Design and Evaluation of Visual Summaries to Improve Readability of Large Network Visualizations

Rezwana Mahfuza
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
rezwana.mahfuza@usask.ca

Debajyoti Mondal
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
d.mondal@usask.ca

Carl Gutwin
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
gutwin@usask.ca

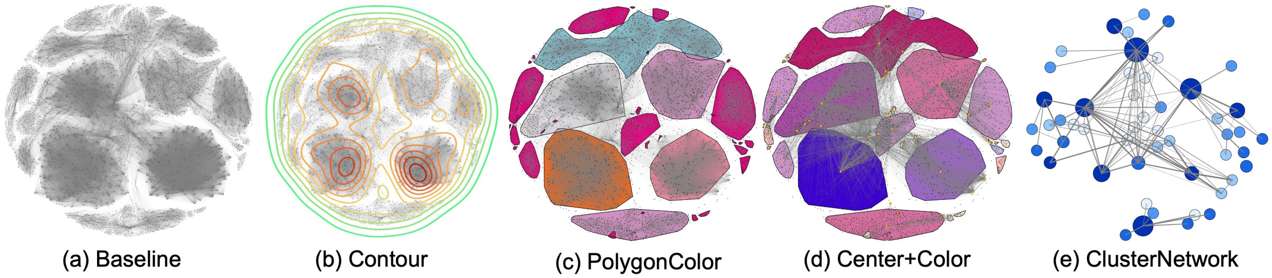


Figure 1: (a) A node-link representation of a Facebook network, and (b–e) its different levels of summarization. (b) Contours showing node distribution. (c) Clusters are colored based on size. (d) The sources of the outgoing links from a cluster are collapsed to a single point. (e) Each cluster is collapsed to a disk with a radius corresponding to the cluster size.

ABSTRACT

Node-link visualizations are commonly used to gain insights into large network data where the entities of the networks (nodes) are represented as points, and relationships (edges) are drawn as straight line segments or links. With growing access to network data and visualization tools, such visualizations are increasingly appearing in infographics and documents intended for non-specialist readers. This necessitates understanding how these visualizations are perceived by end users who are not necessarily domain experts, and determining how visual summaries can be provided to ensure consistent interpretation of the displayed information. In this paper, we investigate the interpretability of node-link visualizations of large graphs: we designed summary representations that could be provided alongside the visualization to improve interpretation, and we evaluated these designs through two user studies. Our results indicate that the information perceived from traditional node-link representations can vary substantially, especially when the nodes are uniformly distributed rather than forming clusters or tangled structures. We observed that visual summaries can enhance the readability of these visualizations – summaries that reduce clutter were preferred by participants and were more accurate for typical interpretation tasks.

CCS CONCEPTS

• **Human-centered computing → Graph drawings**; **Visualization design and evaluation methods**.

KEYWORDS

Large Networks, Node-Link Visualizations, Visual Summaries

ACM Reference Format:
Rezwana Mahfuza, Debajyoti Mondal, and Carl Gutwin. 2025. Design and Evaluation of Visual Summaries to Improve Readability of Large Network Visualizations. In *Proceedings of Graphics Interface 2025 (GI'25).* ACM, New York, NY, USA, 17 pages.

1 INTRODUCTION

Network data generated from real-world contexts, such as social media interactions, text corpora, co-authorship relations, or code dependencies in software databases, are typically large, containing millions of nodes and edges. Node-link visualizations of such networks are widely used to obtain a high-level understanding of the network structure [14]. One crucial piece of information is the identification of clusters, which are groups of nodes that have more connections within the group than with the rest of the graph. In a node-link visualization, clusters are revealed as blobs or tangled structures (sometimes called hairballs). For example, Figure 1(a) illustrates a Facebook network with about 4,000 nodes and 88,000 links, where each node represents an individual and each link represents a friendship relation. We can see several clusters in the visualization, which are not necessarily based on ground-truth communities but are rather formed based on the connectivity information in the graph. We can also see links connecting two different clusters, i.e., when some people from one cluster have friendship relations with a different cluster.

GI'25, May 26–29, 2025, Kelowna, BC, Canada

The increasing availability of visualization tools [1, 8, 35, 38] and their effectiveness in showcasing information have paved the way for these visualizations to be used in everyday settings. Examples include analyzing social media posts to find clusters of supporters of various parties in an election, examining library dependencies in software repositories, or identifying research groups by visualizing co-authorship graphs across various disciplines. Interpreting the structure of a large network from its visualization is challenging, even for simple tasks like estimating relative cluster sizes. Although the clusters are revealed as tangled structures, their relative sizes (i.e., number of nodes and links) are difficult to estimate as the area they occupy in the display may not always correlate with their actual size. Many specialized visual abstractions have appeared in the literature to improve the scientific understanding of large networks [9, 19, 20, 40], but as node-link visualizations become more prevalent, it is crucial to understand how general users, especially those without domain expertise, perceive these structures.

A widely used technique for creating network layout is to use force-based visualization algorithms (FA) [17] that simulate attraction and repulsion forces among the nodes. Prior work [12, 18, 33] evaluated the readability of FA layouts for networks with a few hundred nodes and for fine-grained tasks such as estimating distance between pairwise nodes and finding the neighbors of a given node. This approach does not apply to large networks where such detailed information is not visible, and thus leaves a significant gap in the literature on the readability of these layouts. Furthermore, force-based algorithms may not produce user-centric layouts [5] as they can disproportionately emphasize certain clusters over others due to variations in their network structures.

