Visual Analytics of Genealogy with Attribute-enhanced Topological Clustering

Ling Sun, Xiang Zhang, Xiaan Pan, Yuhua Liu, Wanghao Yu, Ting Xu,
Fang Liu, Weifeng Chen, Yigang Wang, Weihua Su and Zhiguang Zhou*

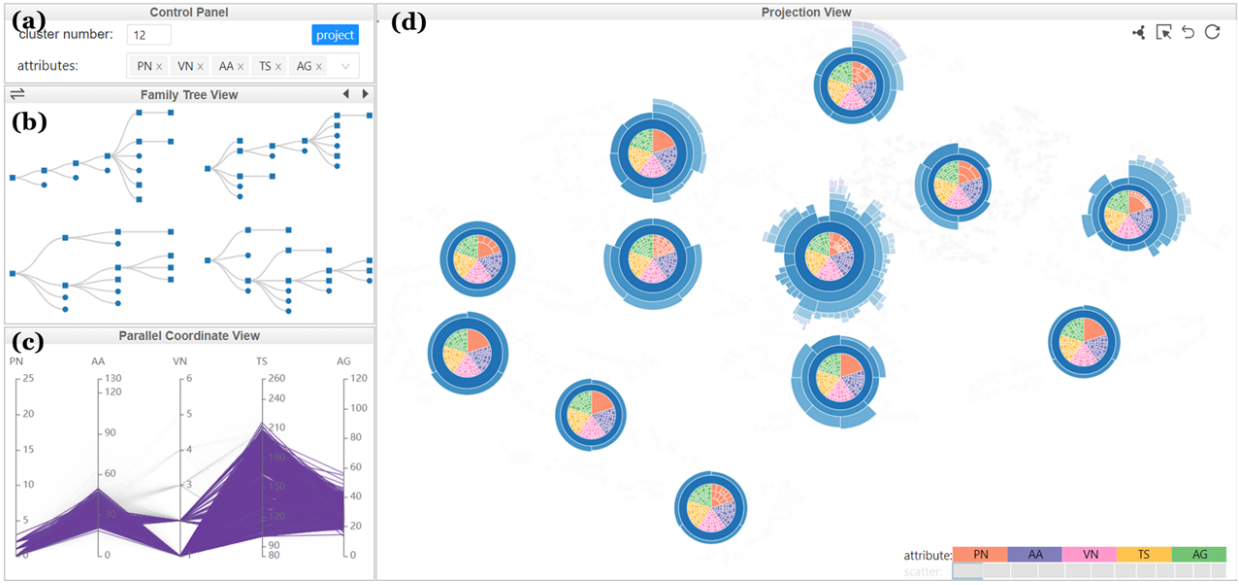

Fig. 1. A screenshot of our system showing an analyst exploring a real-world genealogy dataset, observing the structure and attribute features of clusters. (a) The control panel enables users to specify parameters of our clustering model. (b) The family tree view presents all family trees of a selected cluster in the form of a structure overview. (c) The parallel coordinate view is designed to present multi-dimensional attributes of the family trees. (d) The projection view presents the family clusters, in which the distributions of clusters are determined according to the similarities of the learned vectors and multiple attributes, and the glyphs of clusters are designed encoding multiple attribute features.

Abstract—Clustering is able to present a brief illustration for families of interest and patterns of significance within large-scale genealogical datasets. In the traditional clustering methods, topological features are mostly taken for summarizing and organizing family trees. However, plentiful attributes are ignored which are also important to enhance the understanding and interpretation of genealogical clustering features. Thus, it is a crucial task to combine structures and attributes into a clustering model for exploring genealogy datasets. In this paper, we propose an attribute-enhanced topological clustering method for exploring genealogy datasets based on Partial Least Squares (PLS). Firstly, a graphlet kernel method is utilized to measure the structure difference between family trees. Then, we leverage PLS to combine the learned vectors and multiple attributes, and a joint dimensionality reduction method is applied to project the high-dimensional vectors into a two-dimensional space in which a distance-based clustering method is employed to aggregate the similar family trees taking both the topological structures and attribute features into consideration. Further, we implement a visual analysis system with multi-view collaboration, including glyph, family tree view and parallel coordinate view, to represent, evaluate and explore the clustering features. Case studies and quantitative comparisons based on real-world genealogy datasets have demonstrated the effectiveness of our method in genealogical clustering and exploration.

Index Terms—Visualization in the Humanities; Data Aggregation; Data Clustering; Compression Techniques; Dimensionality Reduction; Hierarchical Data

1 Introduction

- *Ling Sun, Xiaan Pan and Wanghao Yu are with Zhejiang University of Finance and Economics. E-mail: {1055389464, 1967914901, 1968480927}@qq.com.*
- *Xiang Zhang and Fang Liu are with Zhejiang University of Finance and Economics. E-mail: {zkebi, maggie_liufang}@126.com.*
- *Yuhua Liu and Zhiguang Zhou are with Zhejiang University of Finance and Economics. E-mail:{liuyh216, zhgzhou1983}@163.com.*
- *Ting Xu is with Zhejiang University. E-mail: xut@zju.edu.cn.*
- *Weifeng Chen is with Zhejiang University of Finance and Economics. E-mail: cwf818@zufe.edu.cn.*
- *Yigang Wang is with Hangzhou Dianzi University. E-mail: yigang.wang@hdu.edu.cn.*
- *Weihua Su is with Zhejiang Gongshang University. E-mail: swh@zjgsu.edu.cn.*
- **Zhiguang Zhou is the Corresponding Author.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

With the increasingly available genealogy datasets, great opportunities are provided for demography, evolutionary biology, human migration [9], origin of family names, family genetics, etc. Thus, the study of genealogies has attracted more and more attentions ranging from social scientists to historians. However, due to the large data volume, tree structural heterogeneity and associated attribute information, experts often fail to quickly discover families of interest and patterns of significance within the large-scale genealogy datasets. According to the principle of "Overview first, zoom and filter, then details on demand" [21], it is quite important to create an intuitive, simple, yet comprehensive overview of the large collection of family trees. Thus, the experts will easily obtain a visual summary of the entire genealogy dataset, and focus on the families with specific patterns to further get insights into the features of family development and evolution.

Clustering is a commonly-used method to simplify data cognition and expression, which groups similar families into the same clusters, enabling experts to gain general patterns of the entire genealogy dataset at the aggregated level [20]. Previous works mostly focus on summarizing and organizing family trees based on topology-based clustering. However, the currently open datasets of genealogy are often informative. In addition to the kinships between individuals, they embrace plentiful personal information, including the gender, residence, birth/death time, and marital status, etc. Thus, it will be more conducive to cluster the family trees taking both the topological structures and attribute features into consideration, and revealing the correlation between family trees and social attributes [23, 52]. For example, what is the relationship between the average lifespan of family members and the number of generations? How the migration behaviors affect the family population? Can the ages at the first birth of male founders lead to the imbalanced development of family branches? However, there are still three challenges for aggregating the family trees into clusters with the structural information and attribute features taken into consideration:

C1. The structures of trees are always described with topological relationships, which are quite different from that of attribute features, making it a difficult task to combine them effectively.

C2. Clustering features involve abundant information [53], which makes it difficult enough to visually understand the implicit knowledge and capture patterns of interest. It is important to design intuitive visualizations to visually present structural features [39] and attribute information for clusters.

