







Memberships Networks for High-Dimensional Fuzzy Clustering Visualization

Leandro Ariza-Jiménez¹  , Luisa F. Villa² , and Olga Lucía Quintero¹ 

¹ Mathematical Modeling Research Group, Universidad EAFIT, Medellín, Colombia
{larizaj, oquinte1}@eafit.edu.co

² System Engineering Research Group, ARKADIUS, Universidad de Medellín,
Medellín, Colombia
lvilla@udem.edu.co

Abstract. Visualizing the cluster structure of high-dimensional data is a non-trivial task that must be able to deal with the large dimensionality of the input data. Unlike hard clustering structures, visualization of fuzzy clusterings is not as straightforward because soft clustering algorithms yield more complex clustering structures. Here is introduced the concept of membership networks, an undirected weighted network constructed based on the fuzzy partition matrix that represents a fuzzy clustering. This simple network-based method allows understanding visually how elements involved in this kind of complex data clustering structures interact with each other, without relying on a visualization of the input data themselves. Experiment results demonstrated the usefulness of the proposed method for the exploration and analysis of clustering structures on the Iris flower data set and two large and unlabeled financial datasets, which describes the financial profile of customers of a local bank.

Keywords: Fuzzy clustering · Clustering visualization · Membership network · High-dimensional data

1 Introduction

Practical problems and applications related to data clustering not only can benefit from organizing data into unknown groups, but also from understanding how groups are constituted and related to each other [27]. One way to obtain this information is by using methods to visualize clustering results [16, 17]. Since most data of interest are high-dimensional, the visualization of the clustering structure of such data is a non-trivial task that must be able to deal with the large dimensionality of the input data.

For clustering structures obtained from the application of hard clustering algorithms on high-dimensional data, an indirect and straightforward solution to show these cluster structures is projecting the original data down to two- or three-dimensional spaces [21, 23]. Then, different colors or symbols can be

used to represent how data objects were clustered in a scatter plot of the data. Traditional methods to obtain this low-dimensional spaces are principal component analysis (PCA), multidimensional scaling (MDS), and self-organizing maps [2, 25]; or new methods like the *t*-Stochastic Neighbor Embedding algorithm [18, 19]. However, due to these methods reduce somehow the data dimensionality, an accurate representation of the original cluster structure is not guaranteed [29].

On the contrary, the visualization of fuzzy clusterings is not as straightforward as in the case of crisp clustering results. Indeed, unlike the output of a hard clustering algorithm, which is just a list of clusters with their corresponding members, soft clustering algorithms yield complex clustering structures, wherein data objects can belong to one or more clusters with probabilities [30].

So far, several works have addressed the problem of visualizing high-dimensional fuzzy clusterings. In [1, 5], a method is proposed to visualize fuzzy clustering results by performing an iterative process based on an MDS method that maps the cluster centers and the data into a two-dimensional space taking into account the membership values of the data. In [3], a method is presented for the interactive exploration of fuzzy clusters using the novel concept of neighborgrams, which is not well suited for medium-sized data sets. In [12], a technique is introduced to represent high-dimensional fuzzy clusters as intersecting spheres, and it is suitable for larger datasets; however, it heavily depends on the use of a three-dimensional visualization to preserve the overlapping regions appropriately in the original space. In [26, 31], Radviz, a radial non-linear visualization tool that displays multidimensional data in a two-dimensional projection, is used for developing a visualization that expresses the overall distribution of the membership degrees of all data points in a fuzzy clustering; however, as [31] pointed out, scalability is a problem for the Radviz-based visualizations when the number of clusters is large. In [24] fuzzy partitions from data are visualized by using MDS to map the degree of belongingness of data objects to clusters into a metric vector space which has mathematical dimensions.

In this work is proposed a straightforward representation of the fuzzy clustering results from high-dimensional data based on a simple network-based scheme, without relying on a visualization of the data itself. To this end, here is introduced the concept of membership network, which is an undirected weighted network constructed based on the fuzzy partition matrix that represents a particular fuzzy clustering.

The remainder of this paper is organized as follows: Sect. 2 briefly presents a theoretical background for data clustering and networks. Section 3 introduces the proposed network-based method to represent fuzzy clustering structures in high-dimensional data. Results and discussion are presented in Sect. 4, and finally, the concluding remarks can be found in Sect. 5.