Our Contribution. To address the limitations of interpreting FA layouts for large networks, we propose providing an intuitive visual summary alongside an FA visualization. In this paper, we designed several kinds of summaries by detecting and highlighting clusters (Figure 1(b)–(e)) that can be algorithmically computed from a node-link visualization. We then evaluate these visualizations using realistic interpretation tasks on real-life networks through two user studies. Since we identify clusters from the node positions algorithmically, we first want to verify whether the clusters we identify are of good quality and stable in an FA visualization.

RQ1. *Can we find good quality clusters that align with the tangled structures revealed in an FA visualization?*

Once we establish the process for identifying high-quality clusters, we design summary representations with increasingly coarse granularity by encoding cluster properties into different visual cues. In the first user study, we investigate the extent to which participants agree on the identification of clusters in FA visualizations and whether the automatically-detected clusters align with those identified by participants.

RQ2. *How well do participants agree on their identified clusters? Do the automatically detected clusters align with those determined by the participants?*

We also investigate the effectiveness of five designs (the baseline visualization and the four visual summaries) for completing common visual interpretation tasks such as ranking the top three clusters based on the number of nodes or links and finding pairs of clusters that have many between-cluster links.

RQ3. *How can visual cues be utilized to create summaries that enhance the interpretation of cluster structures in FA visualizations? How do different levels of summarization aid in interpreting cluster sizes and between-cluster relationships?*

We conducted two user studies. Study 1 investigated task performance and user preferences for five different visualization designs. The tasks were designed to examine users' agreement on cluster identification in baseline node-link visualization, cluster size estimation and understanding relations between pairwise clusters, both with and without the aid of visualization summaries. We used uniform percentile-based thresholds to map cluster sizes to colors. The first study showed that participants' answers on identifying large clusters can vary substantially – and even more so when the tangled structures are not clearly visible in the visualization. If the tangled structures are clearly visible, then the automatically detected clusters often capture the clusters the participants detected. For cases where clusters are difficult to identify, automatically detected clusters can provide good recommendations. The study also showed that adding summaries can significantly improve accuracy on interpretation tasks. The analysis of the first study's results suggested several design improvements for the visualizations – and in the second user study, we used the refined designs.

Since the summaries were designed by identifying high-quality clusters, we noticed a skewed distribution for the cluster sizes. Therefore, in Study 2 we used geometric-series based percentiles to map cluster sizes to colors. Since the baseline node-link visualization remained unchanged, Study 2 omitted the cluster identification task. However, Study 2 included the remaining tasks of Study 1 and added more challenging tasks such as identifying clusters that have relationships with other clusters in the network and identifying clusters that have many sub-clusters within them. Study 2 further confirmed the effectiveness of summary representations for ranking and cluster-pair search tasks, while also revealing dataset-dependent performance variations, and providing insights into the value of improved color binning. Across both studies, participants favored coarser-grained visual summaries for ease of interpretation.

2 RELATED WORK

In this section, we review the literature on the design and user evaluation of node-link visualizations and discuss recent developments in the context of large networks.

Force-directed layouts [17] are widely used for creating node-link visualizations due to their effectiveness in revealing clusters based on network connectivity. However, force-based visualizations disproportionately emphasize certain clusters, which means that they are not inherently user-centric [5]. Eye-tracking studies have found that force-directed visualizations outperform many other network visualization styles such as orthogonal and layered layouts for various analytical tasks [15, 31]. However, fine-grained tasks such as estimating shortest paths or assessing node degrees are difficult to perform in dense networks with high connectivity, as these often lead to tangled structures that obscure structural relationships [16]. Hence, it is more common to consider overview tasks for large networks as specified in the task taxonomy for the evaluation of network layouts [21]. Given the scarcity of literature evaluating overview tasks in force-directed layouts of large networks, we specifically focus on the context of large networks.

To analyze large networks, various network summarization techniques, and abstractions have been proposed [11, 32]. For example, edge bundling attempts to route the links along a common curve to reduce clutter in the network. A technique called Graph Thumbnails [40] creates icon-like designs of the high-level structure of network data. While the former primarily focuses on revealing a general connectivity skeleton of the network, the latter investigates how many large graphs can be compared at a glance by comparing thumbnails. Other techniques include graph sampling [39], edge-bundled networks that can be explored as geographic maps [26, 28], and space partitioning to separate clusters that can later be explored interactively [39]. All these techniques deviate significantly from traditional node-link visualizations, and as our primary focus is on node-link layouts, we do not incorporate them when designing visualization summaries.

A rich body of past research has evaluated node-link visualizations with fine-grained tasks [12, 18, 29, 33], but for small graphs and without focusing on summarization. Recent research has increased the scalability of computing node-link visualizations [4], and many tools are now available that can quickly create node-link visualization for networks with millions of edges [1, 4, 35]. The availability of such tools has enabled researchers and practitioners from different domains to generate large network visualizations for big datasets and make them accessible to general users. Hence, we focus on designing visual summaries that may be provided alongside the traditional node-link visualization to overcome interpretability challenges for seemingly simple tasks such as assessing cluster sizes and relationships.

