Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Statistical Variability Networks: Transforming Statistical Summaries into Network Representations

Kristyna Plesnikova
VSB - Technical University of Ostrava
Eliska Ochodkova
eliska.ochodkova@vsb.cz

VSB - Technical University of Ostrava
Eva Krieagova
Palacký University Olomouc
Milos Kudelka
VSB - Technical University of Ostrava

Research Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11254023/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Statistical Variability Networks: Transforming Statistical Summaries into Network Representations

Kristyna Plesnikova[1], Eliska Ochodkova[1*], Eva Kriegova[2], Milos Kudelka[1]

[1*]Department of Computer Science, VSB - Technical University Ostrava, 17. listopadu 15/2172, Ostrava, 70800, Czech Republic.
[2]Department of Immunology, Palacky University Olomouc and University Hospital Olomouc, Krizkovskeho, Olomouc, 779 00, Czech Republic.

*Corresponding author(s). E-mail(s): eliska.ochodkova@vsb.cz;
Contributing authors: kristyna.plesnikova@vsb.cz;
eva.kriegova@email.cz; milos.kudelka@vsb.cz;

Cluster analysis is widely used to identify groups of similar observations, but conventional cluster representations provide limited information about the statistical variability associated with these groups. This article introduces Statistical Variability Networks (hereafter referred to as STAT Networks), a network-based framework for representing statistical descriptions of clustered data. For each cluster, a representative centroid and additional variability vectors are constructed by replacing individual centroid values with lower and upper statistical bounds. These vectors are then represented as nodes in a weighted k-nearest-neighbor network, where connections reflect statistical proximity between the resulting vectors. The framework supports several statistical representations, including 95% confidence intervals (CI), one and two standard deviations (1SD and 2SD), interquartile range (IQR), and interquartile whisker range (IWR). The proposed framework is demonstrated on the E. coli, Iris, and Yeast datasets to examine how different statistical representations affect the resulting network structure and relationships between cluster profiles. The results show that the number and structure of inter-cluster connections vary across the statistical representations, with broader variability ranges generally revealing additional statistical

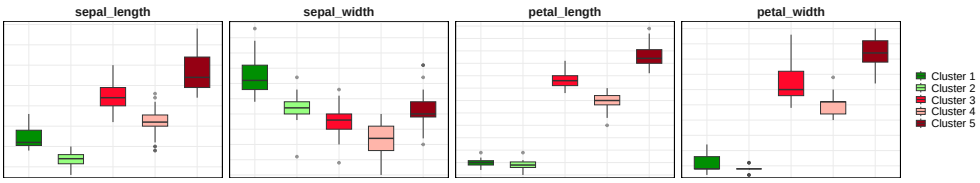


Fig. 1 Boxplots showing the distributions of four attributes across five clusters in the Iris dataset. The figure illustrates how cluster variability and the degree of separation between clusters can be examined for individual attributes.

proximity between cluster profiles. The results also demonstrate cases where cluster profiles that appear separated in the original network become connected when their statistical variability is represented. STAT Networks therefore provide a complementary representation for exploring relationships between cluster profiles and their statistical variability alongside conventional statistical visualizations.

1 Introduction

Clustering is a fundamental technique in exploratory data analysis that identifies groups of similar observations within multidimensional datasets. It is widely used in fields such as biology, medicine, environmental sciences, and social sciences, where large collections of measured variables are common. While clustering algorithms effectively partition data into meaningful groups, interpreting the resulting clusters remains a challenging task. Beyond cluster membership, researchers often seek to understand the internal structure of clusters, identify characteristic attributes, and assess the variability of these attributes [1, 2].

Statistical visualizations such as boxplots or violin plots are commonly used to summarize attribute distributions within clusters. Although these methods provide valuable information about individual variables, they treat attributes independently and can become difficult to interpret when many variables must be considered simultaneously. Existing approaches to cluster visualization also address the challenge of representing high-dimensional cluster structure [3]. However, conventional statistical visualizations provide limited insight into how the statistical profiles of different clusters relate to one another across multiple attributes.

Figure 1 illustrates a conventional representation of cluster variability using boxplots. The distributions of individual attributes can be compared across clusters, making differences in their statistical characteristics and potential overlap visible. However, when several clusters and attributes are considered simultaneously, the relationships between their statistical profiles must be inferred from the individual attribute distributions. It can therefore be difficult to assess the overall statistical relationship between two clusters and identify which aspects of their variability contribute to this relationship.

In particular, researchers may need to understand the statistical profile represented by each cluster and how individual observations are distributed around its representative centroid. For example, in a medical application, a cluster may represent a group

of patients whose centroid describes their typical profile. However, the centroid alone does not describe the variability within the group. It is therefore important to examine how the distributions of different clusters relate to one another, including whether they are clearly separated or to what extent they overlap. In such applications, the size of a cluster is not necessarily the most important characteristic. A relatively small cluster may still be relevant if its statistical profile is clearly separated from those of the other clusters.

This motivates a representation that captures both the variability within individual clusters and the relationships between their statistical descriptions. While conventional statistical summaries and boxplots can reveal the distributions of individual attributes and indicate potential overlap, a relational representation can provide an additional perspective by making relationships between cluster profiles visible at the network level.

Network representations offer a way to visualize complex relationships by representing objects as nodes and their similarities as edges. They have been applied in many areas of data analysis, including the visualization and exploration of relationships in high-dimensional and biological data [4, 5]. Network structure can reveal relationships that may be difficult to identify using conventional representations. However, their use for representing statistical summaries of cluster variability remains limited.

In this context, a network representation does not aim to replace conventional statistical summaries or visualizations, but rather to complement them with an additional relational view of cluster variability. While statistical tables and distribution plots describe individual attributes, the network provides a simultaneous view of the statistical relationships between clusters and their degree of separation.

In this article, we propose a general framework for transforming statistical descriptions of clusters into network representations. For each cluster, a representative centroid is first computed, after which additional vectors are generated according to selected statistical measures describing attribute variability. The resulting vectors are connected according to their Euclidean distances using a k-nearest neighbor graph, where k is set to twice the number of analyzed attributes. The framework can be applied as a post-processing step to clustering results produced by different clustering algorithms.

