed.

To make the result more controllable, we discarded the stochastic factor and compensate for that by offering dynamic suggestions on user editing as described in 3.4.

### 3.2. Plotting data items

#### 3.2.1 Ternary plot

Data items are plotted as ternary plot in the triangles formed by labels. In an triangle with labels $l_i$,$l_j$,$l_k$ on the vertices, for each data item having non-zero possibility vector $(p_i, p_j, p_k)$, its barycentric coordinate is $(w, u, v) := (\frac{p_i}{p_i+p_j+p_k}, \frac{p_j}{p_i+p_j+p_k}, \frac{p_k}{p_i+p_j+p_k})$. Assuming the position vector of the labels are $P_i$, $P_j$, $P_k$ in Cartesian coordinates, the position of that data item is $wP_i + uP_j + vP_k$.

To help users read the exact value of the barycentric coordinates, we also plot grids inside each triangles. Grids have multiple level of granularity: rough, medium, and fine, similar to the heatmap as described 3.2.2.

#### 3.2.2 Ternary heatmap

Rendering ternary plot of large amount of data items can be slow. To improve scalability, the UnTangle Map paper present ternary heatmap, in which triangles are spilited recursively to four sub-triangles, each maintains the number of data items that fall into its area and visualizes the value as heatmap. The counters can be updated in an top-down approach: first decide which sub-triangle the label falls into and update its counter, then recursively split and update the sub-triangle using new vertex position and the item's new barycentric coordinate under the sub-triangle, as shown in Fig.2. We split three times to get tree levels of granularity, each consists of 4, 16, and 64 triangles. In our modification, we also maintain the exact data items that falls in each most fine-grained triangles respectively. Triangles at any granularity can get all its data items by recursively referring to its sub-triangles. This will support the item highlighting interaction as described in 3.4.

### 3.3. Multiple view

In support of UnTangle Map, two other visualization are provided, namely, parallel coordinate plot (PCP) [3] and scatter-plot matrix (SPM) [2]. PCP represent labels as vertical axes that are parallel to each other, and each data item

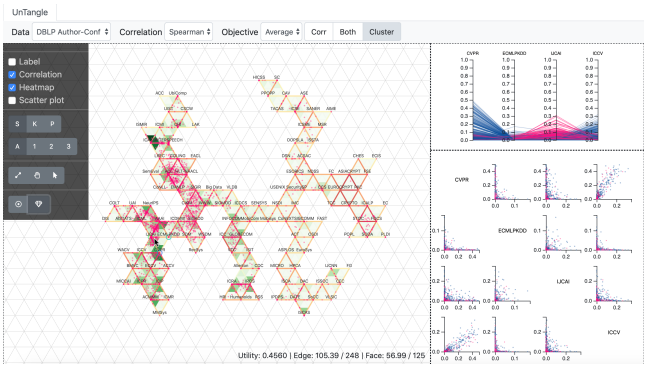



Figure 3: User interface, multiple view and item highlighting

as a ploy-line whose position on an axis represent its possibility of that label. SPM provide a scatter-plot of data items under each pair of label and arrange them in a matrix where rows are labels as y-axes and columns are labels as x-axes. We place the name of the label at the diagonal.

These two classic multi-dimensional visualization can help user to perform tasks that UnTangle Map is not very suitable for, for example, identifying linear correlation between a pair of label. Yet they are not as salable as UnTangle Map where hundreds of labels can be plotted on a relatively small canvas. We allow users to manually select labels of interest in UnTangle Map and plot that subset of lables in PCP and SPM. In this approach, UnTangle Map helps user to browse and filter labels from a wide panorama.

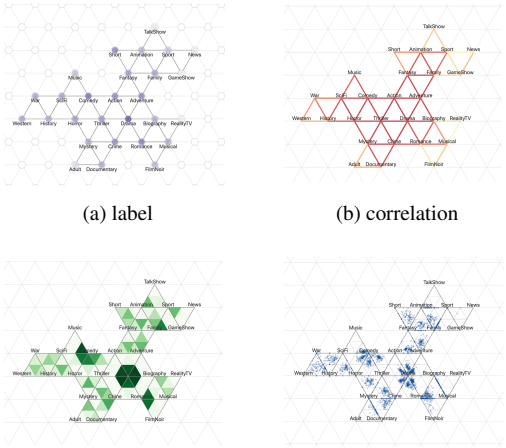
Fig.3 shows the overall interface of our system. User selected four labels to display in PCP and SPM, and use item highlighting tool as described in 3.4 to highlight items in the three views.

### 3.4. Interaction

UnTangle Map encodes information of labels and data items in various aspect, including label summary, label clusters, label-label correlation, and label-item correlation. It also maintains metadata such as number of face and edges, and evaluation such as average pairwise label correlation and average per-triangle label correlation of the graph. To help user capture these information, we provide a rich set of interactions.

**Layer toggling.** We summarize visual channels into four layers, namely, lable layer, correlation layer, heatamap layer, and scatter-plot layer. Users can show or hide these layers to observe any combination of information according to their needs.