3 VISUAL ENCODING

In this section, we describe the. details of the baseline node-link visualization (created by FA [17]) and the four visualization summaries (Figure 1) that progressively summarize the node-link visualization. In the summaries, we worked with the top 50 high-quality clusters. To ensure that the clusters are of high quality and agree with the tangled structures of the node-link visualization, we detect clusters automatically as described below.

3.1 Finding Top 50 Clusters and RQ1

Detecting FA clusters: We first detect clusters from the FA visualization using a hierarchical density-based clustering algorithm called HDBSCAN [25]. HDBSCAN grows clusters by iteratively collecting their nearby points and thus can isolate regions of high point density as clusters. However, if we directly use the node positions of the FA visualization as the input to HDBSCAN, the tangled structures may not always be detected as clusters. This is because the edges play a crucial role in causing visual tangles. Therefore, we placed dummy nodes on short edges (those in the lowest 10th percentile of all edge lengths) and input the node positions, as well as these dummy nodes to HDBSCAN (Figure 2).

Detecting Modularity-Based clusters: A modularity-based clustering algorithm (the Louvain method [3]) partitions a network into clusters so that a metric called modularity is maximized. Modularity is a widely used metric in network science to assess the presence of cluster structures in a network by comparing the network with a random network of the same number of nodes and links. We used the Louvain method to find clusters in the network; however, since the Louvain method does not consider the FA layout, the nodes of a cluster may be distributed into many tangled structures (Figure 3).

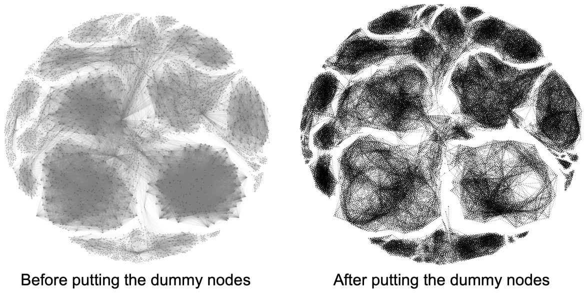


Figure 2: Interpolated dummy nodes added to short edges before applying HDBSCAN.

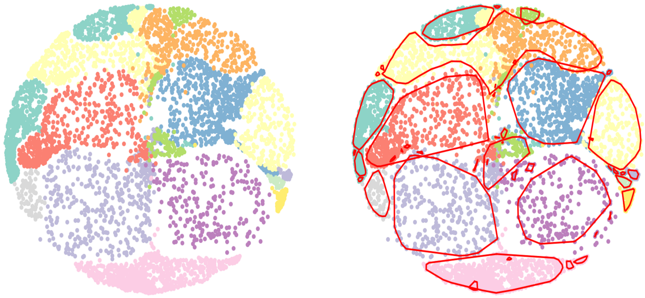


Figure 3: (left) Modularity-based clusters in distinct colors. (right) FA clusters in distinct polygons. Edges are omitted.

Filtering Top 50 clusters: Let c be an FA cluster which may contain nodes from many modularity-based clusters. We assign c the rank $f(c)$, which is the maximum number of nodes in c that belong to the same modularity-based cluster. Finally, we select the top 50 high-rank FA clusters to design our visualization summaries.

Variants of HDBSCAN have previously been used in the literature to detect clusters from network visualizations [24]. However, this may produce many clusters, and we needed a convincing approach to select clusters of good quality. Filtering the top 50 based on the rank we assigned using modularity-based clustering provides us with some justification to retain good-quality clusters. One can observe from Figure 3 that large modularity-based clusters roughly align with FA clusters, which provides evidence towards RQ1 that it is possible to find clusters that align with the tangled structures visible in an FA visualization. We next use these top 50 detected clusters for our visualization designs.

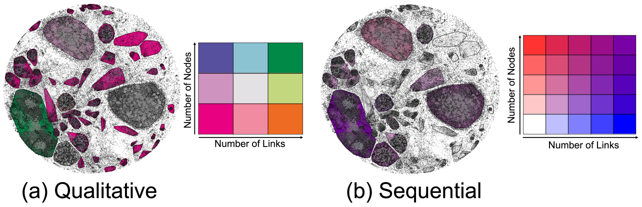
3.2 Node-Link Visualization and Summaries

We used five designs for the study: the traditional node-link visualization as a baseline, and four visualization summaries. A visualization summary is presented as a pair, with the baseline visualization shown at left and the summary representation shown at right. We chose to show the two representations together because (a) we expect a summary to augment understanding but not replace the baseline, and (b) interactive dashboards often make more than one alternative version available to users.

Vis. 1 (`Baseline`) Node positions in the baseline visualization are obtained using the traditional FA algorithm (Figure 1(a)). Nodes are drawn as small black circles, and the connections between them are drawn as grey lines. We rendered the nodes after the links so that they remain visible, and we reduced the opacity of the links so that clusters become visible.

Vis. 2 — Summarize Node Distribution (`Contour`) This design summarizes the distribution of the nodes using contour lines and a color gradient (Figure 1(b)). We used a contour overlay because this technique is often used in large network visualizations [42]. The background of the design is the baseline node-link visualization. We applied the Kernel Density Estimation (KDE) [34] to define the node density at each point of the network, and then overlaid a contour plot where all points of each contour line represent the same node density. This provides an intuitive density distribution of the nodes. For better readability, we colored the contour lines with a continuous colormap from green (low density) to red (high density).