2 Theoretical Background

As mentioned before, this work deals with a network-based representation of fuzzy clusterings on multidimensional data. Consequently, we first introduce here the basics for data clustering and networks.

2.1 Data Clustering

Clustering is the term used for methods whose objective is to partition a given set of unlabeled data objects into groups, known as clusters, such that data objects in the same group are similar and those in different groups are dissimilar to each other [4].

Suppose that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ is a set of N d -dimensional data objects, where $\mathbf{x}_j \in \mathbb{R}^d$. There are two broad forms of clustering, namely hard clustering and soft clustering, also referred to as crisp and soft clustering, respectively. In the first case, clustering methods obtain a K -partition of \mathbf{X} , $C = \{C_1, C_2, \dots, C_K\}$, $K \leq N$, such that $C_i \neq \emptyset$, $C_1 \cup C_2 \cup \dots \cup C_K = \mathbf{X}$, and each data object belongs to exactly one cluster, i.e. $C_i \cap C_j = \emptyset$, for $i, j = 1, \dots, K$ and $j \neq i$.

On the contrary, soft clustering considers that the boundaries between clusters are ambiguous, so that a data object \mathbf{x}_j can belong to more than one cluster C_i with a degree of membership $u_{ij} \in [0, 1]$, such that

$$\sum_{i=1}^K u_{ij} = 1, \quad \forall j = 1, \dots, N \quad (1)$$

and

$$0 < \sum_{j=1}^N u_{ij} < N, \quad \forall i = 1, \dots, K. \quad (2)$$

Based on the above membership degrees, a fuzzy clustering can be represented by a $K \times n$ matrix $U = [u_{ij}]$, known as the fuzzy partition matrix.

The most representative algorithms for hard and soft clustering are the K -means and the Fuzzy c -Means (FCM) algorithms, respectively. Both these algorithms are center-based clustering algorithms, that is, algorithms that look for a predefined number of clusters. These algorithms assume the existence of clusters by using the distances of data points from the cluster centers. Then, these algorithms perform an iterative optimization procedure to seek an optimal clustering of data. FCM is considered the fuzzy counterpart K -means and uses a parameter $m \in [1, \infty]$ that controls the ‘‘fuzziness’’ of the resulting clusters. In particular, when m is close to 1, the entries of the fuzzy partition matrix U converges to 0 or 1, and clusters become crispier; whereas when m increases, the entries of the same matrix decreases and clusters become fuzzier [30].

2.2 Networks

A network is a set of items, which we will call nodes or vertices, with connections between them, called links or edges [22]. Network nodes can be organized into groups, commonly called communities, that have a higher probability of being connected between them than to members of other groups [6].

Formally, a network can be denoted as $G = (V, E)$, where V is the set of nodes and E is the set of links such that $E \subseteq V \times V$. The pair (p, q) is the link connecting the node p with the node q . Links can be undirected, i.e. if $(p, q) \in E$ then $(q, p) \in E$. In addition, links can be associated with a real number, called its weight, by defining a weight function $\omega : E \rightarrow \mathbb{R}$. Finally, $|V|$ and $|E|$ represent the number of nodes and links, respectively.

Importance of networks lies in its ability to represent and facilitate the study and analysis of the interactions between real-world entities [22]. Network visualization can improve the understanding of the structure of those interactions [10]. To this end, layout algorithms that automatically arrange the nodes and links in an aesthetically pleasing way are used (see [10, 13] for a recent review of these algorithms). For visualization purposes, the family of force-directed algorithms is commonly used today [7, 11, 14]. These algorithms model a network as a physical system where nodes are attracted and repelled according to some force [10]. Some of the algorithms included in this family are Fruchterman-Reingold [8], GRIP [9], OpenOrd [20], and ForceAtlas2 [15].

3 Membership Networks

Here is introduced a novel method to visualize clusterings obtained from the application of soft clustering algorithms on high-dimensional data. Rather than using the original data itself, the cluster memberships are used as a clue to link the data objects to the resulting clusters and revealing the similarities among the formers through a membership network. Moreover, this kind of visualization facilitates the understanding of the uncertainties present in the data without any *a priori* knowledge or assumptions.