C3. The correlation between the structures and attributes is difficult to discover for users [51, 54], so that it is an urgent task to develop a visual analytics system, enabling users to conduct quantitative comparison, interactive analysis and in-depth exploration of the genealogy clustering features.

To tackle the above challenges, we propose a structure-attribute fusion model for genealogy clustering in this paper. First, we leverage a graphlet kernel method to measure the structural difference between the family trees. Then, Partial Least Squares (PLS) is utilized to combine learned vectors and multiple attributes, and a joint dimensionality reduction is conducted to project family trees into a low-dimensional feature space, where family trees sharing similar structures and attributes located close to each other are further aggregated based on a distance-based clustering method. Furthermore, we provide various kinds of quality metrics to evaluate the clustering features, and design a set of multi-scale glyphs to visually present their structures and attribute features. In addition, multiple coordinated views, a series of visual cues and interactions are designed enabling users to select those clusters of interest and gain deeper insights. We further demonstrate the effectiveness and usefulness of our system through case studies and expert interviews based on a real-world dataset.

The major contributions of this paper are summarized as follow:

- A family tree clustering method is proposed to aggregate those family trees based on the combination of structures and attribute features.
- A rich set of visualizations and interactions are provided for visual presentation and exploration of the family tree clusters, enabling users to gain deeper insights of those clusters of interest and original large-scale genealogical datasets.
- Case studies and quantitative comparisons based on a real-world genealogical datasets are conducted to evaluate the clustering results and verify the effectiveness and practicability of our system.

This paper is structured as follows. Section 2 briefly introduces the related work. Section 3 summarizes the analytics tasks and presents the system overview of our work. Section 4 lays out details of our attribute-enhance topological clustering. Case studies and quantitative comparisons based on a real-world dataset and discussions are presented in Section5. Finally, Section 6 concludes this paper.

2 Related work

We classify the related work into four categories, including genealogy data analysis, topological data clustering, attribute-enhanced topological data analysis and visual cluster analysis techniques.

2.1 Genealogy Data Analysis

Genealogy has always been a popular activity among people. Kemp [13] found that the audience of genealogy ranged from young to old, as well as ethnic groups. While genealogy is popular with the public, it provides substantial contributions to many studies in the humanities and social sciences. For example, Tsuya et al. [34] studied the European and Asian genealogy to estimate trends in long-term population ratios. SW et al. [8] discussed patriarchal society in Chinese history from dimensions of depth, the number of male members and inclination. Liu et al. [22] analyzed the development of families related to structure, population, migration, and other demographic information.

However, there are many problems in large-scale genealogy data analysis [47] , such as numerous, huge branches and complex hierarchies. Visualization scholars have carried out numerous studies on simplified expression of genealogy. The visualization of family trees can be roughly divided into three categories. **(1) Node-link based visualization.** Nodes represent family members, edges represent parent-child relationships [33], and the layout mode can be the orthogonal layout, indented layout, or radial layout. **(2) Line-based visualization [40].** The individual is represented as a horizontal line, and the length of each line represents the life span of the corresponding individual. Moreover, in order to represent the relationship between individuals (lines), vertical lines are drawn to map the parent-child relationship, and the convergence and divergence between the lines to map the marriage and divorce relationship. **(3) Matrix-based visualization.** The layout can be a diagonally filled matrix, where rows are individuals and columns are core families [1]; it can also be an adjacency matrix, where the non-zero term represents the edge between two corresponding vertices in the graph. There are also some niche representations, such as fan charts, hourglass charts, dual trees, etc.

2.2 Topological Data Clustering

The topological data clustering is one of the main research areas in clustering. A breadth of techniques have been proposed for topology-based data clustering [44], which can be classified under two main categories: graph clustering and tree clustering.

Graph clustering. To date, many graph clustering techniques have been proposed, which are based on various criteria including clustering based on normalized cut, modularity or structure density. There are other clustering methods that rely on similarity functions defined over the graphs. For example, the Editing Distance [12] calculated the distance between undirected acyclic graphs as the sum of costs when efficiently transforming one graph into other. The graph histogram technique [29] captured data features in form of a histogram and then calculated distance between graphs using histogram distance functions. Dexter et al. [7] compared the differences between protein structures based on a RMSD matrix, and grouped them into clusters by the Lance-Williams update algorithm. Graph clustering is an exploratory data analysis task and has been widely applied in many areas. For example, Kong et al. [15] proposed a botnet detection method to analyze the data

packets, based on a graph structure clustering algorithm, MST. Kutz et al. [17] presented the distribution of patents across classes, which improved understanding of the evolvement of patent portfolios over time.

Tree clustering. Like graph clustering, most tree clustering techniques are also based on tree comparison. For example, SW et al. [8] filtered and collected all family trees based on the corresponding criteria, such as inclination. DAViewer [48] allowed for structural comparison of the trees by forcing the elementary discourse units to be the same across parsing algorithms. TreeJuxtaposer [27] compared trees by associating each node in one tree to the best corresponding node in the other tree for categorizing trees. Kosaka et al. [16] proposed a tree-structure speaker clustering algorithm according to a maximum likelihood criterion. Hillies et al. [10] visualized the relationships among sets of phylogenetic trees by comparing the tree-to-tree distance using multidimensional scaling.

2.3 Attribute-enhanced Topological Data Analysis

Associated attribute information in multivariate network datasets can help users to analyze the relationship establishment, community formation and network evolution. For example, Wang et al. [37] explored several factors on the academic influence of scholars, such as the number of papers, citations and cooperation relationships. Ko et al. [14] explored the types and quantities of flight delays between airports by visualizing airline network data with multidimensional spatiotemporal information. Nober et al. [28] visualized the multivariate clinical data in a family to explore how the hereditary and environmental factors affect individual health. Wattenberg [38] designed the PivotGraph, a software tool focusing on the relationship between node attributes and connections of multivariate graphs on a grid layout. Pathfinder [30] used path queries on networks and presented the resulting paths in a linear, ranked list, juxtaposed with rich attribute data for judging paths. Pathline [26] looked at multiple sets of values simultaneously, across time, species, genes and metabolites, in order to compare trends between species. Jin et al. [11] proposed a text clustering algorithm based on the text semantic representation and the graph structure of word co-occurrence to improve the clustering effort. Yang et al. [45] combined the original network structure and individual interest attributes to mine the implicit cluster structure network and then predicted the propagation process of the hot event.

2.4 Visual Cluster Analysis Techniques

In order to assess the quality of a clustering, many quality metrics have been proposed, such as the Calinski-Harabaz index [3], Silhouette Coefficient [32], and Davies-Bouldin index [6]. In addition, visual clustering analysis technology is also widely used in high-quality clustering discovery. For example, VISTA [5] enables users to evaluate the clustering results on a 2D projection visually. However, this method does not support the comparison of multiple clustering results. Clustervision [18] supports finding high-quality clustering results by ranking clustering results utilizing five metrics. XCluSim [24] enables users to interactively generate and compare multiple clustering results with multiple coordinated views. DICON [4], an icon-based cluster visualization method, embeds statistical information into a multi-attribute display to interpret and evaluate clusters. The visualization method of iterative clustering combines automation with interactive methods, allowing users to define seeds(centers) and help users interact with the process [2]. The Clusterix system uses minimization functions to automate parameter selection, allowing users to define clusters and modify features for the clusters to be performed [25] .