Vis. 3 — Summarize Cluster Size (`PolygonColor`) This design summarizes the size (number of nodes and links) of each cluster by coloring the polygon that encloses the cluster (Figure 4); the rationale for using enclosing polygons is that they are commonly used to highlight clusters [6, 10, 33]. To enclose the nodes of a cluster inside a polygon, we use a geometric object called an alpha shape which performs better than a convex hull by carving off unnecessary space [7].



(a) Qualitative (b) Sequential

Figure 4: Vis. 3 — `PolygonColor`, where (a) shows the qualitative colormap and (b) shows the sequential colormap.

For the colormap, we used a 2D color matrix where the vertical and horizontal axes correspond to the number of nodes and links. We tested two different colormaps in our study, one sequential and the other qualitative — both have previously been used in the literature to design bivariate visualization [41], and the qualitative colormap is inspired by the Corners Model [37] to highlight distinct regions of high and low values while minimizing the prominence of intermediate values. We create color bins based on the quantile values.

Vis. 4 — Summarize Cluster edges (`Center+Color`) This design further summarizes the clusters by merging the sources of the links that leave a cluster into the cluster center, which is computed by taking the mean of x- and y-coordinates of the corresponding polygon corners (Figure 1(d)). The starting point for making this design is `PolygonColor`. We do not change the node locations, but the links within the cluster get removed and the outgoing links from a cluster now emanate from the center point. The rationale behind this higher-level summarization strategy is that the identification of clusters and their color already gives some idea of the cluster sizes, and thus one can focus on understanding the external relations that a cluster has with other clusters.

Vis. 5 — Map Clusters to Disks (`ClusterNetwork`) This design maps clusters to disks, with the size of a disk relative to the number of nodes in the cluster (Figure 1(e)). The number of links in a cluster is mapped to a color using the quantile values of the number of links. The opacity of the link between two disks is relative to the number of relations between the corresponding clusters, and each disk is initially positioned at the cluster center. Such node-link networks of clusters commonly appear in network summarization [27, 36]. To prevent overlaps, the disks are iteratively moved away from each other until a valid placement without collision is found. Although this adjustment to disk layout can make it harder for users to match the disks to the corresponding part of the baseline visualization, the clusters can in practice be labeled or highlighted in a mouse-hover operation.

Networks for User Study: We used five large networks: a sample of Facebook friend relations, DBLP author collaborations, a Web hyperlink network, a YouTube social network, and a Brightkite social network; node and link counts for these networks are (4039, 88234), (4973731, 224133), (685228, 262279), (1157617, 137585), and (58228, 201144), respectively. We downloaded the Datasets from Stanford Large Network Dataset Collection (SNAP) [22]. For scalability, we progressively removed low-degree nodes from the network until the number of links was reduced to 300,000.

4 STUDY 1 (CLUSTER IDENTIFICATION AND DESIGN VALIDATION)

Our first study investigates RQ1 – the degree of agreement when identifying clusters (both across participants and between FA clusters and participant choices), and RQ2 – how participants perform when using different visualization designs that present different types of summary information.

4.1 S1: Study Methods

4.1.1 S1: Participants and Apparatus. We recruited 20 participants (15 male, 5 female, mean age 28 years), all of whom were undergraduate or graduate students at a local university. Participants were not required to have prior expertise in network analysis or data visualization. None of the participants reported a color vision deficiency.

The study was conducted in a controlled environment using a 75-inch 4K display and a Windows laptop with a standard mouse. Participants completed the tasks using a custom web-based system developed with the P5JS toolkit, which was used to present the tasks and record responses and completion times.

Table 1: Tasks used in the user study 1 and 2, and their corresponding domains

Study	Type	Task	Domain
1	Identification	Identify the top three clusters separately by node and link count.	Interpret cluster locations in `Baseline`.
1,2	Ranking	Ordering a given set of 5 clusters separately by node and link count.	Interpret and compare sizes of different clusters.
1,2	Pair-Search	Identify and rank the top three cluster pairs by between-cluster links.	Assess the strength of between-cluster relations.
2	Degree-Estimation	Identify the cluster that has links to the highest number of other clusters.	Interpret degree centrality in a network of clusters.
2	Subcluster-Estimation	Identify the cluster with the highest number of sub-clusters.	Interpret fine-grained nested cluster structures.

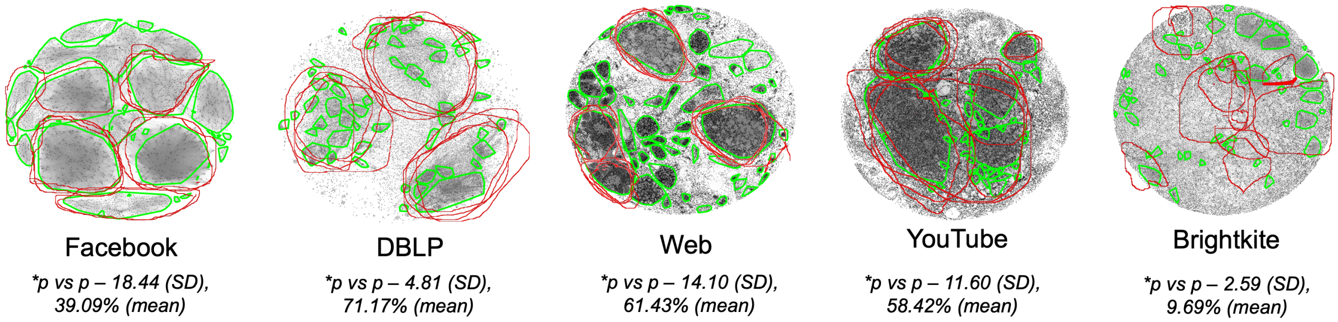


Figure 5: Participants' answers for top-3 cluster identification task (in red) and the top 50 FA clusters (in green).