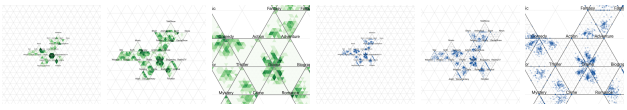
In label layer, each label is drawn as a circle whose color intensity encodes how globally dominant the label is. Higher intensity represent higher score for dominance, which is defined as the average possibility of one label among all data items. When label layer is hidden, only the



(a) label (b) correlation

(c) heatmap (d) scatter-plot

Figure 4: Four Layers



(a) (b) (c) (d) (e) (f)

Figure 5: zooming and granularities of heatamp and grids

label names are shown.

In correlation layer, each edge in the graph are colored according to the correlation coefficient of the pair of label on its two vertices. We use a diverging color scheme where red represent a positive value and blue represent a negative value. User can also select which correlation coefficient function to visualize from Spearman's, Kendall's, and Pearson.

In heapmap layer, data items are plotted as ternary heatmap. User can select from three level of granularity: rough, medium, and fine, each divide one ternary plot into 4, 16, and 64 small triangles. Rough heatmap is helpful for identifying locally dominant and isolated label and overall label popularity. Finer-grained heatmaps depict patterns of item distribution as can be observed in scatter-plot while being much faster to render.

In scatter-plot layer, data items are plotted one by one as semitransparent circles. This layer is lazy-updated only when it is shown, and the elements are removed from the SVG graph when it is hidden to accelerate rendering and interaction.

**Interactive zooming.** When performing different visual task, user may need to view the graph at different level of detail. The original system implemented simple panning and zooming. We allow user to pan on an infinite triangular grid and provide interactive zooming by creating UI that

automatically adjust to current zooming scale. For example, label names are resized so that users can always read them and that they do not interfere with users performing a typical analytical task under that zooming scale. User can also set heatmap and ternary grids to automatically switch to higher or lower granularity as they zoom in or out.

**User editing.** After a layout is generated, users can drag and drop every single label to tweak the layout. Suggestions are provided as a label is being grabbed. Valid slots that can create new boundaries are highlighted as colored circles with the color intensity encode the value of objective function if the label is placed in that slot. Hovering the label over a valid slot, in the lower right corner of the UnTangle Map division, the exact value of objective function, along with edge-wise label correlation, face-wise label correlation, and the number of faces and edges, will be updated assuming that the label is placed in that slot.

As demonstrated in [1], merely maximizing objective function on placing each label may not lead to an overall maximal utility for the whole map, stochastically choosing the second or third best option may lead to a better layout. Also, user may have other needs that are not addressed by the objective function. We allow user to put label on any empty slot, and have the graph updated accordingly.

**Label tooltip.** Users can hover over each label to view detailed information of that label.

**Item highlighting.** Notice that one data item may be plotted multiple times if it has non-zero value for labels that forms multiple triangles. Users view that item distribution by hovering over the heatmap to highlight all the data items that fall within that triangle slot. Through the distribution, user get an intuition on the correlation of labels and clusters that are not adjacent in map, and may gain insight on how to further optimize the layout. In this mode, the items are also highlighted in PCP and SPM view.

# 4. Evaluation

## 4.1. Dataset

We use two datasets, IMDB movie dataset[3] and DBLP dataset.[4]

IMDB dataset consists of movies and their genre, a movie may be marked with one to three genre. We extract two pairs of data item and label: move-genre and year-genre, which is one specific year with the distribution of genre for all the movies in that year. There are 14,762 movies from 116 years, and 28 genres.

For DBLP dataset, we take an author as a data item, a conference as a label. The possibility is the number of publication in a conference to one's total number of publications. We selected 125 conference with H-index over 30, with H-index data from another source[5]. We selected the top 800 active author based on their number of paper published among these conferences. Although both using DBLP, we didn't reproduce the result in the orignal paper, because as far as we know details on data processing in the original paper are not provided.

## 4.2. Tasks

We present the tasks in Table.1 based on the user studies conducted in [1] and [7]. The tasks can be grouped as *data-item-oriented*, *label-oriented*, and *label cluster-oriented*. We describe potential process of using our system to complete these analytical tasks, and estimate the system's strengths and weaknesses. If possible, we might conduct user study to confirm our hypotheses.

**Single data item.** Our system currently do not support selecting single data item, because they are very likely to be overlapping. If such function is provided in the future, user can follow the steps of observing group of items to complete the task.

**Group of data items.** Users can select data items that falls in a same triangle slot in the heatmap and highlight them in UnTangle Map, PCP, and SPM view. They can identify dominant label for these items either by finding highlighted clusters in UnTangle Map or by finding axis with highest value in PCP. They can evaluate the stability of dominant label through the distribution of items in the whole map or reading from PCP.

**Single label.** Using only the UnTangle Map view, users can toggle the label layer to identify global dominant and iolated labels by their color. They can toggle the heatmap layer (or scatter-plot layer) to identify local dominant labels where its surrounding triangle slots are darker in color.