3 Requirement Analysis And System Overview

3.1 Data Description

The real world dataset that we use to ground our study is CMGPD-LN. The China Multigenerational Panel Dataset-Liaoning (CMGPD-LN) is transcribed from the population registers compiled by the Qing Dynasty government in Liaoning Province, northeastern China, from 1749 to 1909. This dataset, with more than 1.5 million records, provides socioeconomic, demographic, and other characteristics for over 260,000 individuals. According to the relationship between individuals, we build approximately 12, 000 family trees. In addition, five attributes are extracted and constructed from member information to characterize the features of families, which include the timespan (TS), average age (AA), position number (PN), villages number (VN) and average gap (AG) between father and son.

3.2 Requirement Analysis

The design of our genealogy clustering system is grounded by interviewing two experts (E_1 and E_2). E_1 is a professor in a research institute, who has extensive experience in graph visualization and mining. E_2 is a university professor in the field of humanities and social sciences, who is knowledgeable on demography, history and human migration. To inform the design and development of our system, we conduct a preliminary design study with domain experts. Specifically, we began by interviewing with experts about their previous work and asked for major issues that were prevalent in their process. It turned out that clustering is an important means to explore and analyze a genealogical dataset, and it is a significant task to take both the topological structures and attribute features into consideration, since these key elements in genealogy are conductive to discover the correlation between family trees and social attributes. Furthermore, a list of requirements was derived from the experts' comments about our system and used to guide the development of our genealogical clustering system. Based on these rounds of interviews with experts, the requirement tasks are summarized as the following.

R1: Fusion of attributes and structure for genealogy clustering. Topology-based clustering methods applied in genealogy require comparing the structure difference between family trees in advance [49]. Family tree is a kind of graph with a special topological structure, so that the traditional graph similarity measurement methods can be used for genealogical data. In addition, the family trees also have rich attribute information and attribute-enhanced clustering is able to give cluster features more practical significance. However, the topological relationships and attribute features are of different magnitudes, so that it is difficult enough to combine them evenly. Therefore, experts require an efficient joint clustering method for synthetically considering both features.

R2: Intuitive visual representation of genealogy clustering results. Clusters contain information on multiple dimensions [19], including structure and attribute. An effective visual representation must convey each dimension of data and ensure that their data is faithfully represented, so that users can visually understand the implicit knowledge. Moreover, in order to capture the pattern of interest and get more insight in the fields of demography, history, etc. from the genealogy dataset, the visual representation of clusters not only needs to support users to view the details of global information between each cluster, but also local information of each family tree in a cluster, and verify their performance in terms of both structures and attributes.

R3: Evaluation of clustering results. In the specific scenario of exploration and analysis of genealogy datasets, the effectiveness of clustering needs to be verified. In the case of different feature combinations, the difference between families within the cluster and families outside the cluster, and between different clusters are expected to be quite distinctive in the clustering results [50]. Therefore, quality metrics and visual cues are required to demonstrate the quality of clustering features, enabling users to evaluate effectiveness from different perspectives and intuitively interpret the clustering features.

R4: The clustering analysis system of genealogical data. The traditional genealogy analysis methods and commercial software used in the field of social science do not support the analysis of genealogy datasets with specific features alone. Besides, the analysis results are limited to numerical statistics, which is not conducive to intuitive understanding and rapid exploration and analysis of large-scale genealogy datasets. Therefore, a visual analysis system of a genealogy dataset is required to support non-expert users to capture family trees of interest and gain deeper insights within large-scale genealogy datasets in a simple exploratory interactive way.

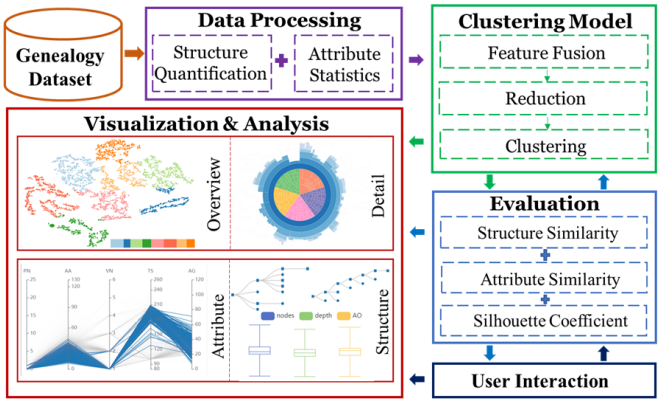

Fig. 2. The pipeline of our genealogy clustering system based on attribute-structure synchronization.

3.3 System Overview

Motivated by the identified requirements, we design a visualization framework enabling users to explore and analyze genealogy datasets based on an attribute-enhanced topological clustering method. The system pipeline is presented in Figure 2. Firstly, a large-scale genealogy dataset is loaded into the visualization system, and features of structure and attribute are captured by graphlet kernel and statistical method respectively in advance. Then, a novel genealogical clustering model is conducted to combine learned vectors and multiple attributes based on PLS, for achieving distance-based clustering after joint dimensionality reduction (**R1**). Furthermore, a set of multi-scale glyphs are designed to visually present the features of clusters from multi-dimensional perspectives (**R2**), and a variety of quality metrics are further calculated to evaluate the effectiveness of the clustering features (**R3**). Multiple coordinated views and a series of interactions are integrated into the visualization system for visual clustering of genealogy, enabling users to explore and analyze the large-scale genealogy dataset (**R4**).

4 Algorithm

In this section, we detail the course about our family tree clustering based on the attribute-structure synchronization.

4.1 Topology-based Genealogical Clustering

Our approach of topology-based genealogical clustering is to compute structure difference between pairwise families by graphlet kernel in advance and convert the family tree data into vector data. However, the vector data have high dimensions [36], so we utilize t-SNE to reduce the dimension of data. The topology-based clustering algorithm mainly consists of two steps as follows.

4.1.1 Structure Similarity Based on Graphlet Kernel

Actually, the graphlet kernel has rapidly developed recently and become an important branch of graph data analysis. The kernel value of two graphs is related to the similarity of the decomposed substructures and can describe the proximately between two graphs. The details of graphlet kernel algorithm and topology-based clustering are as follows.

Given two family trees $T_1(V_1,E_1)$ and $T_2(V_2,E_2)$, a graph decomposition method F, which is the method to sample graphlets based on Metropolis-Hasting random walk [31], the substructures obtained after decomposition are $F(T_1)=\{S_{1,1},S_{1,2},\ldots,S_{1,N_1}\}$; $F(T_2)=\{S_{2,1},S_{2,2},\ldots,S_{2,N_2}\}$. Now we use the graphlet kernel to map graph data into vector data in a high dimension space ϕ , and define the graphlet kernel as follows:

$$k(T_1,T_2)=\langle\varphi(T_1),\varphi(T_2)\rangle \tag{1}$$

where $\varphi(T)$ is a graphlet frequency vector defined such that the component $\varphi_i(T)$ of family tree T corresponds to the relative frequency of graphlet S_i. In essence, the graphlet frequency vector of the tree is the feature vector of the tree. the kernel $k(T_1,T_2)$ can be regarded as the proximity between two family trees. If there are N family trees, the result obtained from the graphlet kernel can be denoted by $K=(k(T_i,T_j))_{N\times N}$ and called a kernel matrix.