4.1.2 S1 Tasks and Datasets. Participants performed three types of tasks (Table 1). The Identification task asked participants to draw outlines around the top three clusters by node count and the top three by link count. In the Ranking task, we pre-labeled five clusters in the visualization summary and asked participants to provide two orderings of the clusters — one based on node count and the other based on link count. The Search task required that participants find the three pairs of clusters that had the highest number of between-cluster links. The Identification task used only the `Baseline` design, while the Ranking and Search tasks were carried out with all five visualization designs. Each participant carried out their tasks with only one of the five network datasets specified in Section 3.2; this meant that each participant completed 11 task instances (one Identification task, five Ranking tasks, and five Search tasks).

4.1.3 S1 Hypotheses. We hypothesized that participants would have low agreement among themselves and also with FA clusters ($h1$). For the Ranking task, we hypothesized that `Baseline` will have lower accuracy, while `ClusterNetwork` will have higher accuracy for the Pair-Search tasks than the others ($h2$). We also hypothesized that `Center+Color` and `ClusterNetwork` will have higher accuracy for the Pair-Search task ($h3$).

4.1.4 S1 Procedure. All participants completed informed consent and demographics forms before the study and were compensated with $15 CAD for their participation. Participants were then given an explanation of each visualization and completed a training session which explained how to interpret each type of visualization using annotated examples. For example, for the `Contour` design, participants were told that `Baseline` would appear on the left and `Contour` would appear on the right, and that the contour lines represented node density, with an example showing a blue dotted boundary highlighting a high-density region. The study system then presented 11 task instances (one instance of the Identification task, and five instances of Ranking and Search) and recorded all relevant study data. For each of the 11 task instances, participants were shown a visualization, accompanied by a brief guideline at the top of the page to remind them how to interpret that specific design. After completing the tasks for each Design, participants answered questions about the difficulty of retrieving information from that design, and completed a NASA-TLX-style questionnaire to assess their perceived effort. At the end of each task, they completed a question asking them which Design was fastest, most accurate, and most preferred. To counterbalance learning effects, Datasets and Designs were paired using a Latin square structure. Participants were instructed to complete the tasks as quickly and accurately as possible. The mean completion time of the entire study was 52 minutes. Ethical approval was received from the research ethics board at the University of Saskatchewan.

4.1.5 S1 Study Design. The study used a mixed factorial design with two factors: the primary factor (within-participants) was the visualization Design (`Baseline`, `Contour`, `PolygonColor`, `Center+Color`, `ClusterNetwork`), and the secondary factor (between-participants) was Dataset (Facebook, DBLP, Web, YouTube, and Brightkite). The ordering of designs and the single dataset used by each participant was counterbalanced using a Latin square. We separately investigated the effects of a third factor, type of ColorMap (Qualitative or Sequential), using only data from the two visualization designs that used color (`PolygonColor` and `Center+Color`).

Performance-based dependent measures were: agreement between participant answers (for the Identification task), accuracy for the Ranking and Search tasks (computed using task-specific metrics such as mean edit distance and mismatch counts), and task completion time for Ranking and Search. Subjective measures included participant ratings of how difficult it was to obtain information from each design, as well as ratings of each design in terms of effort and preference.

4.2 S1 Results

Analyses are organized below by task – first we report the degree of agreement in cluster identification (Identification task), then analyse accuracy and completion time (Ranking and Search tasks), and then present analysis of subjective measures.

For factorial analyses, effect sizes for significant results are reported as generalized eta squared (η^2) [23, 30]. Follow-up tests were corrected using the Holm-Bonferroni method. No data were removed as outliers. In the analyses below, we do not report main effects of Dataset, since differences in this factor were expected due to the varying structure and complexity of the data. We report main effects of visualization Design and interactions between Design and Dataset.

4.2.1 S1: Identification Task. Our RQ2 asks whether participants agree among themselves about clusters, and whether participant clusters agree with automatically-detected clusters. First, our initial hypothesis ($h2$) was that the participants would have a low agreement in the clusters they identified. However, it appears that the answer depends on whether the clusters are revealed as a distinctive tangled structure or not. Figure 5 illustrates participant answers for various Datasets in red where the task was to identify the top three clusters by node count We computed the mean pairwise Jaccard similarity, i.e., $|P_i \cap P_j|/|P_i \cup P_j|$ where P_i and P_j are the union of the polygons drawn by the ith and jth participants. The mean Jaccard similarity was 46.36% – Datasets in which tangled structures were distinctive had a high agreement (DBLP - 71.17%, Web - 61.43%, and YouTube - 58.42%), and Datasets in which clusters were not well separated or less visible had a low agreement (BrightKite - 9.69%). This suggests that guidance based on computational analysis needs to be provided, especially for Datasets where a consensus cannot be reached.