Consider a set of N data objects, $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_j \in \mathbb{R}^d$. Let $A = \{A_1, \dots, A_C\}$ be a clustering provided by a center-based soft partitioning clustering algorithm, e.g. FCM, represented by a C -by- N fuzzy partition matrix $U = [u_{ij}]$ and let a_i be the center of the partition A_i for $i = 1, \dots, C$.

An undirected weighted membership network, $G_U = (V, E)$, that represents the above fuzzy clustering can be constructed as follows:

1. Consider each data object \mathbf{x}_j and each cluster center \mathbf{a}_i as node of G_U , i.e. let $V = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \cup \{a_1, \dots, a_C\}$.
2. Link each data object \mathbf{x}_j with every cluster center a_i to represent the belonging of \mathbf{x}_j to more than one cluster, i.e. let $E = \{(\mathbf{x}_j, a_i), i = 1, \dots, C; j = 1, \dots, N\}$.
3. Associate to each link $(\mathbf{x}_j, \mathbf{a}_i)$ a weight $\omega(\mathbf{x}_j, \mathbf{a}_i)$ equal to the degree of membership of \mathbf{x}_j in the cluster A_i , i.e. let $\omega(\mathbf{x}_j, \mathbf{a}_i) = u_{ij}$.

A mathematical property of the obtained membership network is that it allows computing a cluster validity index for fuzzy clusterings, called partition coefficient (PC) [28], as $PC = (2|V|)^{-1}\text{Tr}(W^2)$, where W is the symmetric $|V|$ -by- $|V|$ matrix obtained when all of the weight edges of the graph G_U are recorded in a single matrix.

Figure 1 provides an example to illustrate the proposed method for fuzzy clustering visualization. Eight multidimensional data objects were clustered into two groups using a fuzzy clustering algorithm. The clustering result is a fuzzy partition matrix U . The undirected weighted membership network that represents this fuzzy clustering has as many data objects as nodes. Two additional nodes, nodes 9 and 10, represent the centers of the fuzzy clusters. Data object nodes are connected to the cluster center nodes using links having weights given by the entries of the matrix U . For clarity, the color and the thickness of the links between data object nodes and cluster center nodes is proportional to the degree of membership of each data object to the different clusters. Darker and thicker links indicate a higher degree of membership.

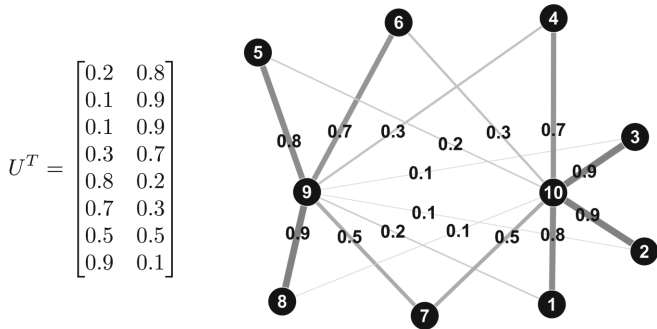


Fig. 1. Representing a fuzzy clustering result (left) as an undirected weighted membership network (right) using the proposed method.

Once a network membership is built, it can be visualized using a layout algorithm that automatically arranges the nodes and links in an aesthetically pleasing way. In particular, the OpenOrd algorithm [20] is used here for this task, since it is an algorithm suitable for drawing undirected weighted networks, and able to provide layouts for large-scale real-world networks wherein clusters can be better distinguished.

4 Results and Discussion

Experiments were performed to demonstrate the feasibility of the network-based approach for visualizing soft clusterings obtained from different datasets. As the purpose of this paper is mainly on presenting a clustering visualization method,

the traditional FCM algorithm was used to obtain fuzzy clusterings from the data sets. In particular, the number of clusters C and the fuzzification parameter m were manually set depending on the case.

4.1 Iris Flower Dataset

The proposed visualization method is first demonstrated using the widely known Iris flower data set. This multidimensional data set consist of four morphological measurements taken on 50 samples from three different species of Iris flowers: *setosa*, *versicolor*, and *virginica*. In particular, it is known that all the samples of *Iris setosa* are linearly separable from the samples of the other two species, while the latter species are not linearly separable from each other.