4.1.2 Dimension Reduction Based on t-SNE

With graphlet kernel, the similarity of N family trees is represented as a kernel matrix $K=(k(T_i,T_j))_{N\times N}$ in a high-dimensional space. High dimension data often have redundant features, or irrelevant features, and all of these will increase the computational complexity and affect the efficiency of the clustering algorithm, or even lead to dimension disaster, so dimension reduction is necessary. t-Distributed Stochastic Neighbor Embedding (t-SNE) [35] is an effective dimensionality reduction method, which is capable of enhancing local features while retaining global features in the dimensionality reduction space [42, 46]. After dimension reduction, we get the two-dimensional vector data which keeps the main structural characteristics of the genealogy data simultaneously [43]. We then conduct clustering on the dimensionally reduced family trees by K-means, the distance-based clustering algorithm, in the projection space.

4.2 Attribute-enhanced Genealogical Clustering

We will introduce how to extract the attribute features from family trees, and enhance them based on the topological clustering.

4.2.1 Attribute Feature Extraction

The attributes of family trees are extracted based on the domain knowledge of genealogy, including micro level and macro level. The macro-level attributes refer to global features of family trees, such as the timespan, which reflect the family-oriented information. In contrast, the micro-level attributes can be obtained by the sum or average of the values associated with nodes, such as the village number, the average age, etc. which reflect the individual-oriented information. These macro-level and micro-level attributes constitute an attribute vector of each family tree, $A_i=\{a_{i1},a_{i2},\ldots,a_{iN_A}\}$, where N_A refers to the number of attributes which are extracted from the family tree T_i.

4.2.2 PLS-based Clustering

Based on the above extraction of structure and attribute features, supposed we have obtained two groups of feature sets in a large-scale genealogy dataset. Given a family tree T_i , two corresponding feature vectors are $s_i\subset\mathbb{R}^q$ and $a_i\subset\mathbb{R}^p$, where p and q are the dimensions of those vectors respectively. The structure and attribute feature vectors of all family trees constitute two data matrices $S\in\mathbb{R}^{p+n}$ and $A\in\mathbb{R}^{q+n}$, where n is the total number of family trees. We decentralized variables so that the columns of A^T and S^T are zero-mean. Then we define the between-set covariance matrix of A and S as $(1/n-1)$ Sas, where Sas $=AS^T$.

PLS aims to find a pair of weight vectors, input weight α and output weight β, respectively for A and S, to maximize the covariance between input and output variables, A and S.

$$\{\alpha;\beta\}=arg\max_{\alpha^T\alpha=\beta^T\beta=1}Cov(A^T\alpha\ ,\ S^T\beta)=arg\max_{\alpha^T\alpha=\beta^T\beta=1}\alpha^T Sas\beta \tag{2}$$

This method decomposes high collinearity variables into one-dimensional variables, A-score T and S-scores U, which is formulated as follows:

$$A=TP^T+E\ ,\ S=UQ^T+F \tag{3}$$

Where P and Q are loading matrices, and E and F are residuals. PLS algorithm is an iterative process, which can be formulated as follows:

$$E_{h-1}^T F_{h-1}F_{h-1}^T E_{h-1}\alpha_h=\lambda_h^2\alpha_h \tag{4}$$

We initialize $E_0=A^T$, $F_0=S^T$ and $h=1$. After the above iteration and convergence ultimately, the A-scores T can be given as $T_h=E_{h-1}\alpha_h$. U_h and β_h can be derived in the same way. All

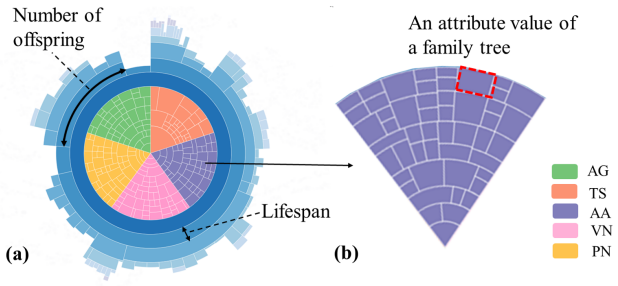

Fig. 3. Glyph design for a cluster. The outer sector (a) represents the structure of the representative family tree in a cluster and the inner circle (b) encodes five attribute values of families in a cluster. The fan-shaped sectors of yellow, purple, pink, orange, and green are mapped to attributes "PN", "AA", "VN", "TS", and "AG" respectively.

m pairs of weight vectors, $\{\alpha_i;\beta_i\}_1^m$, constitute two projection matrices $W_A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ and $W_s = \{\beta_1, \beta_2, \ldots, \beta_m\}$ respectively. Through the above PLS algorithm, we can acquire two sets of reduced vectors respectively. And each pair of PLS components can be get by

$$A^{'} = W_A^T A \ , \ S^{'} = W_S^T S. \tag{5}$$

Finally, for each family tree in the large genealogy dataset, the original attribute features and structural features can be transformed by PLS mentioned above to a pair of new vectors which maximized the covariance between two features. Then we directly concatenated those as an attribute-structure fused vector. Based on the above feature fusion using PLS, all family trees in a large genealogy dataset are represented as high-dimensional fused vectors by combining structure and attributes. Same with topology-based clustering mentioned above, then we use the t-SNE to project them into a two-dimensional space. We conduct distance-based clustering on the two-dimensional vectors by K-means.

5 Visual Design

5.1 Cluster Visualization

When a set of similar family trees are grouped together into a cluster, the attribute and structure features of family trees in a cluster must be combined into a single glyph representation, as shown in Figure 3. Considering in part to the aesthetic appeal of fan charts as well as their compact appearance relative to the more common node-link graphs, we provide another version of family trees in the outer sector of the glyph, as shown in Figure 3(a). The incremental layout method used in the radial space-filling tree follows the general format of the traditional Sunburst visualization. For each arc, the shade of filled color maps generation, the length encodes the number of descendants, and the bandwidth maps the lifespan.

The goal of counter-based treemap view is to present attributes of family trees in a cluster glyph. We extend an iterative slice-scale process on the radial space [41] to fit a contour-based treemap, as shown in the inner circle of glyph. The inner circle is equally divided into five fans with different colors, and each sector represents one kind of attribute distribution of all families in a cluster, as shown in Figure 3(b). Specifically, the fan is cut into many facets, whose number equals to the mapped family number which can be can be set to 50, 100, or the number of all families in a cluster, according to the attribute values of mapped families. In addition, the area of facets also encodes the attribute value. By observing the shape of the counter, users can get an intuitive idea about an overview of all the family attributes in a cluster.

5.2 Genealogical Clustering Evaluation

To evaluate the validity of the clustering results, we provide a couple of multi-scale metrics from the perspectives of structure and attribute.