Second, we considered whether FA clusters align with participant-identified clusters. Figure 5 shows the top 50 FA clusters in green. The FA clusters often matched the top three clusters identified by participants, but were often more granular than the participants' clusters. The latter case is again more prominent when the clusters in the visualization are less distinctive.

4.2.2 S1: Ranking Task. Participants ranked five pre-labeled clusters twice: once by node count and once by link count. We measured accuracy in the rankings by calculating edit distance (using NLTK's default edit distance function [2]) between a participant's ranking and the correct ranking (lower values indicate better accuracy).

Accuracy of node-count ranking. Figure 6 summarizes the edit distance results for the ranking task, and Table 2 presents findings from the Design × Dataset ANOVA. The ANOVA found a significant main effect of Design – as shown in Figure 6, `ClusterNetwork` (mean edit distance 0.85) had lower edit distances across most datasets, and significantly lower mean than `Baseline` (2.30) and `Contour` (2.60).

We also found a significant Design × Dataset interaction (Table 2). To explore this interaction, we carried out one-way ANOVAs to look for effects of Design within each Dataset. We found significant effects of Design on edit distance in the Facebook Dataset ($F_{4,15} = 5.02, p < .05$) and the Brightkite dataset ($F_{4,15} = 5.29, p < .05$). Follow-up t-tests showed that `ClusterNetwork` (0.50) had lower edit distance than `Contour` (4.0) in the Brightkite dataset; there were no pairwise differences found for the Facebook dataset.

Accuracy of link-count ranking. For the link-ranking task, Figure 7 summarizes the mean edit distances, and Table 2 shows the ANOVA results. As shown in Figure 7, edit distances across the different designs were similar, and we did not find a main effect of Design. However, we found a significant interaction between Design and Dataset. As seen in Figure 7 (right), there is a substantial accuracy difference across designs in the Web dataset (with `Baseline` performing much worse) compared to the other datasets. To explore this interaction, we carried out one-way ANOVAs for each dataset. We found a significant effect of Design on edit distance in the Web dataset ($F_{4,15} = 5.43, p < .05$), but no pairwise differences were found.

Completion time for ranking tasks. A single completion-time measure was gathered that included both the node-ranking and the link-ranking tasks. Table 2 summarizes the Design × Dataset ANOVA (we found no effect of Design on completion time, nor any interaction).

Qualitative vs. Sequential colormaps. To assess whether the type of colormap affected accuracy in the node-count or link-count tasks, we compared the Qualitative and Sequential colormaps. Figure 8 summarizes the edit distance results for the ranking task, and Table 5 presents ANOVA results for factor ColorMap on accuracy and completion time (no effects were found).

4.2.3 S1: Search Task. Participants identified the top three cluster pairs with the highest between-cluster link counts, from a set of five pre-defined clusters (which form a total of 10 possible pairs). We measured mismatch count by the number of selected pairs that were not in the solution set (lower values indicate better accuracy).

Accuracy of search. Figure 9 summarizes the mismatch results, and Table 4 presents the Design × Dataset ANOVA. We did not find a main effect of Design or a significant interaction between Design and Dataset. As seen in Figure 9 (right), however, there is a substantial accuracy difference across designs in the YouTube dataset compared to the other datasets, which can be explored in future studies.

Completion time for search. Figure 10 summarizes the completion time for the search task, and Table 4 presents findings from the Design × Dataset ANOVA. The ANOVA found a significant main effect of Design - as shown in Figure 10, `Baseline` (2.35) took significantly more time than `Contour` (1.65), `Center+Color` (1.56), and `ClusterNetwork` (1.54).

We also found a significant Design × Dataset interaction (Table 4) for completion time. To explore this interaction, we carried out a one-way ANOVAs to look for effects of Design within each Dataset. We found significant effects in the Facebook dataset ($F_{4,15} = 5.12, p < .01$), the Web dataset ($F_{4,15} = 3.77, p < .05$), and the Brightkite dataset ($F_{4,15} = 4.27, p < .05$). Follow-up t-tests

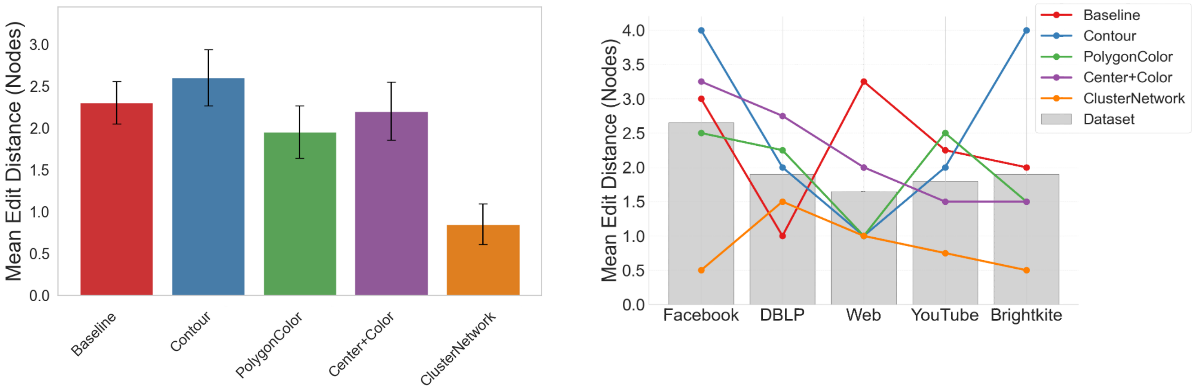


Figure 6: S1 Node ranking: (Left) mean edit distance ± SEM, by Design; (Right) mean edit distance ± SEM, by Dataset and Design.

