ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijsem.org

Vol 9, Issuse.4 Dec 2021

Exploring Multilevel Clusters as a Network

[1]*G Shiva Prasad,*
[2]*Bandam Naresh,*
[3]*E Muralidhar Reddy,*
[4]*Dr. T.Sreenivasulu, .*

Abstract

Visual analytics has gained a lot of attention for its ***ability to identify meaningful patterns in large datasets. The identification and depiction of clusters is one of the most typical activities. In heterogeneous datasets, on the other hand, this is more difficult since the data must be seen from several angles. A dataset with a high degree of variability may in fact conceal trends that lie under the surface. An analytic framework for examining the clustering at various degrees of detail is provided by the use of dendrograms (graphical representations of agglomerative hierarchical clustering). Nevertheless, as the dataset becomes larger, dendrograms get congested, and a single cut might be inadequate in diverse datasets to demarcate multiple clusters. Visual analytics technique dubbed MCLEAN is proposed in this study to assist the user in discovering and exploring clusters. Data may be represented in a wide variety of ways by modifying its spatialization using graph-based transformations of relational data. Thus, multilevel representations of the clustered dataset are combined with techniques for locating the communities that live there. User exploration and data analysis might begin with the presentation of heuristics findings to the public. Users are invited to compare the findings of MCLEAN and the dendrogram while exploring a diverse dataset in order to assess our suggested technique. Qualitative findings show that MCLEAN is a useful tool for helping people identify clusters in diverse datasets. An R programme implementing the suggested approach is readily accessible.***

Introduction

The technique for addressing the clustering issue is separate from the problem of determining the number of clusters in a dataset. Many times, it is difficult to determine how many groups a dataset should have based on its structure and size, as well as the required level of clustering resolution by a user. There are several factors to consider when deciding how many clusters to utilise; in general, it's a trade-off between the most compression and the highest resolution possible.For example, k-means, DBSCAN, and hierarchical clustering all use a variety of clustering methods for determining the number of clusters. In certain cases, these factors may directly or indirectly determine the number of clusters that are generated by the algorithm. Pre-existing data knowledge or time-consuming trial and error are required to set these values. It's also possible that a single cutoff might obscure intriguing structures behind it. Automated clustering approaches sometimes overlook unique aspects of clusters, such as their density or sparsity, since there may not be a single logical cutoff in the actual world.

"The clustering process is not complete until it is examined, verified, and approved by the user," says the statement. As a result, visual validation and exploration may help clarify the clustering structure and uncover patterns, outliers, and clusters that otherwise would be difficult to see. These visualisations aid in swiftly assimilating the data and give insights that supplement textual outputs or statistics summaries. How well-defined are the clusters, how far apart they are, what their size is, and whether or not the observations belong strongly to the cluster or are just marginally associated with it? There are many potential clustering situations to explore, and it might be difficult for the user to identify related groupings of records (i.e., patterns).

Assistant Professor, Mail ID:gspgsp26@gmail.com
Assistant Professor ,Mail ID:nareshbandam4@gmail.com
Assistant Professor ,Mail ID:krishna81.reddy@gmail.com
Associate Professor ,Mail ID:@gmail.com
Department of CSE Engineering,
Pallavi Engineering College Hyderabad, Telangana 501505

A frequently used and successful approach for answering these concerns is hierarchical clustering, which uses a to impose a hierarchy on the clustering in order to give multiple degrees of information for examining the grouping. The graphical depiction of the tree's cutoff selection process provides insights that show the solution's adequacy, however hierarchical clustering has several drawbacks: (1) the dendrogram representation gets clogged with big datasets; (2) a single cut of the dendrogram is adequate for a homogenous dataset. It is possible, however, that several cuts at various levels will be necessary when the dataset is heterogeneous. The cutoff will conceal all but one pattern if there are many tiers of patterns.Clustering approaches often follow a predetermined sequence: loading a dataset, specifying parameters, executing an algorithm, and visualising the results. To put it another way, clustering is most often employed to analyse data rather than to investigate it. It is feasible to make the clustering process dynamic by integrating visualisation and algorithm into the same model. Data mining benefits greatly from the interactive visual clustering (IVC) architecture proposed by since it enables users to participate in the clustering process by using their visual perception and domain expertise. As suggested by, we think that by combining the clustering technique with an adaptation of the visualisation environment, we can give users with a highly natural manner to explore datasets.