Structure-related views are the family tree view and the boxplot view. When the user selects a cluster, the family tree view will display the

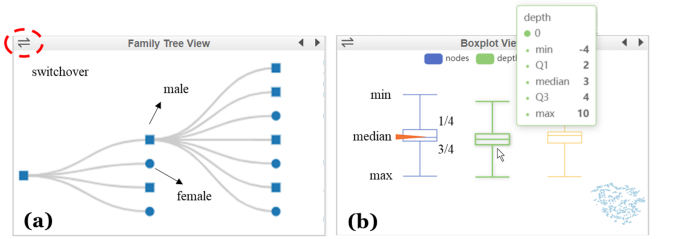

Fig. 4. The structure-related view. The hierarchical structures of all families in the selected cluster are displayed in the family tree view (a), and three statistical indicators of the structure are shown in the form of boxplots (b).

structures of family trees in detail in the form of a node-link graph, as shown in Figure 4(a). For each node, its shape encodes the gender. In addition, the structural statistical indexes of a cluster are displayed in the form of a boxplot. Boxplot can be used to help summarize three structural statistical indicators of all family trees in a cluster, including the average number of descendants (AO), the number of family members (nodes) and generations of a family (depth). In the box plot, as shown in Figure 4(b), the box bounds the first and third quartiles of the genealogy data. The horizontal line inside the box is the middle or median value of the structure feature value of the family trees in a cluster. The dispersion of the genealogy data above and below this range is marked by vertical tails that extend to the most extreme values within a border at 1.5 times the interquartile range.

The attribute-related view is the parallel coordinate view. The parallel coordinate view is adopted in our system for multi-dimensional attribute visualization. We have identified several important attributes of genealogy, including timespan, village number, position number and some other statistical features. Parallel coordinate view has vertical axes that represent each feature of family trees and draw a line crossing the axes for each family tree. The colors of the lines and clusters are unified to better distinguish the lines corresponding to different clusters, as shown in Figure 1(c).

5.3 Interactions

We developed a set of interactions to help users switch between multiple coordinated views. First, users can change the number of clusters in the projection views (Fig.1d) by adjusting a NumericUpDown in the control panel (Fig.1a). In addition, the projected elements can be freely specified by ticking the corresponding attribute check box in the control panel (Fig.1a). After setting these two parameters, users are ready to explore and analyze. Specifically, users can select a cluster by clicking the colored stripe of the clustering summary glyph in the lower right corner of the projection view, as shown in Figure 5, where each colored patch represents a cluster and whose width is proportional to the number of families in each cluster. When selecting a cluster, not only will the structure-related view (Fig.1b) be displayed the hierarchy and statistics of the genealogical tree, but the parallel coordinate view (Fig.1c) will also automatically highlight lines of selected families. Besides, to observe the overall features of the selected cluster, users can click the button "glyph" in the upper right corner of the projection view for investigating the glyph above mentioned.

6 Evaluation

In this section, the real-world genealogy dataset, CMGPD-LN, is experimented to evaluate the effectiveness of our clustering method. Then, case studies are conducted based on two real-word datasets to verify the validity and convenience of our visual clustering visual system.

6.1 Quantitative Comparison

We have conducted a set of experiments to verify the validity of our clustering strategies. In order to mitigate the impact of randomness, each clustering operation is performed 20 times and takes the average.

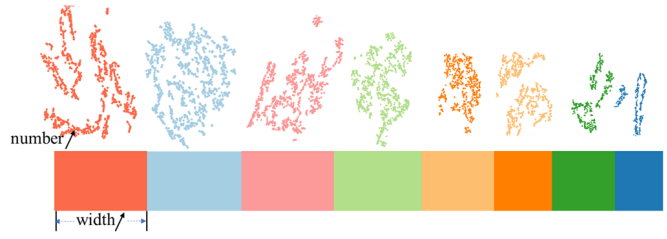

Fig. 5. The clustering summary glyph. Each cluster is summarized as a stripe with the same color. The stripes are sorted in descending order according to the number of families in the clusters and widths are proportional to the number.

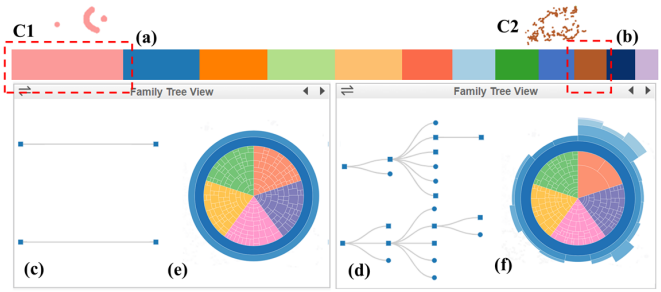

Fig. 6. Evaluation and comparison of different clusters with different numbers of families in the context of topology-based clustering. (a) presents a small cluster with a large number of simple families. (b) presents a large cluster with fewer complex families.

we compare our method with two clustering strategies: **(1) Structure (Str)** quantifies the structure of the family tree through graphlet kernel without consideration of attributes. **(2) Attribute (Attr)** clusters genealogy based on multi-dimensional attributes, without structures.

To further compare the effectiveness of different clustering strategies, we respectively evaluate the clustering quality by the following metrics: **(1) Structure Similarity (SS).** SS of a cluster is defined as the average structural similarity between any two points in the cluster, which is measured by graphlet kernel and ranges from 0 to 1. Well-defined clustering has a higher SS value. **(2) Attribute Similarity (AS).** AS is defined as the average of standard deviations of the attribute values of all families in each cluster. A low value of AS indicates that the family tree is well matched to its own cluster in terms of attributes. **(3) Silhouette Coefficient (SC)**. The SC of a clustering [32] is defined as the ratio of the difference between within-cluster distance and nearby-cluster distance and the maximum of these two distances, where distance is defined by the Euclidean distance of two-dimensional vectors. A high value of SC indicates that the family tree is well matched to its own cluster and poorly matched to neighboring clusters. Table 1 summarizes the experimental results across different clustering strategies under various numbers of clusters (C).

In the comparisons of the similarity in clusters, it can be found that our method performs better than Attr and performs inferior to Str in terms of structural similarity. With the increase of C, the gaps between our method and Attr are gradually widening, and the gaps with Str are gradually narrowing. Besides, in terms of attribute similarity, we found that our method performs better than Str and inferior to Attr. This indicates that our method combines the structure and attributes information of family trees towards striking the balance between structure similarity and attribute similarity. In addition, our method almost outperforms Str and Attr in terms of Silhouette Coefficient, which demonstrates that our clustering method not only improves the family similarity within clusters but also reduces the similarity between clusters under the context of the combination of structural and attribute features. The above results demonstrate that our clustering method effectively maintains the similarity of structures and attributes.

Table 1. Quantitative comparison

ClusterNum	Strategy	SS	AS	SC
C=5	Str	**0.7539**	11.5954	0.9825
	Attr	0.5907	**4.9957**	0.9836
	Our	0.6274	9.6254	**0.9859**
C=10	Str	**0.8235**	11.1262	0.9768
	Attr	0.6267	**3.9518**	0.9769
	Our	0.6943	7.9920	**0.9782**
C=15	Str	**0.8475**	10.8877	0.9685
	Attr	0.6368	**3.4887**	0.9707
	Our	0.7240	6.9365	**0.9714**

6.2 Case Study

We implement a clustering visualization framework based on a large-scale genealogy dataset that integrates the above clustering and the visual designs, and also provides a series of interactions to enable users to explore genealogy. To study the effectiveness of our system, we invited experts in the field of network mining and genealogy (E_1 and E_2) to conduct several seminars. First, we introduced the visual designs and interactions of our systems to the experts, and then ask them to use our system in free to explore the real-world genealogy dataset. The involved interactions and evaluation of our clustering method, and several cases found by our experts are formulated into case studies.