S1: Factor	DF (n,d)	F	*p*	η^2	Pairwise Contrasts (mean), t-test result
Edit Distance - Node Count					
Design	4,75	6.15	**p<.001**	0.174	ClusterNetwork (0.85) < Baseline (2.30), **p<.05** ClusterNetwork (0.85) < Contour (2.60), **p<.05**
Design × Dataset	16,75	2.08	**p < .05**	0.236	Contrasts reported in text
Edit Distance - Link Count					
Design	4,75	1.38	0.25	-	
Design × Dataset	16,75	2.07	**p<.05**	0.222	Contrasts reported in text
Completion Time - Combined Node Count and Link Count					
Design	4,75	1.60	0.18	-	
Design × Dataset	16,75	0.86	0.61	-	

Table 2: S1: Factorial Analysis of Accuracy and Completion Time, Node and Link Ranking Tasks

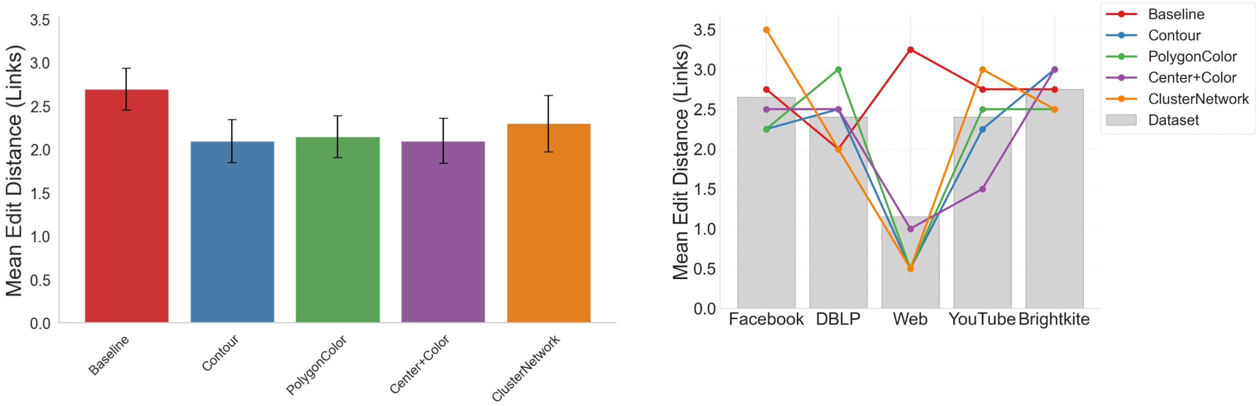


Figure 7: S1 Link ranking: (Left) mean edit distance ± SEM, by Design; (Right) mean edit distance ± SEM, by Dataset and Design.

showed that ClusterNetwork (1.13) had significantly lower completion time than Baseline (2.36) in the Web dataset; no significant pairwise differences were found for the Facebook or Brightkite datasets.

Qualitative vs. Sequential colormaps. To assess whether the type of colormap affected accuracy in the search task, we compared the Qualitative and Sequential colormaps. Figure 9 summarizes the mismatch result, and Table 5 presents findings from the one-way ANOVA. We did not find any significant main effect of ColorMap, either on mismatch counts or on completion time.

4.2.4 S1 Subjective Measures. We gathered two types of subjective measures. First, participants rated each Design based on how well it supported four analyses: identifying good clusters, estimating

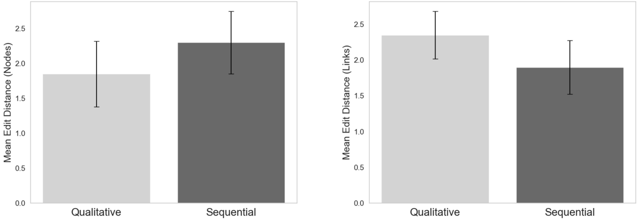


Figure 8: S1: Mean edit distances (± SEM), by ColorMap, for node-count ranking (left) and link-count ranking (right).

S1: Measure	**DF (n,d)**	**F**	p	η^2
Edit Distance – Node Count	1,38	0.95	0.34	–
Edit Distance – Link Count	1,38	1.66	0.21	–
Completion Time	1,38	0.04	0.83	–

Table 3: S1: Factorial Analysis of accuracy and completion time for factor ColorMap (ranking tasks).

cluster size based on node count, estimating size based on link count, and estimating between-cluster links.

Figure 12 summarizes users' ratings across different designs. As shown in the figure, participants gave better ratings to designs with a higher degree of summarization across all four analysis types (with the ClusterNetwork visualization consistently ranked best).

Second, after each task, participants ranked the Designs in terms of speed, ease of use, and preference, along with a workload assessment using a TLX-style survey [13]. As shown in Figure 13, ClusterNetwork consistently outperformed the other designs, being rated as fastest, easiest to use, and most preferred. Center+Color also demonstrated strong usability, particularly in ease of use and preference, while Baseline consistently ranked the worst (Figure 14).