The MCLEAN (Multi-Level Clustering Exploration As Network) approach is a novel and generic clustering and exploration approach that allows for: (1) exploration of the dataset using an overview-plus-detail representation, (2) simplification of the dataset using aggregation based on the similarity of data elements, (3) detection of substructures using community detection algorithms, and (4) inclusion of the substructures. Synergistic approaches to data exploration that mix the power of computers, community identification tools and people' perceptual abilities to see trends are used in our approach. Using this strategy, the user may engage with the algorithm results in a visible way. Hierarchical clustering methods are used to determine the best clusters, which are then shown in a simplified network form using interactive tools. Understanding the patterns of interaction between things, discovering entities with intriguing functions, and identifying intrinsic groupings or clusters of entities may all be accomplished using network visualisations. An R package implementing the MCLEAN approach can be found atIn the following sections, you'll find a breakdown of the content. A brief history of multi-level clustering and graph visualisation approaches is provided in the section under "Background" in this paper. A detailed description of the suggested visualisation strategy for clustering exploration is given in the section 'Methods, followed by an assessment of our approach in the part entitled 'Evaluation.' The section titled "Conclusions and Future Work" concludes the paper by outlining the findings and outlining potential future paths.work.

Background

Users may apply tacit knowledge in the clustering process so that substructures can be discovered. Multilevel data visualisation is made possible by the use of an overview-plus-detail approach that combines an overall view with graphs to show the relationship between several groups of data. An example of visual multilevel clustering and a network transformation of data to discover patterns are provided to put our work into perspective. Clustering in a multi-tiered fashion

Even though there are a variety of clustering approaches, only a select number allow for visual inspection. More than a handful of the clusters may be explored interactively at varying degrees of depth. Clustering analysis, however, is becoming more and more dependent on visual engagement, since experts are able to lead the analysis to create more relevant findings. Algorithms aren't always able to account for the user's tacit knowledge, which frequently drives their judgments. Therefore, a human being must be involved in the decision-making process and in the analysis.

In numerous domains, including biology, social sciences, and computer vision, hierarchical clustering has long been employed because of the simplicity with which the user can comprehend the result. A single similarity criteria is used to choose the clusters, and the tree is chopped at the same height. For vast and diverse datasets, a more flexible method is needed to enable the user to experiment with alternative clustering situations. There are a variety of ways to chop down the tree. developed a method that automatically divides the dendrogram into levels depending on the morphology of the branches.

suggested guided piecewise snipping as a method for finding clusters in a dendrogram. With the piecewise rather than the fixed-height cut and the incorporation of external data to decide on the ideal cut, this technique addresses the shortcomings of the fixed-height approach. Similarly, is a visual aid for dendrograms of diverse data sets at various degrees of detail in the same field of inquiry.

Splitting a data set into k clusters based on certain criteria is the goal of k-means and CLARANS approaches. provided semi-interactive data exploration via iterative clustering of multidimensional datasets. User participation in clustering tasks is made possible by their framework, which links users to the data mining process. Using Looney's method, tiny clusters are removed and re-assigned to more dense areas in an iterative manner. This improves the accuracy of the clustering findings. Also offered an interactive method for exploring huge numbers of paths using clustering methods by integrating the user's preferences into the clustering process using 2D data projections. Refinements are made by users in order to organise trajectories in a more efficient manner.

Graph representation

Visual depiction of dendrograms is not scalable to huge datasets. For example, Chen, MacEachren and Peuquet (2009) suggested a method in which the dendrogram representation is simplified by using a uniform threshold. Using this method, the dendrogram may be summarised and shown in a more compact form. Multi-level cuts and data exploration are not supported by this tool.

Many approaches exist for determining a graph representation from a matrix whose elements describe the degree of similarity between data points. Hierarchical clustering techniques often employ the idea of a graph to represent data elements. For example, Ploceus (Liu, Navathe, &Stasko, 2014) provides a way for doing network-based visual analysis on tabular data in a more generic manner. Direct manipulation of data tables allows users to design and modify networks in a variety of ways. Ploceus provides a seamless analytic experience by combining dynamic network modification with visual exploration.