**Pair-wise label correlation.** Toggling correlation layer, user can identify the correlation coefficient for edges in the graph. To view the exact relationship, such as linear correlation, they can select labels and use SPM view which shows the distribution of data item under pairs of labels.

**Conditional probability.** Ternary plot is specialized for identifying conditional probability among three labels. User can put three labels of interest in one triangle, zoom in, and toggle scatter-plot layer. Users can observe the distribution of data items and read the exact ternary coordinates with the help of grids.

**Clusters of labels.** By maximizing the objective function, UnTangle Map generates laybel layout that may capture the clustering of labels. For example, Fig.6 found such topics as robotics, computer vision, language processing, and human computer interaction. The dominance of label cluster could be identified from the color intensity of heatmap.

[3] https://www.kaggle.com/orgesleka/imdbmovies.

[4] https://dblp.org/xml/

[5] http://www.guide2research.com/

Table 1: Analytical tasks for multi-dimensional data

| Task | | Questions |
|---|---|---|
| data-item-oriented | single data item | Maximum and minimum possible label |
| | | Stability: Does item belongs to a label with dominant high possibility? |
| | group of data items | Stability: Are there more stable items than unstable ones? |
| label-oriented | single label | Dominant/Isolated label: Which label is the strongest or weakest in the probability vectors? |
| | pair of labels | Which two labels are more correlated? |
| | | Which two labels mostly reflect linear correlation? |
| | triple of labels | Conditional probability: Given 1 prior label, what is the probability of another 2 labels? |
| | | Conditional probability: Given 2 prior label, what is the probability of another label? |
| label-cluster-oriented | cluster of labels | Are there any sets of labels that are more correlated to each other? |
| | groups of label clusters | Given groups of label clusters, are they positively correlated? |

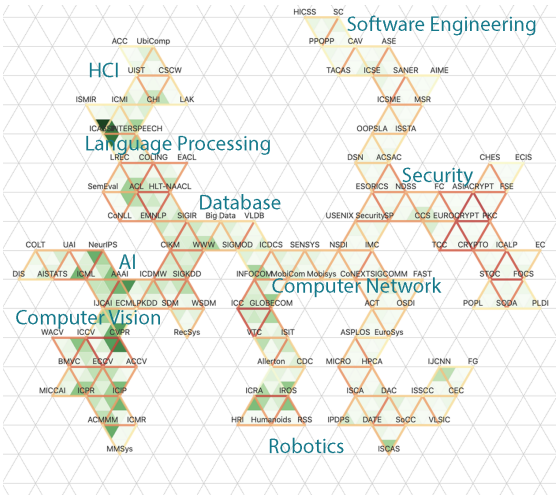



Figure 6: rough clusters of labels from generated layout

Notice that there is no systematic definition on how good the clusters are. On the one hand, optimal layout to maximize objective function is hard to find. On the other hand, maximizing the objective function does not guarantee the layout to be ideal because it only consider label correlation. Many outliers could be found in Fig.6. User may need to understand the labels to make effective editing on the generated map.

**Groups of labels clusters.** User can get an intuition of data item distribution in multiple data clusters through data item highlighting. Some clusters may tend to contain the same selection of data items.

## 5. Conclusion

In this paper, we reviewed the implementation details of UnTangle Map, enhanced its visual channel and interaction, and analyzed our system's potential in solving typical analytical tasks.

We argue that UnTangle Map is a novel and useful tool for probabilistic multi-label data analysis. It demonstrates multiple aspects information in one graph, and can be very scalable to label and data items. The original UnTangle also has draw-backs. The label layout can be hard to interpret. Some common analytical task may not be suitable for it. We make up for these draw-backs with interactive user editing, multiple-view, along with label-level and data-item-level interactions.

For future work, we may test our system with multi-label data from other sources, such as LDA and fuzzing clustering. We can further enrich user interactions and conduct real user study. Another direction is to make our code a reusable D3.js plug-in.

## References


[1] N. Cao, Y.-R. Lin, and D. Gotz. Untangle map: Visual analysis of probabilistic multi-label data. *IEEE transactions on visualization and computer graphics*, 22(2):1149–1163, 2015.

[2] W. S. Cleveland. *Visualizing data*. Hobart Press, 1993.

[3] A. Inselberg. Parallel coordinates: visualization, exploration and classification of high-dimensional data. In *Handbook of Data Visualization*, pages 643–680. Springer, 2008.

[4] B. Nagy and K. Abuhmaidan. A continuous coordinate system for the plane by triangular symmetry. *Symmetry*, 11(2):191, 2019.

[5] C. Sievert and K. Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[6] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1283–1292, 2009.

[7] Y. Zhao, F. Luo, M. Chen, Y. Wang, J. Xia, F. Zhou, Y. Wang, Y. Chen, and W. Chen. Evaluating multi-dimensional visualizations for understanding fuzzy clusters. *IEEE transactions on visualization and computer graphics*, 25(1):12–21, 2018.