Three soft clusters for the Iris data set were initially obtained using FCM with $m = 2$. Figure 2A shows the membership network constructed based on the fuzzy partition matrix that represented the resulting clustering. Three groups of nodes representing the 3-cluster structure are visible in this network-based representation. Groups at the top of the figure represent two clusters that contain *virginica* and *versicolor* samples, while the remaining group represents a cluster with *setosa* samples. Each group of nodes has an anchor node which represents the cluster center, and, for clarity purposes, the size of the cluster center node is larger than the nodes representing data objects. Darker and thicker links indicate a higher degree of membership. As expected, intra-cluster links have a greater membership weight in comparison with extra-cluster links when clusters are well separated from each other. However, as the boundary between the clusters of *virginica* and *versicolor* samples is ambiguous, there exist nodes in this boundary belonging both clusters with a no well-defined degree of membership, and thus they are linked to both cluster center nodes with darker and thicker connections.

Since m is a crucial parameter of FCM, membership networks representing fuzzy clusterings obtained after running FCM on the Iris data set for different values of m were also constructed. Figure 2 also shows the behavior of the built membership networks under these conditions. As stated before, when m is close to 1 the entries of matrix U converge to 0 and 1, i.e., clusters become crispier. This is represented as well-separated groups whose nodes are compactly arranged around the cluster center node, except for those nodes which are in the boundary of two overlapping clusters (Fig. 2B). As m is increased, the cluster fuzziness also increased, and thus groups start to overlap each other due to the strength of the association between the nodes, and the clusters are not clearly defined (Fig. 2C and D).

4.2 Financial Datasets

The applicability of the proposed visualization method on large real-world data with unknown clustering structure is demonstrated using two datasets which describe the financial behavior of 18.583 customers of a local bank, during a particular time window. The first data set describes each customer using four

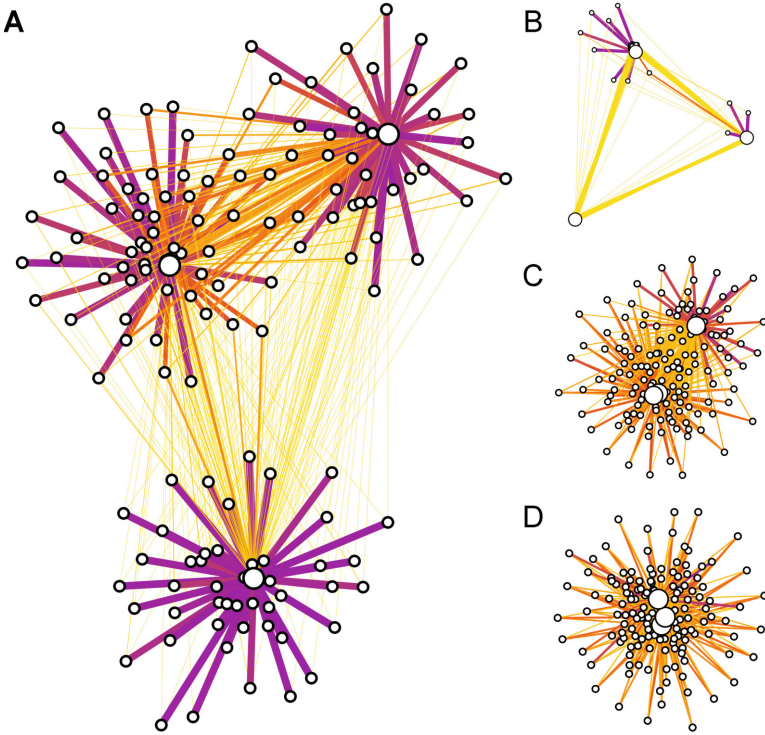


Fig. 2. Membership network representing a 3-cluster fuzzy structure discovered by the FCM algorithm on the Iris flower data set for different values of the fuzzification parameter m . (A) $m = 2$. (B) $m = 1.1$. (C) $m = 5$. (D) $m = 8$.

variables which characterize its transactions with other customers, and the second data set consists of ten variables describing the financial statements of each customer.

Since these datasets have an unknown underlying structure, they were arbitrarily clustered into ten groups of customers using the FCM algorithm with $m = 2$. Figures 3 and 4 show the resulting network-based representation for each soft clustering.