The proposed framework extends preliminary work presented at the 6th International Symposium on Complex Systems (ISCS 2026) under the title "CI and IQR Networks: A Network-Based Visualization of Cluster Variability." In this extended study, the framework is evaluated on benchmark datasets of varying complexity and expanded to examine how different statistical representations affect the resulting network structure and reveal relationships between cluster profiles and their statistical variability. The resulting networks complement traditional statistical visualizations by providing a relational view of cluster variability.

2 Related Work

Visual interpretation of clustering results is an important step in exploratory data analysis, allowing researchers to understand cluster characteristics, identify representative attributes, and communicate results to domain experts. Numerous visualization techniques have been proposed for this purpose, ranging from traditional scatter plots and heatmaps to interactive visual analytics and graph-based representations. Interactive systems such as the Hierarchical Clustering Explorer introduced by Seo and Shneiderman [2] demonstrated the value of combining clustering with coordinated visualizations, enabling users to explore hierarchical structures and attribute distributions simultaneously.

More recent research has focused on reducing the visual complexity associated with large and high-dimensional datasets. Clusterplot, introduced by Malkai et al. [3], represents each cluster as a two-dimensional graphical object that captures its geometric characteristics, overlap, and relationships with neighboring clusters. By emphasizing cluster-level properties instead of individual samples, the method reduces visual clutter while supporting the interpretation of cluster organization and separation. Similarly, recent surveys on cluster visualization emphasize the growing importance of techniques that summarize cluster structure while preserving the interpretability of multidimensional data [1].

Graph-based visualizations have also become increasingly popular for exploring complex datasets. Rather than relying solely on geometric projections, graph representations model relationships directly through network topology, enabling the analysis of structural properties such as connectivity and community organization. An example is TooManyCells, proposed by Schwartz et al. [4], which constructs hierarchical graph representations for the exploration of single-cell transcriptomic data. Their approach demonstrates how network structures can facilitate the interpretation of relationships between clusters while remaining scalable to very large datasets.

Although these methods significantly improve the visualization of clustered data, their primary objective is to represent observations, cluster geometry, or hierarchical relationships. Our approach addresses a complementary problem by representing statistical descriptions of cluster variability as network structures. The proposed framework, referred to as STAT Networks, transforms statistical summaries into centroid-based networks that enable the analysis of intra-cluster variability through network topology. Rather than replacing traditional statistical visualizations, these networks provide an additional perspective on the relationships among attribute variations within individual clusters.

3 Methodology

3.1 Overview

The proposed framework assumes that the input data have already been partitioned into clusters. In the experiments presented in this study, the original vector-based data are first transformed into a similarity network using the LRNet algorithm [6], which constructs edges based on local representativeness. Clusters are then identified by

applying the Louvain community detection algorithm [7]. The resulting partitioning provides the cluster structure used as input to the proposed STAT Network framework. The framework itself is independent of the clustering algorithm and can therefore be applied as a post-processing step to any clustering result. For each cluster, a representative centroid is computed. Additional vectors are then generated by systematically replacing individual centroid values with selected statistical descriptors of attribute variability. These vectors become the nodes of the resulting network. Regardless of the statistical representation used, the network is constructed from pairwise Euclidean distances between the generated vectors using a k-nearest-neighbor graph, where k is set to twice the number of analyzed attributes. Under ideal separation, the statistical vectors associated with each cluster form a clique, while connections between vectors belonging to different clusters represent relative statistical proximity within the set of generated vectors. The same network-construction procedure is used for all statistical representations considered in the framework, enabling direct comparison of the resulting STAT Networks.

3.2 Generation of Variability Vectors

Let

$$X = \{x_1, x_2, \ldots, x_n\}$$

denote the set of observations belonging to a particular cluster, where each observation is described by m attributes.

For every cluster, a representative centroid vector

$$\mathbf{c} = (c_1, c_2, \ldots, c_m)$$

is computed. The centroid represents the typical attribute configuration of the cluster and is defined according to the selected statistical representation.

To characterize attribute variability, two additional vectors are generated for each attribute by replacing the corresponding centroid value with its lower and upper statistical bounds while keeping all remaining attributes unchanged. The definition of these bounds depends on the selected network type.

Consequently, each attribute produces two variability vectors, resulting in

$$2m + 1$$

vectors for every cluster, including the centroid itself. Each variability vector therefore represents a controlled modification of one attribute of the centroid while preserving the values of all remaining attributes.

3.3 Statistical Variability Networks

Different statistical summaries emphasize different characteristics of attribute variability. The proposed framework therefore supports multiple network variants that differ in the statistical descriptors used to construct the centroid and variability vectors. The overall network construction remains identical across all variants.

Table 1 Statistical variability network variants supported by the proposed framework.

Network	Centroid	Variability bounds	Distribution
CI	Mean	95% confidence interval	Normal
1SD	Mean	$\mu \pm \sigma$	Normal
2SD	Mean	$\mu \pm 2\sigma$	Normal
IQR	Median	Q_1, Q_3	Distribution-free
IWR	Median	Tukey whiskers	Distribution-free

The labels of the variability vectors in the resulting networks follow a consistent notation. The network type is given first, followed by the cluster number, the direction of the statistical bound, and the corresponding attribute. For example, *1SD 2+ alm* denotes the upper one-standard-deviation vector of Cluster 2 for the *ALM* attribute, while *1SD 2- alm* denotes the corresponding lower-bound vector.

The CI, 1SD, and 2SD networks use the arithmetic mean as the centroid and describe variability relative to the mean. These representations are based on standard deviation or confidence-interval calculations and are therefore sensitive to the distributional characteristics of the data, particularly skewness and outliers. The CI, 1SD, and 2SD networks differ in the statistical bounds used to generate the variability vectors.

For skewed distributions or datasets containing outliers, more robust alternatives can be employed. IQR networks use the median together with the first and third quartiles, while IWR networks extend these bounds to the lower and upper whiskers defined by Tukey's rule, thereby representing the typical range of non-outlier observations. Both IQR and IWR provide distribution-free descriptions of variability.