The WhatsOnWeb system (Di Giacomo et al., 2007) makes use of the graph-based visualisations provided by the results of a Web search engine.. If documents are sufficiently semantically linked, the system creates a network from a search query. This graph then connects any nodes it finds. Using an edge weight and topological clustering technique, the network hierarchy is formed and various layers of information are shown, creating a visual representation of the relationship's strength.

Duman, Healing &Ghanea-Hercock (2009), Desjardins, MacGlashan&Ferraioli (2007, 2008) and Beale (2007, 2012) use a force-directed graph architecture to encode distance between items as forces in their grouping and exploration systems. Using partitioning-based approaches, it is possible to allocate clusters by projecting distances into a smaller dimension. Nodes are spatially represented by omitting links, which makes it easier to interpret the spatialization of the nodes. In contrast to typical approaches like projection pursuit or multi-dimensional scaling, they provide an alternative.

There are methods given for navigation of clustering results for large-scale graph visualisation systems such as for MCLEAN's network exploration. Expanding and compressing nodes in a graph is made possible by these tools (meta-nodes). However, while browsing clustered graphs with deeper hierarchies, people typically lose their bearings.

Methods

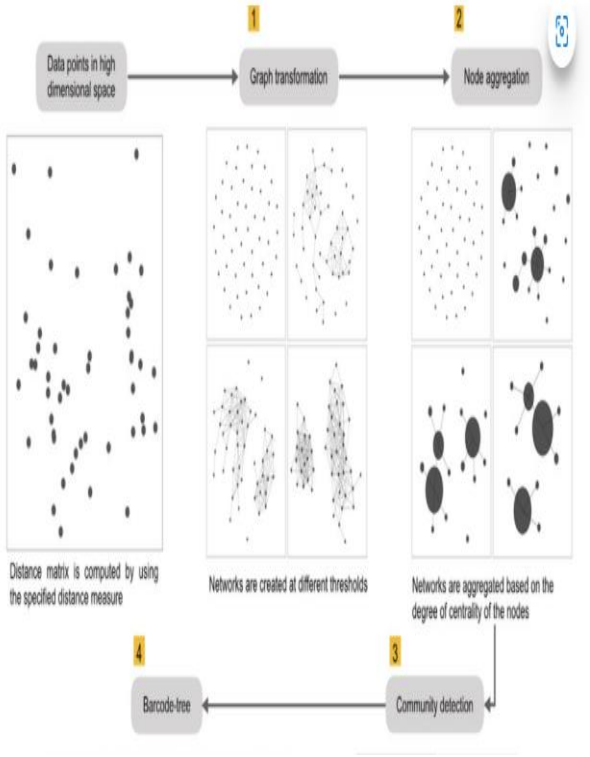
In the MCLEAN approach, a similarity matrix of all data records is used as input, and a simplified graph representation shows a greater abstraction of the clustering process is produced. Visual representations are combined throughout MCLEAN's work. If you want to see how changing values of the parameter affect the overall cluster structure, an overview plot (barcode-tree), which is connected to a dendrogram and the topological barcode plot, is what you're looking for first. Node-link plots, on the other hand, show the clustering findings based on a specific. There are two levels of clustering information in this node-link diagram. To begin, data clusters over this level belong to graph related components. Another way to think about it is that various colours inside a single component imply that this particular subnetwork would be separated into many segments if it were to be analysed using a more strict.

It is a subgraph in which all of the nodes are linked either directly or indirectly. Clusters in the dataset are defined using linked components. Community discovery technique is also used by MCLEAN in order for subclusters to be discovered inside related components It is thus possible to determine if a cluster is unique from another by using user knowledge (tacit or explicit). In diverse data sets, this uncertainty is widespread.

We employ an agglomerative algorithm, like many other clustering methods, which relies on a single parameter. There is a threshold () in this parameter that determines the separation of two data pieces in order for them to be united. The MCLEAN technique

and topological data analysis have a lot in common (TDA). A topological structure perspective on multidimensional spaces, interpreting the persistent homology by calculating the number of connected components (b0 from betti numbers) and using the persistence concept to define the optimal threshold of network representation prove that although the goals are different, they share a common philosophy of analysis (Topaz, Ziegelmeier& Halverson, 2015).