Data exploration based on organizing both datasets into ten arbitrarily soft clusters and using the network-based approach to visualize both clustering results demonstrate how different are their corresponding underlying structures. The clustering structure provided by FCM on the first data set consist of both large and predominant groups, and small groups (Fig. 3), while the same algorithm partitioned the second data set into regular size clusters (Fig. 4). Furthermore, larger groups discovered in the first data set are more closely connected in comparison with the smaller groups (Fig. 3), which could indicate that the former groups represent data objects lying in large regions with almost uniform data density, while the latter groups consist of data objects that are outliers. A

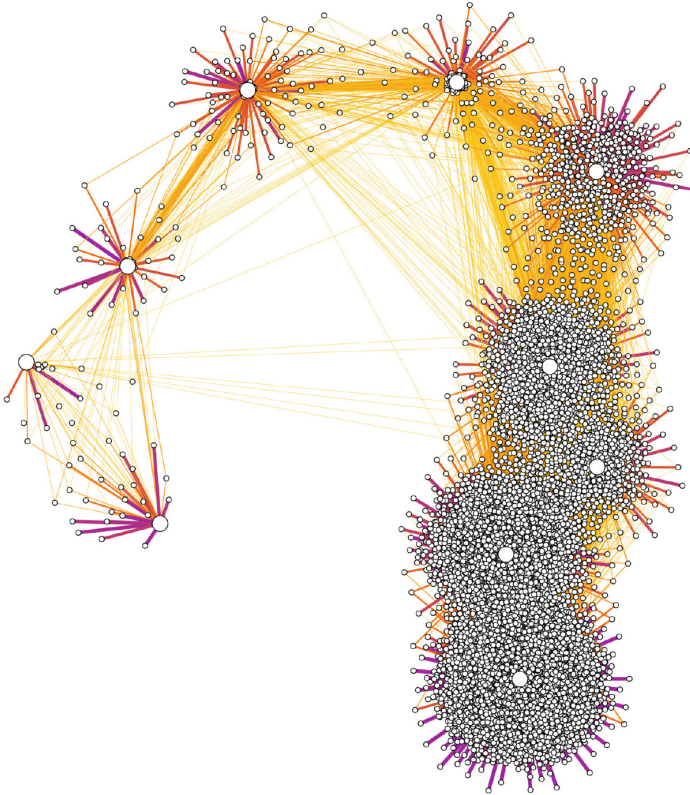


Fig. 3. Membership network representing a 10-cluster fuzzy structure discovered by the FCM algorithm on the first financial data set for $m = 2$.

subsequent examination of these group of outliers indicated that they correspond to clients who participate in numerous and large transactions, and thus these clients could be of interest for receiving special banking services related with financial transactions. On the other hand, the well-connected groups of nodes discovered in the second data set (Fig. 4) may also consist of data objects sharing the same feature like the larger groups in the first data set (Fig. 3). However, this result could also be interpreted as there is no evidence of the existence of natural groups in the second data set, and thus, the FCM algorithm only performed segmentation of the bank customers. Finally, the closeness between the three cluster center nodes in the middle of Fig. 4 could indicate that a cluster has been wrongly split into three clusters, as FCM necessarily must partition the data into a given number of clusters.