3.4 Network Construction

Each representative and variability vector is represented as a node in the resulting network. The Euclidean distance is calculated between all generated vectors to quantify the distance between their statistical profiles. Based on these distances, a weighted k-nearest-neighbor graph is constructed. For each edge, the Euclidean distance between the connected vectors is transformed into an edge weight by dividing the distance by the maximum Euclidean distance and subtracting the resulting ratio from 1. Consequently, smaller Euclidean distances result in stronger edge weights, indicating greater statistical proximity between the connected vectors. The resulting edge weights range from 0 to 1, enabling comparison of edge strengths across different STAT Networks.

Instead of connecting every node to all other nodes, each node is connected only to its k nearest neighbors. The value of k is determined as

$$k = 2m,$$

where m is the number of attributes in the dataset. This choice is motivated by the construction of the variability vectors, as each attribute generates two statistical vectors. Consequently, in the ideal case of perfectly separated clusters, each statistical

cluster forms a clique, while only a limited number of edges connect different statistical clusters. This preserves the local structure of individual clusters while retaining relationships between neighboring clusters.

The resulting STAT Network can be visualized using standard graph layouts and further analyzed using common network analysis and community detection techniques. Because the graph construction is independent of the selected statistical representation, all STAT Network variants are generated using the same procedure and can therefore be directly compared.

3.5 Interpretation of Statistical Networks

STAT Networks provide a representation of the statistical relationships among clusters. Rather than visualizing individual observations, the network represents statistical descriptions of clusters and the statistical proximity between them. The topology of the network therefore reflects how closely the statistical characteristics of different cluster profiles are related.

Clique Structure

The network construction is designed so that, in the ideal case of perfectly separated clusters, each statistical cluster forms a clique. Such a structure indicates that the statistical vectors belonging to the same cluster are mutually connected. Figure 2 illustrates an example of an ideal clique, where every node is connected to all remaining nodes within the cluster. In this situation, all nodes have the same ordinary degree, making degree alone less informative for distinguishing nodes within the clique.

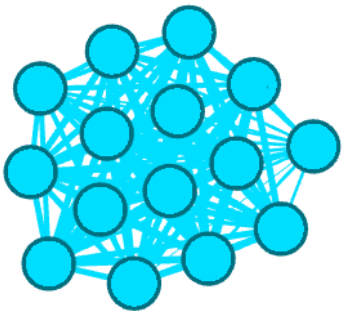


Fig. 2 Example of an ideal clique representing a perfectly separated statistical cluster.

Strong Inter-Cluster Connections

While clique structures describe the connectivity of statistical vectors within a cluster, edges between different clusters reveal statistical proximity between their corresponding vectors. Multiple strong connections indicate that several statistical vectors from

two clusters are closely related. Figure 3 shows an example of two statistical clusters connected by several strong edges, indicating greater statistical proximity between their statistical profiles.

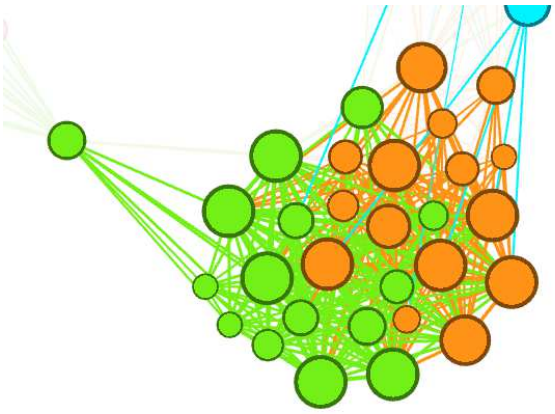


Fig. 3 Strong inter-cluster connections indicate statistical proximity between neighboring clusters.

Weak Inter-Cluster Connections

In contrast, statistical clusters connected by only a few weak edges exhibit less statistical proximity between their corresponding vectors. Such structures indicate greater separation between the statistical profiles represented by the connected vectors. Figure 4 presents two statistical clusters connected by only a small number of weak edges. The single purple node connected to the other cluster represents a variability vector associated with one specific attribute, indicating that the variability of this attribute is the component through which the ideal separation between the two clusters is disrupted. The weak connection therefore indicates limited statistical proximity between the corresponding statistical profiles.

Weighted Degree

Since all nodes within an ideal clique have identical ordinary degree, the weighted degree provides additional information about the strength of their connections. Unlike the ordinary degree, the weighted degree combines the number of connections and their strengths, where the edge weights reflect the statistical proximity between connected vectors. Consequently, nodes with higher weighted degrees have stronger overall connections to their local neighborhood than nodes with lower weighted degrees. Figure 5 highlights an example of a node with a high weighted degree.

Nodes with High Degree

Although the ordinary degree is identical within an ideal clique, real STAT Networks may contain nodes with higher degree due to additional connections outside their

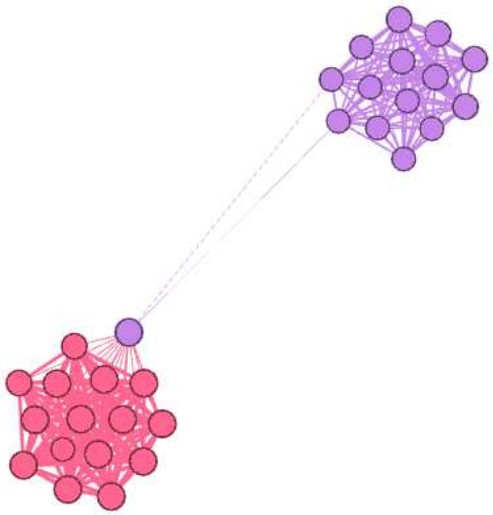


Fig. 4 Weak inter-cluster connections indicate limited statistical proximity between cluster profiles.

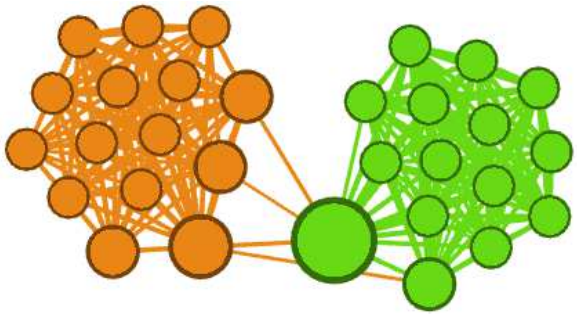


Fig. 5 Example of a node with a high weighted degree.

local cluster. Such nodes may occupy structurally important positions within the network and can indicate regions where statistical relationships extend across cluster boundaries. Figure 6 presents an example of nodes with a high degree.