Figure 1 illustrates the four elements of the MCLEAN method: Node-link representations of the distance matrix are transformed into node-link representations depending on the threshold chosen; the network is simplified, and community identification techniques are used to identify substructures; and the resultant networks are explored for various threshold values.



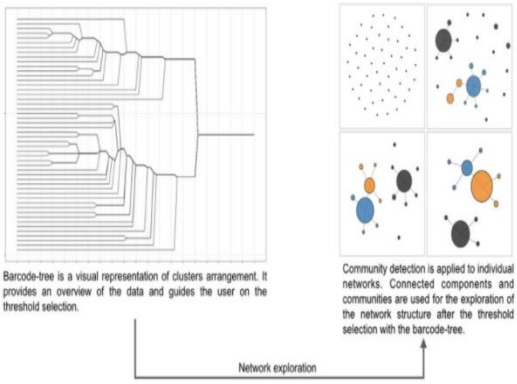
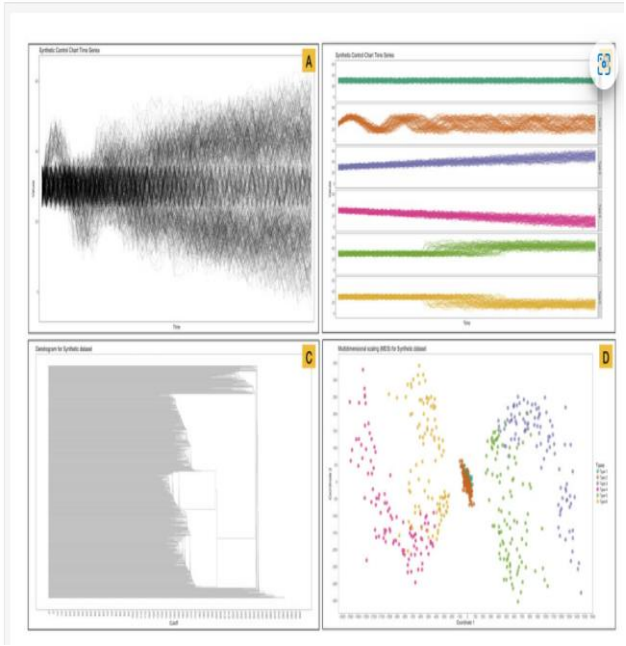


Figure 1: Workflow diagram of MCLEAN algorithm, consisting of four steps: (1) graph transformation, (2) node aggregation, (3) community detection and, (4) barcode-tree creation

This section uses a dataset from the UCI repository to demonstrate the process (see Fig. 2). In accordance with Alcock&Manolopoulos, this collection provides 600 instances of control charts synthesised from scratch (1999). The temporal sequences were compared using Dynamic Time Warping (DTW). A dendrogram (Fig. 2C) and a scatterplot of the first two dimensions of multidimensional scaling are shown in Figure 2, as are representations of the raw data (Figs. 2A and 2B) (Fig. 2D).



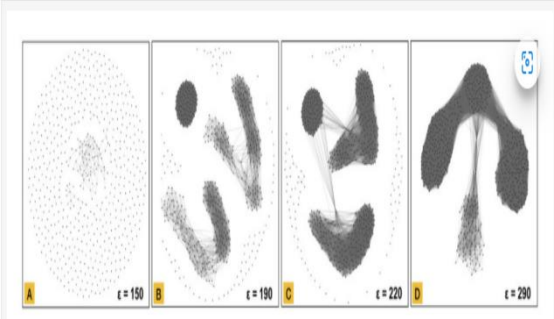
Graph transformation

The data items are projected in a reduced dimension ordination space using Multidimensional Scaling (MDS). The dimensionality of substructures in the data isn't as important as the ease of display when choosing between two and three dimensions. As a

result of these projections' varying distances and optical restrictions, certain patterns might be obscured. Complex datasets may be addressed by changing the spatialization (such as network visualisation). The MDS applied to the synthetic dataset in Fig. 2D is an illustration of these flaws.