6.2.1 Evaluation of Clustering Results

After loading the data into the system, E_1 firstly set the number of clusters to "12" and pitched on "project" in the control panel. The projection view immediately presents the colorful clusters of all families in the two-dimensional space, where experts predicted that those families located nearby share similar structures. Then E_1 clicked on the icon representing glyph, and the projection view displayed all glyphs according to the location of the center point of the cluster. He became interested in the glyph on the pink cluster, which has only two arcs, as shown in Figure 6(e), and then fully examined the families of the cluster in the projection view. The expert found that the most of these families are two-generation structures with one father and one child, as shown in Figure 6(c). "The structure of the families in a cluster is similar to each other and to the outer sector of the glyph, so we can quickly understand the features of all family trees through all glyphs at first." said the expert.

Furthermore, in order to evaluate the quality of clusters based on attribute-enhance topological clustering, experts ticked five attribute checkboxes and the "Project" button successively in the control panel, and then observed the distribution of families under all feature combinations in the projection view, as shown in Figure 7(a). E_1 initially clicked a cluster with orange color and found that the structures of mostly families were very similar in the family tree view, and the distribution of attribute values on each vertical axis of the parallel coordinate view is also relatively tight, as shown in Figure 7(b). E_1 praised that "The multi-dimensional features are well preserved in a cluster." Subsequently, E_1 randomly clicked different colored stripes of the clustering summary glyph and found a strange phenomenon. On the one hand, there is no significant difference between the distribution of orange and dark blue clusters in the five kinds of attributes which can be seen from the parallel coordinate view and counter-based treemap. However, on the other hand, the distance between corresponding clusters in the two-dimensional space is very great, as shown in Figure 7(a). Out of this confusion, E_1 carefully observed the fan chart and the family tree view, and found that the structures of families in these two clusters were greatly different. Specifically, the families in the orange cluster presented the structural features of more generations and fewer descendants, while the families in the dark blue cluster presented fewer

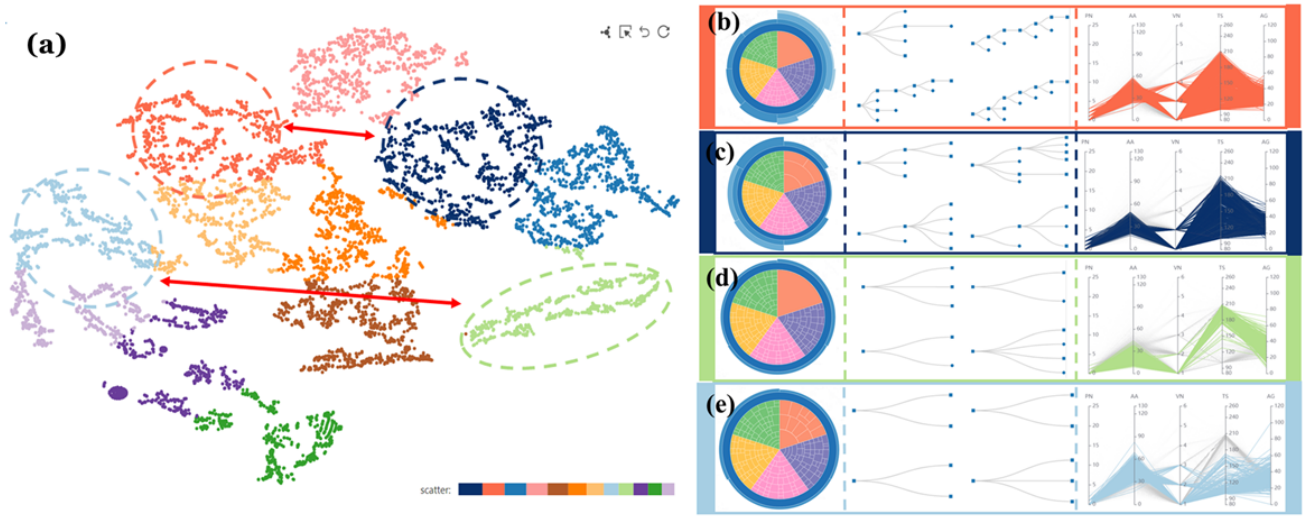

Fig. 7. Exploration of clusters on the projection view (a) with structures and attributes considered together. (b), (c), (d) and (e) present the detailed features of the corresponding clusters.

generations and more descendants, as shown in Figure 7 (b, c). In the process of further exploration, the expert found that the situation is completely opposite to the above, that is, the families in the green and light blue clusters which are far away from each other in the projection view presented similar structures and distinguishing attributes, as shown in Figure 7 (d, e). E_1 said, "This finding is an excellent demonstration of the effectiveness of the PLS-based clustering method, which takes structures and attributes into consideration synchronously."

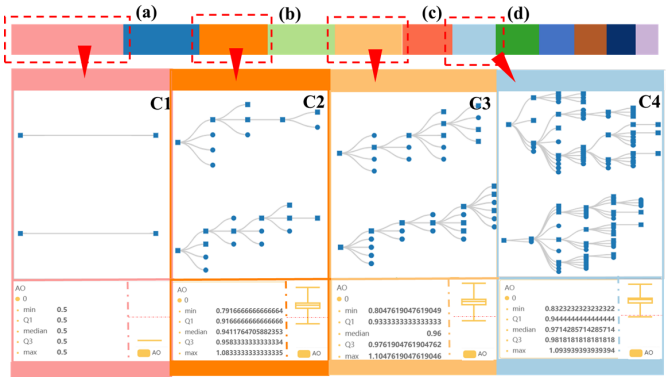

Fig. 8. Evaluate the quality of clustering and gain insights from structure-related views. (a), (b), (c) and (d) are the different clusters in the projection view based on the topology-based clustering.

6.2.2 Insight Discovery

The experts are interested in finding clusters of similar families to extract meaningful groups of families with specific features. For example, there are many families in which members may have the same one attribute or structure feature, but will respond to other features differently. They believe that if they could classify families into groups of similar features, they could gain deeper insights of clusters of interest.

In the process of exploring the structural projection view, E_2 discovered a strange issue that the area of a cluster corresponding the widest color patch in the summary glyph is unexpectedly small, as shown in Figure 6(a). Then, through careful observation of the family tree view, it is found that most of the families in that cluster are a simple hierarchical structure with a father and a son. "Since the structures of families in this cluster is so simple that these are almost identical, which cause many points to overlap, the cluster which should have occupied the largest area is only shown as a very small cluster." Expert said. The opposite situation is shown in the Figure 6(b). Then, E_2 clicked on the following stripes in turn, and found that with the decrease of the number of families in each cluster, the number of offspring in the family shows an increasing trend. In short, the number of families with multiple offspring is often less than that with fewer offspring. To verify this finding, the expert compared the boxplots of each cluster in the average offspring (AO), and found that the conclusion is valid, as shown in Figure 8. "The visual presentation of the hierarchical structure and the structural statistical indicators can help us capture and verify some patterns." said the expert.