4 Results and Analysis

The proposed STAT Networks are evaluated on the E. coli, Iris, and Yeast datasets to examine how different statistical representations affect the relationships between cluster profiles and their attribute variability. The analysis considers both the overall structure of the resulting networks and the specific inter-cluster connections revealed by the statistical variability vectors. The clustering solutions used in the experiments are not intended to represent the only possible partitions of the datasets. They were selected as illustrative examples for demonstrating the construction and interpretation

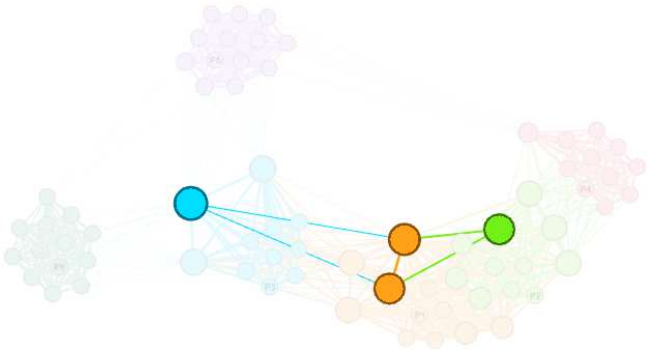


Fig. 6 Example of nodes with a high degree.

of the proposed STAT Networks. The STAT Networks are intended as an initial visual support for assessing a given clustering result. They provide a complementary view of the relationships between the resulting cluster profiles and their statistical variability, which can help identify patterns that may warrant further examination using conventional statistical and clustering analysis. The network representations provide a global view of relationships between cluster profiles, while the corresponding boxplots and Origami plots [8] provide complementary detail on the attribute distributions and mean profiles underlying these relationships.

The following experiments illustrate these relationships across datasets with different characteristics. The E. coli dataset demonstrates how the relationships between cluster profiles change across the different statistical representations. The Iris dataset provides a relatively compact setting for comparing the network relationships with the corresponding attribute distributions. Finally, the Yeast dataset provides a more complex setting, with seven clusters and eight attributes, in which the different statistical representations produce more varied and interconnected network structures.

4.1 E. coli Dataset

Figure 7(a) presents the original network constructed from the original vector data, with communities detected in this network, together with the STAT Networks obtained using the IQR, IWR, 1SD, 2SD, and CI representations. The original network provides an initial view of the cluster structure, while the STAT Networks show how the relationships between the six cluster profiles change when attribute variability is represented. From a global perspective, Clusters 1 (orange) and 2 (green) form a particularly close pair, while the remaining clusters form a more heterogeneous part of the network. The IQR network provides the clearest separation between the cluster profiles, whereas the IWR, 1SD, 2SD, and CI networks reveal additional inter-cluster relationships.

The close relationship between Clusters 1 and 2 is particularly evident when the IQR and IWR networks are compared. In the IQR network in Figure 7(a), the two