In spite of the lack of explicit network semantics in the distance matrix, MCLEAN employs this method to modify the encoding of distances via the utilisation of network linkages To prevent overlapping between nodes, the method used in the drawing of the network (i.e., force-directed graph) is optimised.

As with the DBSCAN approach (Ester et al., 1996), the graph transformation stage of MCLEAN uses a parameter that determines the radius that classifies points as being in each other's vicinity. It's possible to specify the minimum number of points that may make up a cluster in DBSCAN by setting the second option, numPts, to 0. But in MCLEAN, all data points are considered network nodes, and data points that are close enough to each other are interconnected. This stage yields a network in which a route exists between two nodes only if they are members of the same linked component. Clusters are shown as linked components in topological space at this level of the process. On the same dataset, four distinct snapshots of the graph modification process are shown in Figure 3. Number of connections between nodes rises as increases.



Results and analysis

A barcode tree was used to find patterns, a dendrogram was used to compare thresholds, and finally a network representation and a barcode tree were combined to find patterns. After completing the activity, a questionnaire was given to gather user feedback and satisfaction.

Fig. 2B shows the varied underlying patterns that we tested participants' ability to recognise in a preliminary assessment. In the temporal dataset, five users (participants A–E) used the barcode-tree only for steps of 5 from 0 to 300 to identify four distinct patterns (Fig. 7). With the same picture in mind, Participant F detected two distinct patterns, one of which included all three of the other patterns (see Fig. 9). Dendrogram investigation yielded the same findings. Pattern A4 was deemed an anomaly by participants D and E as well. Users A-E combined Type 3 and Type 5 signals (see Fig. 2B) to form a single pattern (pattern A2; see Fig. 9A), and Type 4 and Type 6 signals (see Fig. 9A) to form pattern A3, when recognising four patterns. However, in both circumstances, the pairings act in comparable ways, but in opposite directions: either a constant rise and fall, or a sudden change. Due to the DTW sequence alignment, the global distance between the various sorts of patterns in each pair is modest compared to the rest. Between three and five patterns were detected by participants A–E when analysing the dendrograms (Fig. 9B). In the centre, there were two or three groups, and users B and E mistook the heterogeneous data components (pattern B1) for two separate clusters because of where the branches were located. One cluster included Type 1 and Type 3 signals, whereas Type 4 and Type 6 signals were found by Participant C in the second cluster. Based on these findings, it seems that a dendrogram's perception has deteriorated somewhat compared to that of a barcode tree, and that the dendrogram may be misinterpreted owing to its branching structure. Participants' perception of changes in tree resolution did not alter when the resolution was increased, but it did when it was lowered, as seen in Figs. 1 and 2 (Fig. 8A and 8B). When we examined the number of related components in increments of 20, three people (B, C, and E) found six patterns, as shown in Fig. 8F. Due to this characteristic, there are several ways to interpret the same data depending on the resolution used.

While participants A-E picked one cutoff between 180 and 195, characterising two or three clusters and ungrouped data-elements, only participant F struggled with cutoff selection. Participants A-E chose the same threshold using the barcode-tree. Participant D went on to examine a second limit of 220. The threshold 285 was selected by participant F with the express purpose of investigating the network's representation. Because no restriction was placed on the number of cutoffs, the user was free to experiment with other partitioning strategies. Overall, the barcode-tree gave consumers greater confidence in selecting thresholds than the dendrogram. A persistent segment begins at the 185-point threshold and continues until the 202-point criterion is reached by the joining of three clusters. Discussion with the participants suggested that this persistence in the barcode-tree improves readability and hence increases the level of trust in the selection of the threshold. An incorrect interpretation of cutoff

selection may occur when the dendrogram's leaves are not optimally ordered and their binary union is not optimised, leaving certain components outside of a viable cluster.