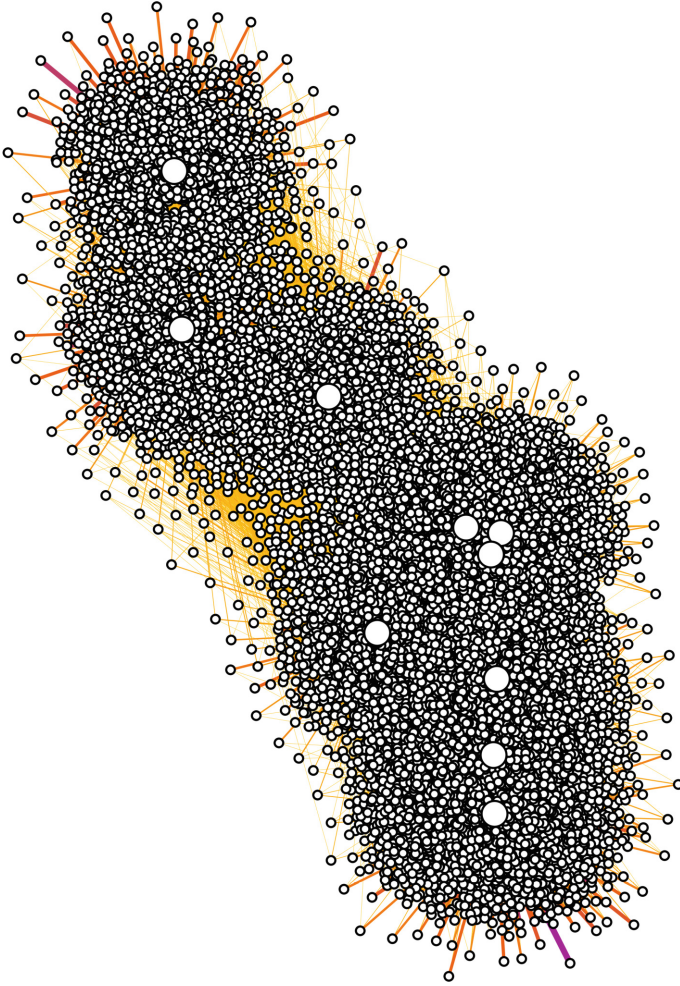


Fig. 4. Membership network representing a 10-cluster fuzzy structure discovered by the FCM algorithm on the second financial data set for $m = 2$.

5 Conclusions

In this work is presented a method to visualize clustering structures obtained from the application of fuzzy clustering algorithms on high-dimensional data. The proposed method uses undirected weighted membership networks to represent fuzzy partition matrices. This simple network-based method allows understanding visually how elements (data objects and clusters) involved in this kind of complex data clustering structures interact each other, without relying on a visualization of the input high-dimensional data itself. The proposed visualization method provides a means to represent near-crisp and very fuzzy clustering

structures and to explore large real-world data with unknown clustering structure. As future work, we plan to extend our analysis to other methods, besides FCM, that consider the notion of data objects belonging to multiple groups, such as the Gustafson-Kessel algorithm and the Possibilistic c -Means algorithm, among others.

Acknowledgements. This research work was supported by Centro de Excelencia y Apropiación en Big Data y Data Analytics -Alianza CAOBA- and Universidad EAFIT.

References

1. Abonyi, J., Babuska, R.: FUZZSAM - visualization of fuzzy clustering results by modified Sammon mapping. In: IEEE International Conference on Fuzzy Systems, vol. 1, pp. 365–370 (2004). <https://doi.org/10.1109/FUZZY.2004.1375750>
2. Bécavin, C., Benecke, A.: New dimensionality reduction methods for the representation of high dimensional ‘omics’ data. *Expert Rev. Mol. Diagn.* **11**(1), 27–34 (2011). <https://doi.org/10.1586/erm.10.95>
3. Berthold, M.R., Wiswedel, B., Patterson, D.E.: Interactive exploration of fuzzy clusters using neighborgrams. *Fuzzy Sets Syst.* **149**(1), 21–37 (2005). <https://doi.org/10.1016/j.fss.2004.07.009>
4. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: *Cluster Analysis*. Wiley, Hoboken (2011)
5. Feil, B., Balasko, B., Abonyi, J.: Visualization of fuzzy clusters by fuzzy Sammon mapping projection: application to the analysis of phase space trajectories. *Soft Comput.* **11**(5), 479–488 (2007). <https://doi.org/10.1007/s00500-006-0111-5>
6. Fortunato, S., Hric, D.: Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44 (2016). <https://doi.org/10.1016/j.physrep.2016.09.002>
7. Francalanci, C., Hussain, A.: Influence-based Twitter browsing with NavigTweet. *Inf. Syst.* **64**, 119–131 (2017). <https://doi.org/10.1016/j.is.2016.07.012>
8. Fruchterman, T.M., Reingold, E.M.: Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**(11), 1129–1164 (1991). <https://doi.org/10.1002/spe.4380211102>
9. Gajer, P., Goodrich, M.T., Kobourov, S.G.: A multi-dimensional approach to force-directed layouts of large graphs. *Comput. Geom.* **29**(1), 3–18 (2004). <https://doi.org/10.1016/j.comgeo.2004.03.014>
10. Gibson, H., Faith, J., Vickers, P.: A survey of two-dimensional graph layout techniques for information visualisation. *Inf. Vis.* **12**(3–4), 324–357 (2013). <https://doi.org/10.1177/1473871612455749>
11. Heberle, H., Carazzolle, M.F., Telles, G.P., Meirelles, G.V., Minghim, R.: Cell NetVis: a web tool for visualization of biological networks using force-directed layout constrained by cellular components. *BMC Bioinform.* **18**(S10), 395 (2017). <https://doi.org/10.1186/s12859-017-1787-5>
12. Höppner, F., Klawonn, F.: Visualising clusters in high-dimensional data sets by intersecting spheres. In: Proceedings of 2006 International Symposium on Evolving Fuzzy Systems, EFS 2006, vol. 2, no. 2, pp. 106–111 (2006). <https://doi.org/10.1109/ISEFS.2006.251180>
13. Hu, Y., Shi, L.: Visualizing large graphs. *Wiley Interdiscip. Rev. Comput. Stat.* **7**(2), 115–136 (2015). <https://doi.org/10.1002/wics.1343>