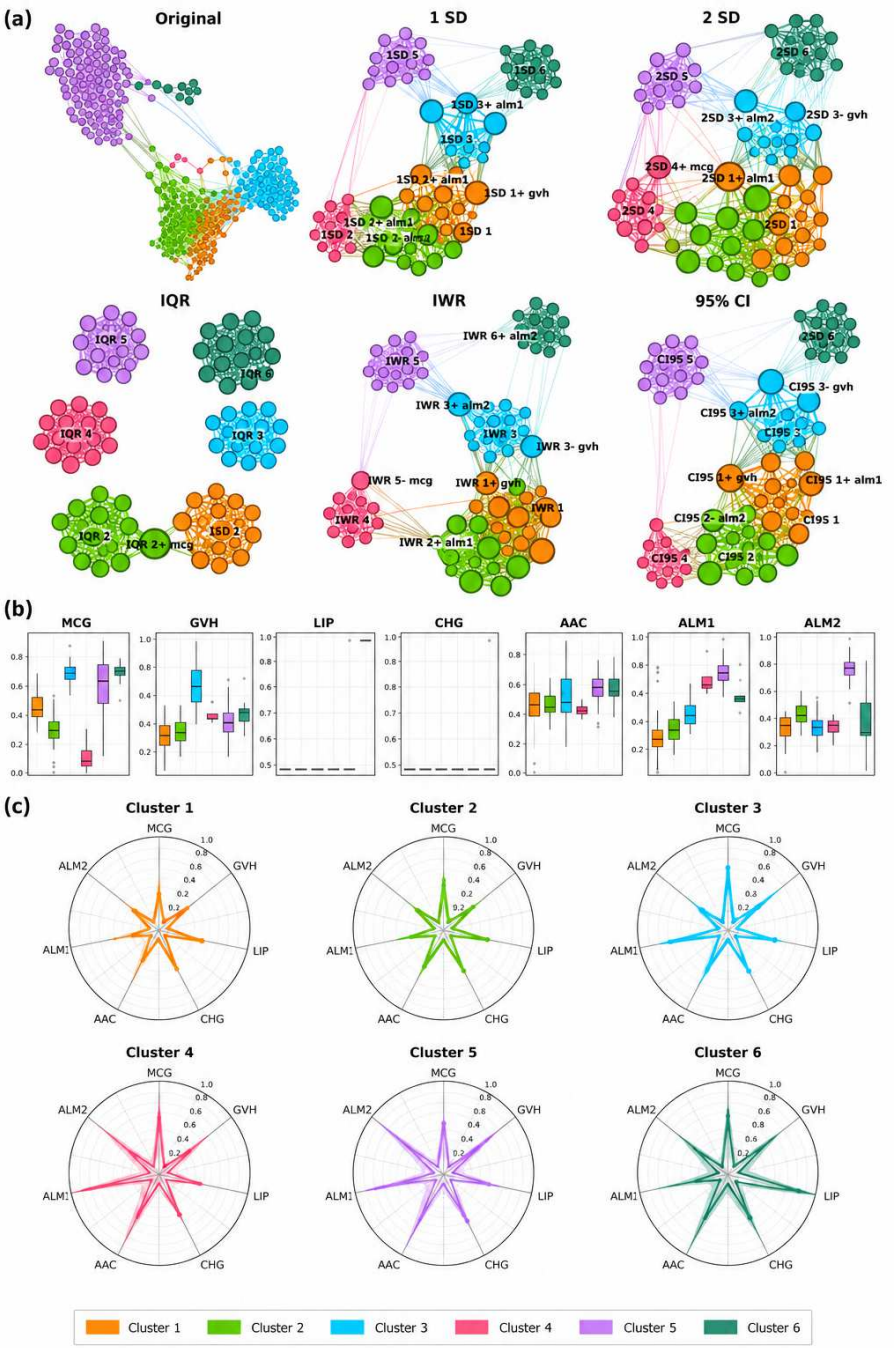


Fig. 7 Visualization of the *E. coli* data. (a) Network representations for the original data and the 1SD, 2SD, IQR, IWR, and 95% CI variants. (b) Distributions of MCG, GVH, LIP, CHG, AAC, ALM1, and ALM2 attributes across the six clusters. (c) Cluster-specific Origami plots. Colors denote Clusters 1–6 consistently across all panels.

clusters are connected through only a single variability component, involving the upper-quartile representation of the attribute MCG (*IQR 2+ mcg*), while the remainder of their statistical representations remains separated. The boxplots in Figure 7(b) shows that the MCG attribute distributions of Clusters 1 and 2 occupy relatively close ranges compared with several other cluster pairs. Their Origami plots in Figure 7(c) also show similar mean-profile patterns across the attributes, although their absolute magnitudes differ. The single connection in the IQR network therefore identifies MCG attribute as a specific component through which the two cluster profiles become statistically proximate.

When the broader Tukey-whisker range is considered, the relationship between Clusters 1 and 2 becomes more densely connected. In the IWR network in Figure 7(a), multiple connections occur between the two clusters, with attribute ALM1 repeatedly involved in the corresponding variability vectors. The change from the single IQR connection to the denser IWR connectivity indicates that additional parts of the two cluster profiles become sufficiently close when a broader portion of their non-outlier ranges is considered. This is consistent with the boxplots in Figure 7(b), where the broader ranges of the distributions extend towards one another. The Origami plots in Figure 7(c) provide a complementary profile-level view of the relationship, showing related patterns across the attributes despite differences in their absolute values.

A similar increase in connectivity can be observed when the 1SD and 2SD networks are compared. In the 1SD network in Figure 7(a), several inter-cluster connections involve attribute ALM1, including *1SD 1+ alm1*, *1SD 2+ alm1*, and *1SD 3+ alm1*. These connections indicate that particular standard-deviation bounds of the ALM1 distributions bring the corresponding statistical vectors closer together. The 2SD network contains additional connections involving ALM1, ALM2, GVH, and MCG attributes, including *2SD 1+ alm1*, *2SD 3- gvh*, *2SD 3+ alm2*, and *2SD 4+ mcg*. The greater connectivity in the 2SD network therefore shows that extending the represented range can expose additional statistical proximity between the cluster profiles.

The broader connectivity observed in the 1SD and 2SD networks also highlights relationships involving clusters that appear relatively separated in the global network structure. In particular, Cluster 6 (dark green) forms a comparatively distinct group, but connections with Cluster 3 are visible in the IWR, 1SD, 2SD, and CI networks in Figure 7(a). These connections show that the apparent separation of Cluster 6 does not occur uniformly across the different statistical representations. Instead, particular variability vectors become sufficiently close to corresponding vectors from Cluster 3 under these representations. The boxplots in Figure 7(b) and the Origami plots in Figure 7(c) provide complementary views of the attribute distributions and mean profiles associated with these connections.

In contrast to the relationships observed for Cluster 6 (dark green), Cluster 4 (pink) forms a relatively small and dispersed group in the network structure shown in Figure 7(a). Its dispersed distribution is also reflected in the boxplots in Figure 7(b) and in the corresponding Origami plot in Figure 7(c). Since Cluster 4 contains relatively few observations, its confidence intervals should be interpreted in relation to its cluster size, as the width of a confidence interval is affected by the number of

observations. Nevertheless, Cluster 4 remains relatively well separated in the network structure, despite the dispersed nature of its profile.

Overall, the E. coli networks show that the relationships between the six cluster profiles change across the statistical representations. The IQR-IWR comparison shows how extending the represented range can reveal additional connections between Clusters 1 and 2, while the 1SD-2SD comparison shows a similar increase in connectivity. The connections involving Cluster 6 (dark green) and Cluster 3 (blue) further illustrate that separation observed in one representation may not be maintained across all statistical representations. The boxplots and Origami plots in Figure 7(b)-(c) provide complementary detail on the distributions and mean profiles underlying these network relationships.

4.2 Iris Dataset

The Iris dataset provides a relatively compact setting for examining the relationships between five cluster profiles. The original network in Figure 8(a) provides an initial view of the cluster structure, with node sizes representing the average attribute values. From a global perspective, the network suggests two broader groups, but the differences in node size indicate that the five clusters are not equivalent in their average profiles. These differences provide additional support for retaining the five-cluster structure within the two broader groups. The STAT Networks in Figure 8(a) then show how the separation between these profiles changes when different measures of variability are considered. The IQR and CI networks maintain relatively clear separation, whereas the IWR, 1SD, and 2SD networks reveal more inter-cluster connections. The boxplots in Figure 8(b) show clear differences between the clusters, particularly in petal length and petal width, while greater overlap remains among some of the sepal attributes.