Conclusion and future work

An interactive, multi-resolution study of clustering findings in complicated datasets is provided in this work. According to results from usability studies, making the data more transparent and reassuring to the user may help them better absorb the information. A user-centric approach to information discovery may benefit from the fact that the quantity and quality of clusters are closely correlated with user activity, and we feel this is truly a strength of the system (one that was deliberately designed for). As useful as these network and barcode-tree representations are to the user, there are several obvious areas for further research. There are certain datasets where average or full linkage clustering could be more beneficial than single linkage clustering, which is what we're doing now (especially where the distance matrix does not exhibit gaps). There are also visual artefacts (parallel lines merging with a cluster) in the existing visual encoding of the barcode-tree. Further ways for assessing how data items are integrated across thresholds should be investigated, too.
Clustering results may keep their multi-level patterns if the domain user is included into the process itself. MCLEAN makes it easier to incorporate tacit or other user knowledge into the understanding and investigation of clustering results while also simplifying the representation of groups, particularly in the presence of noise or outliers. We propose that the MCLEAN method offers new possibilities for cluster visualisation and exploration that go beyond previous approaches.

References

Abello J, Van Ham F, Krishnan N. 2006. Ask-graphview: a large scale graph visualization system. *IEEE Transactions on Visualization and Computer Graphics* **12**(5):669-676

Alcock RJ, Manolopoulos Y. 1999. Time-series similarity queries employing a feature-based approach. In: 7th Hellenic conference on informatics. River Edge. World Scientific Publishing. 27-29

Archambault D, Munzner T, Auber D. 2008. GrouseFlocks: steerable exploration of graph hierarchy space. *IEEE Transactions on Visualization and Computer Graphics* **14**(4):900-913

Archambault D, Munzner T, Auber D. 2009. TugGraph: path-preserving hierarchies for browsing proximity and paths in graphs. In: Visualization symposium, 2009. PacificVis' 09. IEEE Pacific. Beijing, China. Piscataway: IEEE. 113-120

Beale R. 2007. Supporting serendipity: using ambient intelligence to augment user exploration for data mining and web browsing. *International Journal of Human-Computer Studies* **65**(5):421-433

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

Bruneau P, Otjacques B. 2013. An interactive, example-based, visual clustering system. In: Information visualisation (IV), 2013 17th international conference. London. Piscataway: IEEE. 168-173

Chen J, MacEachren AM, Peuquet DJ. 2009. Constructing overview+ detail dendrogram-matrix views. *IEEE Transactions on Visualization and Computer Graphics* **15**(6):889-896

Desjardins M, MacGlashan J, Ferraioli J. 2007. Interactive visual clustering. In: Proceedings of the 12th international conference on intelligent user interfaces. Honolulu, Hawaii. New York: ACM. 361-364

Di Giacomo E, Didimo W, Grilli L, Liotta G. 2007. Graph visualization techniques for web clustering engines. *IEEE Transactions on Visualization and Computer Graphics* **13**(2):294-304

Duman H, Healing A, Ghanea-Hercock R. 2009. An intelligent agent approach for visual information structure generation. In: Intelligent agents, 2009. IA'09. IEEE symposium on. Nashville. Pisacataway: IEEE. 55-62

Eades P, Feng Q-W. 1996. Multilevel visualization of clustered graphs. In: International symposium on graph drawing. Springer. 101-112

Eades P, Huang ML. 2000. Navigating clustered graphs using force-directed methods. *Journal of Graph Algorithms and Applications* **4**(3):157-181

Ester M, Kriegel H-P, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96 proceedings of the second international conference on knowledge discovery and data mining. Portland, OR. Palo Alto: AAAI Press. 226-231

Fisher R, Marshall M. 1936. Iris data set. *UC Irvine Machine Learning Repository*.

Friedman J, Hastie T, Tibshirani R. 2001. *The elements of statistical learning*. New York: Springer Series in Statistics New York. Vol. 1

Jain AK, Murty MN, Flynn PJ. 1999. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31**(3):264-323

Keim DA, Mansmann F, Thomas J. 2010. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter* **11**(2):5-8

Langfelder P, Zhang B, Horvath S. 2007. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**(5):719-720**Lee H, Kihm J, Choo J, Stasko J, Park H. 2012** . Ivisclustering: an interactive visual document clustering via topic modeling. In: *Computer graphics forum*. New Jersey: Wiley Online Library. Vol. 31:1155-1164

1. **Liu Z, Navathe SB, Stasko JT. 2014**. Ploce us: modeling, visualizing, and analyzing tabular data as networks. *Information Visualization* **13**(1):59-89