Next,the expert E_2 turned his attention to the combination of different features. Through the control panel, E_2 selected "AA" and "VN" in the attribute box for projection. Then the different color clusters were presented in the projection view. By comparison of the features of different clusters from the family tree view and the parallel coordinate view, he found some interesting stories. For example, In the largest cluster, AA is between 20 and 32, and VN is almost 1, and the family structure is relatively simple. In the second large cluster, AA is between 35 and 45, and VN is almost 2, most of which are large families with many generations. In the smaller cluster, AA is between 30-50, and VN is almost 3. The complexity of the family structure in these clusters showed an increasing trend. This fully proves that families with longevity and more villages will be huger. Besides, from the comparison of the number of families in the cluster, it is found that the family structure at that time was not too complicated. From these, E_2 inferred two reasons according to the social background at that time, one is that it was difficult to record genealogy at that time, and the records might be lost or inaccurate, the other is that the war was in chaos, and the family's continuity was greatly affected by frequent changes. "It can be seen from both the glyph and the parallel coordinate view, the value of AG in each cluster are similar to each other, which conforms to a simple mathematical rule." said the expert. Finally, the expert concluded that our genealogical clustering system was just conducive to explore and analyze the dataset based on the fusion of structure and attributes.

6.3 Expert Interview

After the experts had explored the real-world graph datasets with our system, we conducted a semi-structured interview to collect the experts' feedbacks about the system usability, visual designs, and interaction, which are listed as follows.

System capability and effectiveness. The two experts were very

impressed by our system, especially the clustering model that supports the fusion of structure and attribute. E_1 commented that our system made it easier to quickly analyze the family trees that meeting the structure and attribute similarity. Besides, he added "Multiple attribute combinations provided by the system is convenient for me to specify attributes of interest and get the genealogical clustering under different similarity measures."

Visual design and interactions. The visualization views were appreciated by our experts. They found the cluster glyph designed in the cluster detail view was visually appealing and conveyed complementary information regarding structure and attribute. Experts commented that the fusion of each view and their logic interaction in the system interface was clear. In addition, experts also appreciated the collaboration of multi-view and interactive evaluation, which provide an intuitive and convenient method to represent, evaluate and explore cluster quality.

6.4 Discussion

Compared with traditional clustering strategies, our genealogy clustering model enhances the attribute features of the family tree. The main advantages are that we conduct a PLS model to fuse structures and attributes and design a glyph to represent the features of clusters. But there are still some issues not well resolved in this paper.

(1) In this paper, we utilize graphlet kernel and statistical method to measure the similarity of structure and attribute respectively between family trees. However, the features of genealogy have the problem of inconsistent expression, so it is difficult to fuse them effectively. Although PLS is utilized in this paper, the specific meanings of features are not fully considered to some extent. In the future work, we will explore a more effective feature fusion method to effectively and fully express the multi-dimensional features of data.

(2) In this paper, we provide a new glyph to jointly express the structure and attribute of clusters in the same space, but this presents a visual scaling problem. When the structure of a family tree is too complex or the number of families in a cluster is too large, our glyphs will face great pressure in terms of visual perception. In the future work, we will investigate more refined designs for feature representation to enhance the visual expansion, or integrate more elements into our visual designs to enrich the expression of designs.

(3) Clustering is a significant part of a simplified analysis of genealogical data, but it is not the end. According to the requirement of domain experts, the analysis of genealogical data needs to further explore the correlation between genealogy and humanities, society, history, economy and other fields. In the further work, we will introduce more algorithmic models to conduct an in-depth correlation analysis of the above relationships, and draw causal relationships to analyze social phenomena and gain more insights.

7 Conclusion

In this paper, the Graphlet kernel is utilized to represent the structures of graphs and a Partial Least Squares (PLS) model is designed to combine the structure and attribute information into a fused embedding space. Then we employed a distance-based genealogical clustering scheme on a real-word genealogy dataset. In addition, a set of visual interfaces are provided enabling users to interactively perform genealogical clustering and visually evaluate the similarity of clustering results from various perspectives. Quantitative comparisons and case studies based on real-world datasets have demonstrated the effectiveness of our system.

8 Acknowledgments

We would like to thank the reviewers for their thoughtful comments. The work is supported in part by the National Natural Science Foundation of China (No.61872314 and 61802339), the Open Project Program of the State Key Lab of CADCG of Zhejiang University (No.A2001), the Natural Science Foundation of Zhejiang Province, China (LY21F020029 and LY19F020011) and the Public Welfare Technology Applied Research Project of Zhejiang Province (No.LGF20G010003).

References

[1] A. Bezerianos, P. Dragicevic, J.-D. Fekete, J. Bae, and B. Watson. Geneaquilts: A system for exploring large genealogies. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1073–1081, 2010. doi: 10.1109/TVCG.2010.159

[2] L. Boudjeloud-Assala, P. Pinheiro, A. Blansché, T. Tamisier, and B. Otjacques. Interactive and iterative visual clustering. *Information Visualization*, 15(3):181–197, 2016.

[3] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[4] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE transactions on visualization and computer graphics*, 17(12):2581–2590, 2011.

[5] K. Chen and L. Liu. Vista: Validating and refining clusters via visualization. *Information Visualization*, 3(4):257–270, 2004.

[6] D. Davies. L., bouldin, d., w., a cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2)*, pp. 224–227, 1979.

[7] S. Dexter, G. Yarmish, and P. Listowsky. Parallel clustering of protein structures generated via stochastic monte carlo. In *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)*, pp. 410–413, 2016. doi: 10.1109/SMRLO.2016.71

[8] S. Fu, H. Dong, W. Cui, J. Zhao, and H. Qu. How do ancestral traits shape family trees over generations? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):205–214, 2018. doi: 10.1109/TVCG.2017.2744080

[9] T. Gu, M. Zhu, W. Chen, Z. Huang, R. Maciejewski, and L. Chang. Structuring mobility transition with an adaptive graph representation. *IEEE Transactions on Computational Social Systems*, 5(4):1121–1132, 2018.

[10] D. M. Hillis, T. A. Heath, and J. K. St. Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482, 2005.

[11] C.-X. Jin and Q.-C. Bai. Text clustering algorithm based on the graph structures of semantic word co-occurrence. pp. 497–502, 06 2016. doi: 10.1109/ISAI.2016.0112

[12] Kaizhongzhang, J. Wang, and Dennisshasha. On the editing distance between undirected acyclic graphs. *International Journal of Foundations of Computer Science*, 07, 11 2011. doi: 10.1142/S0129054196000051

[13] T. Kemp. Genealogy: Finding roots on the web. 60:X1–X2, 01 1999.

[14] S. Ko, S. Afzal, S. Walton, Y. Yang, and D. Ebert. Analyzing high-dimensional multivaríate network links with integrated anomaly detection, highlighting and exploration. In *Visual Analytics Science Technology*, 2015.

[15] X. Kong, Y. Chen, H. Tian, T. Wang, Y. Cai, and X. Chen. A novel botnet detection method based on preprocessing data packet by graph structure clustering. pp. 42–45, 10 2016. doi: 10.1109/CyberC.2016.16

[16] T. Kosaka and S. Sagayama. Tree-structured speaker clustering for fast speaker adaptation. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. i, pp. I/245–I/248 vol.1, 1994. doi: 10.1109/ICASSP.1994.389309

[17] D. Kutz. Examining the evolution and distribution of patent classifications. pp. 983– 988, 08 2004. doi: 10.1109/IV.2004.1320261

[18] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE transactions on visualization and computer graphics*, 24(1):142–151, 2017.