The relationship between Clusters 1 (light green) and 2 (dark green) provides a clear example of the effect of the variability range considered. In the IQR network in Figure 8(a), the two clusters remain separated, indicating that their central 50% ranges do not produce inter-cluster connections. In contrast, the IWR network in Figure 8(a) contains several connections between the two clusters, involving *IWR 1- sepal_width*, *IWR 2+ sepal_width*, and *IWR 2+ sepal_length*. Extending the represented range from the interquartile range to the broader Tukey-whisker range therefore reveals additional statistical proximity between the two cluster profiles. This is consistent with the boxplots in Figure 8(b), where the interquartile ranges of Clusters 1 and 2 remain separated while their broader ranges extend towards one another. The Origami plots in Figure 8(c) show similar profile proportions across the attributes, although the overall magnitudes of the two profiles differ. The network therefore identifies specific variability components through which the two profiles become closer when a broader range is considered.

The same effect of extending the represented variability range can be observed for Clusters 3 (red), 4 (pink), and 5 (dark red). In the 1SD network in Figure 8(a), these clusters are connected through a limited number of variability components, including *1SD 3+ petal_length*, *1SD 4+ petal_length*, and *1SD 5- petal_length*. The 2SD network in Figure 8(a) contains additional connections, including *2SD 3- petal_length*, *2SD 3+ petal_length*, and *2SD 5- petal_length*. The 2SD network therefore reveals broader

statistical proximity among these clusters than the 1SD network. In contrast, the CI network in Figure 8(a) shows no inter-cluster connections among Clusters 3, 4, and 5. The corresponding boxplots in Figure 8(b) and Origami plots in Figure 8(c) provide complementary views of the distributions and mean profiles underlying these relationships.

The distributional characteristics of the Iris attributes should also be considered when interpreting the 1SD, 2SD, and CI networks. The normality analysis did not support normal distributions for the attributes. Consequently, these representations should be interpreted with caution, whereas the IQR and IWR networks provide distribution-free descriptions of variability.

Overall, the Iris networks show that the relationships between the five cluster profiles vary across the statistical representations. The IQR–IWR comparison highlights how extending the represented range reveals additional proximity between Clusters 1 and 2, while the 1SD–2SD comparison shows a similar increase in connectivity among Clusters 3, 4, and 5. The boxplots and Origami plots provide complementary detail on the distributions and mean profiles underlying these network relationships.

4.3 Yeast Dataset

The Yeast dataset provides a more complex setting for examining the relationships between seven cluster profiles described by eight attributes. From a global perspective, the network structures in Figure 9 suggest two broader parts of the dataset, with different degrees of separation among the cluster profiles. Cluster 4 (pink) forms a small and dispersed group, while the remaining clusters exhibit varying degrees of connectivity across the network representations. The IQR network shows the clearest overall separation, whereas the IWR, 1SD, 2SD, and CI networks reveal additional inter-cluster connections.

The contrast between the original network and the STAT Networks is particularly visible for Clusters 1 (green) and 3 (blue). Although the two clusters appear relatively separated in the original network, this separation is less pronounced in the STAT Networks, where connections between their variability vectors can be observed. The apparent separation of the cluster profiles in the original network therefore does not necessarily extend to their represented variability. The boxplots and Origami plots in Figure 9 provide complementary views of the distributions and mean profiles underlying this relationship.

A contrasting pattern is observed for Cluster 4 (pink). It forms a small and dispersed group, as reflected in the boxplots and Origami plot in Figure 9. Despite the dispersed nature of its profile, Cluster 4 is fully separated from the other clusters in both the IQR and IWR networks. This separation is maintained in the 1SD, 2SD, and CI networks. It is particularly evident in the 2SD network, where all other clusters are interconnected while Cluster 4 remains disconnected from the rest. The separation of Cluster 4 is therefore maintained even when the represented variability is extended to a broader range, highlighting the distinct position of this cluster despite its relatively small and dispersed distribution.

Cluster 7 (dark green) shows an almost opposite pattern. In the original network in Figure 9, Cluster 7 appears relatively well separated from the other clusters, whereas