14. Ishida, Y., Itoh, T.: A force-directed visualization of conversation logs. In: Proceedings of Computer Graphics International Conference - CGI 2017, pp. 1–5. ACM Press, New York (2017). <https://doi.org/10.1145/3095140.3095156>
15. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **9**(6), 1–12 (2014). <https://doi.org/10.1371/journal.pone.0098679>
16. Leisch, F.: A toolbox for K-centroids cluster analysis. *Comput. Stat. Data Anal.* **51**(2), 526–544 (2006). <https://doi.org/10.1016/j.csd.2005.10.006>
17. Leisch, F.: Neighborhood graphs, stripes and shadow plots for cluster visualization. *Stat. Comput.* **20**(4), 457–469 (2010). <https://doi.org/10.1007/s11222-009-9137-8>
18. van der Maaten, L.: Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014). <http://jmlr.org/papers/v15/vandermaaten14a.html>
19. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008). <http://www.jmlr.org/papers/v9/vandermaaten08a.html>
20. Martin, S., Brown, W.M., Klavans, R., Boyack, K.W.: OpenOrd: an open-source toolbox for large graph layout. In: Proceedings of SPIE, p. 7868, January 2011. <https://doi.org/10.1117/12.871402>
21. Metsalu, T., Vilo, J.: ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res.* **43**(W1), W566–W570 (2015). <https://doi.org/10.1093/nar/gkv468>
22. Newman, M.E.J.: The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003). <https://doi.org/10.1137/S003614450342480>
23. Pison, G., Struyf, A., Rousseeuw, P.J.: Displaying a clustering with CLUSPLOT. *Comput. Stat. Data Anal.* **30**(4), 381–392 (1999). [https://doi.org/10.1016/S0167-9473\(98\)00102-9](https://doi.org/10.1016/S0167-9473(98)00102-9)
24. Sato-Ilic, M., Ilic, P.: Visualization of fuzzy clustering result in metric space. *Proc. Comput. Sci.* **96**, 1666–1675 (2016). <https://doi.org/10.1016/j.procs.2016.08.214>
25. Serra, A., Galdi, P., Tagliaferri, R.: Machine learning for bioinformatics and neuroimaging. *Wiley Interdisc. Rev.: Data Min. Knowl. Discov.* **8**(5), 1–33 (2018). <https://doi.org/10.1002/widm.1248>
26. Sharko, J., Grinstein, G.: Visualizing fuzzy clusters using RadViz. In: Proceedings of International Conference Information Visualisation, pp. 307–316 (2009). <https://doi.org/10.1109/IV.2009.74>
27. Wang, K.J., Yan, X.H., Chen, L.F.: Geometric double-entity model for recognizing far-near relations of clusters. *Sci. China Inf. Sci.* **54**(10), 2040–2050 (2011). <https://doi.org/10.1007/s11432-011-4386-5>
28. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. *Fuzzy Sets Syst.* **158**(19), 2095–2117 (2007). <https://doi.org/10.1016/j.fss.2007.03.004>
29. Xu, R., Wunsch, D.: Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**(3), 645–678 (2005). <https://doi.org/10.1109/TNN.2005.845141>
30. Xu, R., Wunsch, D.C.: Clustering algorithms in biomedical research: a review. *IEEE Rev. Biomed. Eng.* **3**, 120–54 (2010). <https://doi.org/10.1109/RBME.2010.2083647>
31. Zhou, F., et al.: A radviz-based visualization for understanding fuzzy clustering results. In: Proceedings of 10th International Symposium on Visual Information Communication and Interaction, pp. 9–15. ACM, New York (2017). <https://doi.org/10.1145/3105971.3105980>