[19] H. Liao, Y. Wu, L. Chen, and W. Chen. Cluster-based visual abstraction for multivariate scatterplots. *IEEE transactions on visualization and computer graphics*, 24(9):2531–2545, 2017.

[20] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):91–100, 2016.

[21] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[22] Y. Liu, S. Dai, C. Wang, Z. Zhou, and H. Qu. Genealogyvis: A system for visual analysis of multidimensional genealogical data. *IEEE Transactions on Human-Machine Systems*, 47(6):873–885, 2017. doi: 10.1109/THMS.2017.2693236

[23] Y. Liu, Z. Guo, X. Zhang, R. Zhang, and Z. Zhou. uncertainty visualization in stratigraphic correlation based on multi-source data fusion. *Journal of Visualization*, 22, 08 2019. doi: 10.1007/s12650-019-00579-0

[24] S. L'Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, and J. Seo. Xclusim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC bioinformatics*, 16(11):1–15, 2015.

[25] E. Maguire, I. Koutsakis, and G. Louppe. Clusterix: a visual analytics approach to clustering. In *Symposium on Visualization in Data Science at IEEE VIS*, 2016.

[26] Meyer, Munzner, Styczynski, Wong, and Pfister. Pathline: A tool for comparative functional genomics.

[27] T. Munzner, F. Guimbretiere, S. Tasiran, L. Zhang, and Y. Zhou. Treejuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. *Acm Transactions on Graphics*, 22(3):p.453–462, 2003.

[28] C. Nobre, N. Gehlenborg, H. Coon, and A. Lex. Lineage: Visualizing multivariate clinical data in genealogy graphs. *IEEE Transactions on Visualization Computer Graphics*, 25(03):1543–1558, mar 2019. doi: 10.1109/TVCG.2018.2811488

[29] A. Papadopoulos and Y. Manolopoulos. Structure-based similarity search with graph histograms. In *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99*, pp. 174–178, 1999. doi: 10.1109/DEXA.1999.795162

[30] Partl, Gratzl, Streit, A. M, Wassermann, Pfister, Schmalstieg, and Lex. Pathfinder: Visual analysis of paths in graphs. *Computer graphics forum : journal of the European Association for Computer Graphics*, 2016.

[31] M. Rahman, M. A. Bhuiyan, M. Rahman, and M. Hasan. Guise: a uniform sampler for constructing frequency histogram of graphlets. *Knowledge Information Systems*, 38(3):511–536, 2014.

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

[33] Shaw, PD, Graham, M, Kennedy, J, Milne, I, Marshall, and DF. Helium: visualization of large scale plant pedigrees. *BMC BIOINFORMATICS*, 2014.

[34] N. Tsuya, F. Wang, G. Alter, and J. Lee. *Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700-1900*. 01 2010. doi: 10.7551/mitpress/8162.001.0001

[35] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[36] J. Wang, J. Wu, A. Cao, Z. Zhou, and Y. Wu. Tac-miner: Visual tactic mining for multiple table tennis matches. *IEEE Transactions on Visualization and Computer Graphics*, PP, 2021.

[37] Y. Wang, C. Shi, L. Li, H. Tong, and H. Qu. Visualizing research impact through citation data. *ACM Trans. Interact. Intell. Syst.*, 8(1), Mar. 2018. doi: 10.1145/3132744

[38] M. Wattenberg. Visual exploration of multivariate graphs. pp. 811–819, 01 2006. doi: 10.1145/1124772.1124891

[39] D. Weng, C. Zheng, Z. Deng, M. Ma, J. Bao, Y. Zheng, M. Xu, and Y. Wu. Towards better bus networks: A visual analytics approach. *CoRR*, abs/2008.10915, 2020.

[40] D. R. White and P. Jorion. Representing and computing kinship: A new approach. *Current Anthropology*, 33(4):454–463, 1992. doi: 10.1086/204097

[41] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni. Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):935–944, 2016. doi: 10.1109/TVCG.2015.2467194

[42] J. Xia, T. Chen, L. Zhang, W. Chen, Y. Chen, X. Zhang, C. Xie, and T. Schreck. Smap: A joint dimensionality reduction scheme for secure multi-party visualization. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 107–118. IEEE, 2020.

[43] J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A. K. Tung. Ldsscanner: Exploratory analysis of low-dimensional structures in high-dimensional datasets. *IEEE transactions on visualization and computer graphics*, 24(1):236–245, 2017.

[44] J.-z. Xia, Y.-h. Zhang, H. Ye, Y. Wang, G. Jiang, Y. Zhao, C. Xie, X.-y. Kui, S.-h. Liao, and W.-p. Wang. Supoolvisor: a visual analytics system for mining pool surveillance. *Frontiers of Information Technology & Electronic Engineering*, 21:507–523, 2020.

[45] M. Yang, C. Wu, and T. Xie. Information propagation dynamics model based on implicit cluster structure network. pp. 1253–1257, 06 2020. doi: 10.1109/ITOEC49072.2020.9141733

[46] S. Ye, Z. Chen, X. Chu, Y. Wang, and Y. Wu. Shuttlespace: Exploring and analyzing movement trajectory in immersive visualization. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2020.

[47] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu. A survey of visual analytics techniques for machine learning, 08 2020.

[48] J. Zhao, F. Chevalier, C. Collins, and R. Balakrishnan. Facilitating discourse analysis with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2639–2648, 2012. doi: 10.1109/TVCG.2012.226

[49] Y. Zhao, H. Jiang, Y. Qin, H. Xie, Y. Wu, S. Liu, Z. Zhou, J. Xia, F. Zhou, et al. Preserving minority structures in graph sampling. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[50] Y. Zhao, X. Luo, X. Lin, H. Wang, X. Kui, F. Zhou, J. Wang, Y. Chen, and W. Chen. Visual analytics for electromagnetic situation awareness in radio monitoring and management. *IEEE transactions on visualization and computer graphics*, 26(1):590–600, 2019.

[51] F. Zhou, X. Lin, C. Liu, Z. Ying, P. Xu, L. Ren, T. Xue, and L. Ren. A survey of visualization for smart manufacturing. *Journal of Visualization*, 22, 11 2018. doi: 10.1007/s12650-018-0530-2

[52] Z. Zhou, C. Shi, X. Shen, L. Cai, H. Wang, Y. Liu, Y. Zhao, and W. Chen. Context-aware sampling of large networks via graph representation learning. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 10 2020. doi: 10.1109/TVCG.2020.3030440

[53] Z. Zhou, Z. Ye, Y. Liu, F. Liu, Y. Tao, and W. Su. Visual analytics for spatial clusters of air-quality data. *IEEE computer graphics and applications*, 37(5):98–105, 2017.

[54] Z. Zhou, Z. Ye, J. Yu, and W. Chen. Cluster-aware arrangement of the parallel coordinate plots. *Journal of Visual Languages Computing*, 46, 10 2017. doi: 10.1016/j.jvlc.2017.10.003