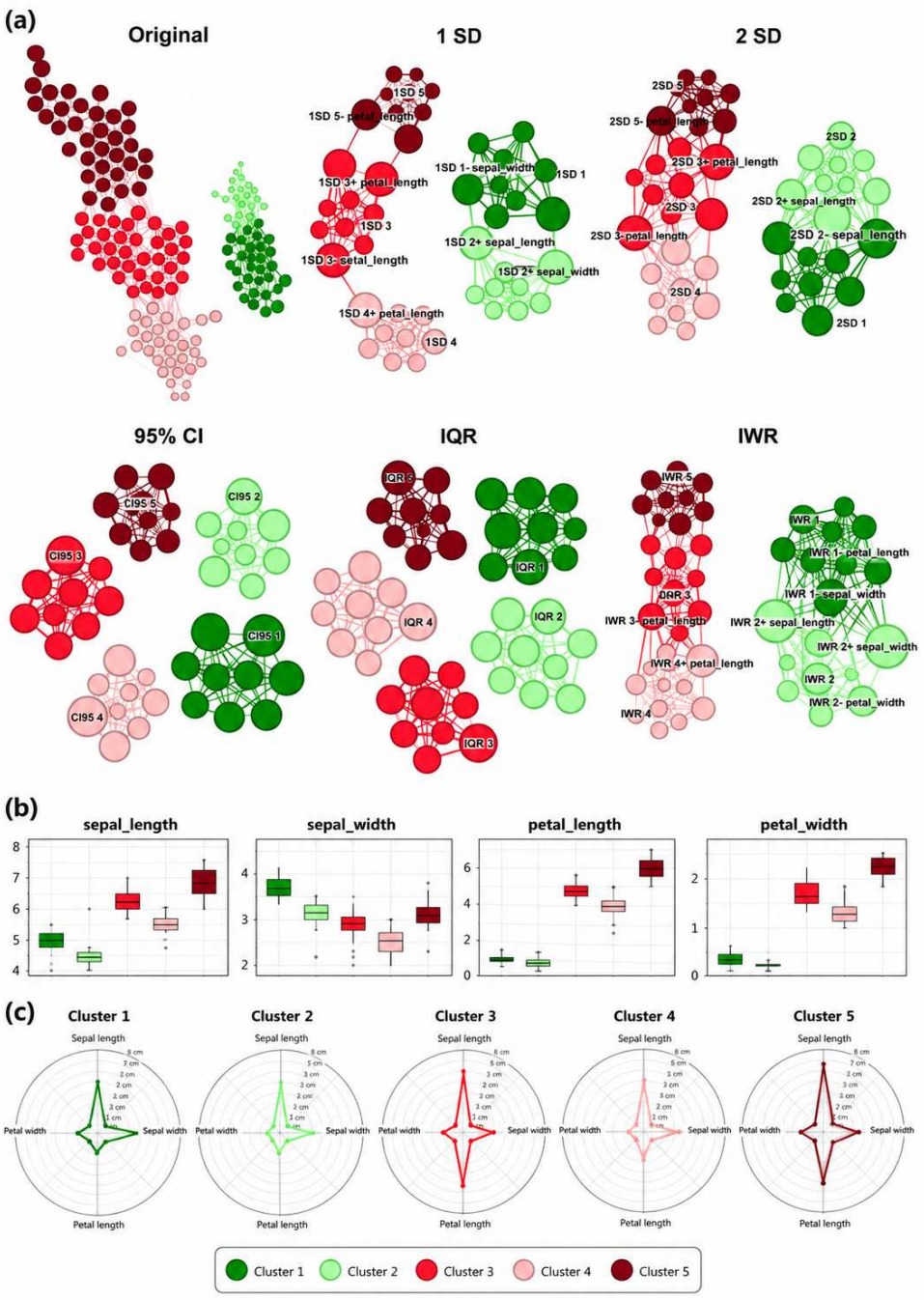


Fig. 8 Visualization of the *Iris* data. (a) Network representations for the original data and the 1SD, 2SD, 95% CI, IQR, and IWR variants. (b) Distributions of the four attributes sepal_length, sepal_width, petal_length, and petal_width across the five clusters. (c) Cluster-specific Origami plots. Colors consistently denote Clusters 1–5 throughout the figure.

this separation becomes less pronounced in the STAT Networks, where Cluster 7 forms connections with other cluster profiles across several representations. The contrast with Cluster 4 (pink) is particularly notable: although Cluster 4 has a relatively dispersed profile, its variability remains separated from the other clusters, whereas Cluster 7 appears more separated in the original network but becomes more connected when the variability of the profiles is taken into account. This comparison illustrates how the original network and the STAT Networks capture different aspects of cluster separation.

A further pattern is the persistent connection between Clusters 1 (green) and 2 (orange) across the statistical representations. Although the strength and structure of the connections vary between networks, these two clusters remain connected in almost all of the STAT Networks, with ALM repeatedly involved in their connections. This indicates that their statistical proximity is not limited to a single representation or isolated attribute, but extends to several parts of their statistical profiles. In a dataset with seven clusters and eight attributes, this type of relationship is difficult to assess from individual boxplots alone, whereas the network representation makes the recurring connections between the cluster profiles visible at the global level.

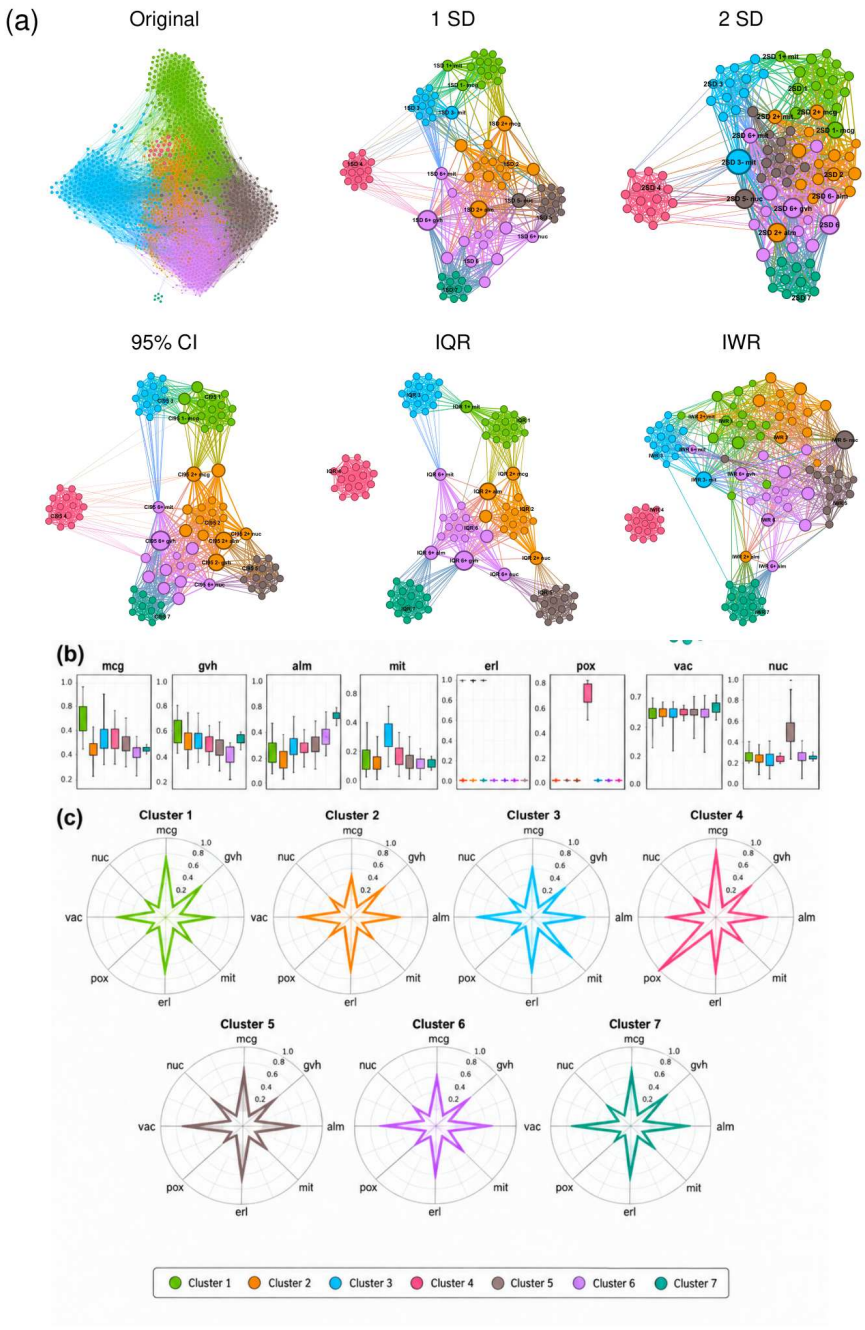


Fig. 9 Visualization of the *Yeast* data: (a) network representations for the original data and the 1SD, 2SD, 95% CI, IQR, and IWR variants; (b) boxplots of MCG, GVH, ALM, MIT, ERL, POX, VAC, and NUC attributes across the seven clusters; (c) cluster-specific Origami plots. Colors denote Clusters 1–7 consistently across all panels.

5 Discussion

The results demonstrate that STAT Networks can expose relationships between cluster profiles that are not fully apparent from the original network alone. The comparison between the original network and the STAT Networks, particularly in the Yeast dataset, shows that separation of cluster profiles does not necessarily correspond to separation of their statistical variability.

The choice of statistical representation also affects the resulting network structure. The IQR, IWR, 1SD, 2SD, and CI representations emphasize different parts of the underlying distributions, leading to different patterns of inter-cluster connections. IQR and IWR provide distribution-free alternatives to the mean-based representations, while 1SD, 2SD, and CI describe variability relative to the mean.

These results indicate that the STAT Networks should be interpreted together with the underlying distributions and mean profiles rather than as a replacement for conventional statistical visualizations.

6 Limitations

The proposed framework and its current evaluation have several limitations that should be considered when interpreting the results. One practical limitation concerns the number of clusters represented in the network. As the number of clusters increases, the number of variability vectors and potential inter-cluster connections also increases, which can make the overall network structure more difficult to interpret in a single visualization. The framework can represent these relationships, but larger numbers of clusters may require additional approaches for organizing or filtering the displayed connections.

A similar issue arises when the number of attributes increases. For a dataset with m attributes, the proposed construction generates $2m + 1$ vectors for each cluster. Consequently, datasets with a large number of attributes can produce networks with a substantial number of nodes and connections, increasing the visual complexity of the resulting representation. This can make it more difficult to identify relevant relationships directly from the network and suggests the need for approaches that can reduce visual complexity while preserving important statistical relationships.

The use of static visualizations is another limitation of the current presentation. While static network figures are suitable for illustrating the framework and comparing the resulting representations, they become less practical as the size and complexity of the network increase. An interactive visualization environment could allow users to explore connections, inspect individual variability vectors, and filter the network according to clusters, attributes, or statistical representations. Tools such as Gephi[1] could provide one possible environment for this type of exploration.

Finally, the current study does not provide a formal statistical evaluation of the observed network relationships. The experiments are intended to demonstrate and illustrate the proposed framework rather than to test the statistical significance of individual connections or to establish whether the framework improves clustering quality.

[1] https://gephi.org/

No statistical hypothesis testing or systematic quantitative validation of the network patterns was therefore performed. The contribution of this work is the proposal and initial evaluation of the Statistical Variability Network framework, while a more extensive assessment using additional datasets, statistical analyses, and systematic variation of network-construction parameters remains for future work.

7 Conclusion

This study introduced STAT Networks as a network-based approach for representing the statistical variability of clustered data. By constructing variability vectors from different statistical summaries and connecting them according to their statistical proximity, the proposed framework provides a way to examine relationships between cluster profiles beyond their original network structure.

The experiments on the Iris, E. coli, and Yeast datasets demonstrated that the resulting network structure depends on the statistical representation used. The IQR, IWR, 1SD, 2SD, and CI networks revealed different patterns of inter-cluster connections, showing that different representations capture different aspects of the underlying cluster distributions. The results also demonstrated that relationships between cluster profiles may become apparent when their variability is considered, even when the corresponding profiles appear more separated in the original network.

Overall, STAT Networks provide a complementary representation for exploring cluster variability and relationships between cluster profiles. The choice of statistical representation should be guided by the characteristics of the data and the aspect of variability of interest. Further studies using additional datasets and alternative network-construction parameters would be useful to evaluate the robustness and general applicability of the proposed framework.

References

[1] Visualization of the Clustering Results, pp. 47–80. Birkhäuser Basel, Basel (2007). https://doi.org/10.1007/978-3-7643-7988-9_2 . https://doi.org/10.1007/978-3-7643-7988-9_2

[2] Seo, J., Shneiderman, B.: Interactively exploring hierarchical clustering results. IEEE Computer **35**(7), 80–86 (2002) https://doi.org/10.1109/MC.2002.1016905

[3] Malkai, O., Lu, M., Cohen-Or, D.: Clusterplot: High-dimensional cluster visualization. arXiv preprint arXiv:2103.02992 (2021)

[4] Schwartz, G.W., Zhou, Y., Petrovic, J., Fasolino, M., Xu, L., Shaffer, S.M., Pear, W.S., Vahedi, G., Faryabi, R.B.: Toomanycells identifies and visualizes relationships of single-cell clades. Nature methods **17**(4), 405–413 (2020)

[5] Anlauf, T., Kubikova, K., Ochodkova, E., Kriegova, E., Kudelka, M.: Simnetx: tinkering with patient similarity networks to understand biomedical data. Applied Network Science **10**(1), 54 (2025)

[6] Ochodkova, E., Zehnalova, S., Kudelka, M.: Graph construction based on local representativeness. In: International Computing and Combinatorics Conference, pp. 654–665 (2017). Springer

[7] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

[8] Duan, R., Tong, J., Sutton, A.J., Asch, D.A., Chu, H., Schmid, C.H., Chen, Y.: Origami plot: a novel multivariate data visualization tool that improves radar chart. Journal of clinical epidemiology **156**, 85–94 (2023)

Declarations

Funding. This work was supported by SGS, VŠB-Technical University Ostrava, under grant no. SP2026/008, Ministry of Health CZ NW24-10-00395 and DRO FNOL, 0098892.

Conflict of interest. The authors declare that they have no conflict of interest.

Ethics approval and consent to participate. Not applicable.

Data availability. There are no restrictions to the availability of data.

Materials availability. Not applicable.

Code availability. The LRNet algorithm is available at https://homel.vsb.cz/~kud007/lrnet_files/

Author contribution. The authors contributed equally to the work.