4.3 S1 Discussion

Performance of* Baseline *compared with other designs. For the ranking task (*h2*), we expected that Baseline would have lower accuracy and that ClusterNetwork to perform best. The results partially support this hypothesis: ClusterNetwork significantly outperformed Baseline and Contour when ranking clusters by node count. However, for link count, all Designs performed similarly (although Baseline had a slightly higher mean edit distance). The significant interaction between Design and Dataset suggests that visual encoding effectiveness varies depending on whether cluster structures are clearly visible. However, a high agreement in cluster identification may not translate to accurate ranking (as seen with the DBLP Dataset).

Does between-cluster link summarization perform better on the search task? For the search task, we expected CenterColor and ClusterNetwork to have higher accuracy (*h3*); however, no significant effect of Design on mismatch count was found. However, there were differences in task completion time, with Baseline taking longer than Contour, Center+Color, and ClusterNetwork.

Did performance differ across the two colormaps? We found no significant effects of ColorMap for accuracy (on either the node-count or link-count ranking) or for completion time.

Did user preferences vary across the Designs? User preference and TLX assessments further highlight perceived usability differences, with ClusterNetwork rated as the most efficient and Baseline as the least. TLX scores show that Baseline imposed greater cognitive effort, which aligns with its lower accuracy and efficiency.

4.4 Design Refinements for Study 2

Our findings led us to refine our Designs before we ran Study 2. When we provided five candidate clusters for the ranking and pair-search tasks, we outlined the clusters manually for Baseline and used automatically detected clusters for the visual summaries. This is because the polygon boundary of an automatically detected cluster may not exactly match if we overlay it on Baseline, which could be a potential source of confusion among participants. However, this creates a problem when we compare accuracy for Base line with other Designs. Hence we decided to use automatically-detected polygons to also highlight clusters in Baseline. We also noticed that Center+Color had similar accuracy to that of Base line when ranking clusters by nodes, despite having less clutter. Hence we investigated and corrected a rendering issue that placed links on top of nodes, potentially affecting accuracy in estimating node count. Additionally, to overcome the visual clutter in the summaries Contour and PolygonColor, we omitted links that are not part of the top 50 clusters. We also noticed that PolygonColor and Center+Color performed similarly to all other Designs when ranking clusters by link count; many clusters were classified as having high link counts (i.e., darker colors) despite using a percentile-based binning due to a skewed distribution. When the color of two clusters is the same, participants have no choice but to compare their polygon area or appearance in Baseline. Therefore, in Study 2, we selected percentiles by choosing thresholds in a geometric series, i.e., p_1, p_2, p_3 where $p_i = 1 - \left(\frac{1}{4^i}\right), \quad i \in \{1, 2, 3\}$. This creates a larger bin size at lower values and a finer bin size at higher values, effectively handling skewed distributions and providing more color variations.

5 STUDY 2 (INTERPRETING CLUSTER SIZES AND STRUCTURES)

Study 2 was designed to further investigate how different visual summaries support participants in interpreting cluster sizes and structures, with a wider variety of analysis tasks.

5.1 S2: Study Methods

5.1.1 S2 Participants and Apparatus. We recruited 20 new participants (11 male, 9 female, mean age 26.8 years) who had not gone through Study 1; participants were not required to have prior expertise in network analysis or data visualization. None of the participants reported a color vision deficiency.

The study was conducted in the same environment used for Study 1, including the same 75-inch 4K display, Windows laptop, and standard mouse. Tasks were presented with an updated version of

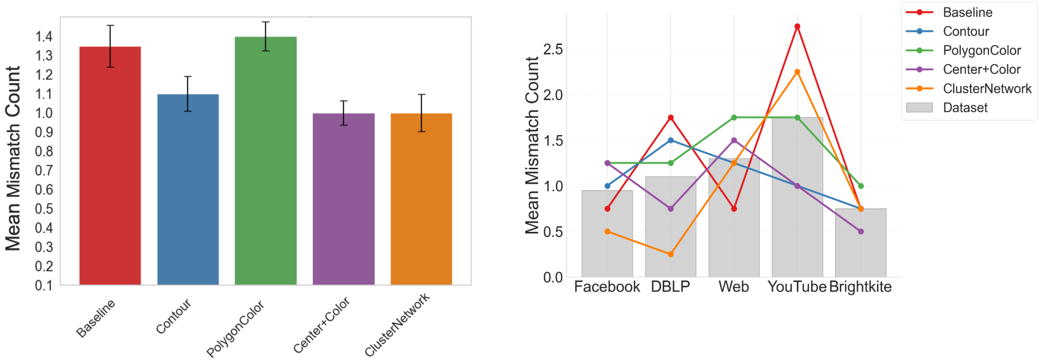


Figure 9: S1 Accuracy for Search task: Mean mismatch ± SEM, by Design (left) and by Dataset and Design (right).

S1: Factor	DF (n,d)	F	p	η^2	Pairwise Contrasts (mean), t-test result
Mismatch Count					
Design	4,75	1.19	0.32	-	
Design × Dataset	16,75	1.68	0.07	-	
Completion Time					
Design	4,75	7.61	**p<.001**	0.1839	`Contour` (1.65), `Center+Color` (1.56), `ClusterNetwork` (1.54) < `Baseline` (2.35)
Design × Dataset	16,75	2.80	**p<.01**	0.2705	Contrasts reported in text

Table 4: S1: Factorial Analysis of Accuracy and Completion Time, Search Task

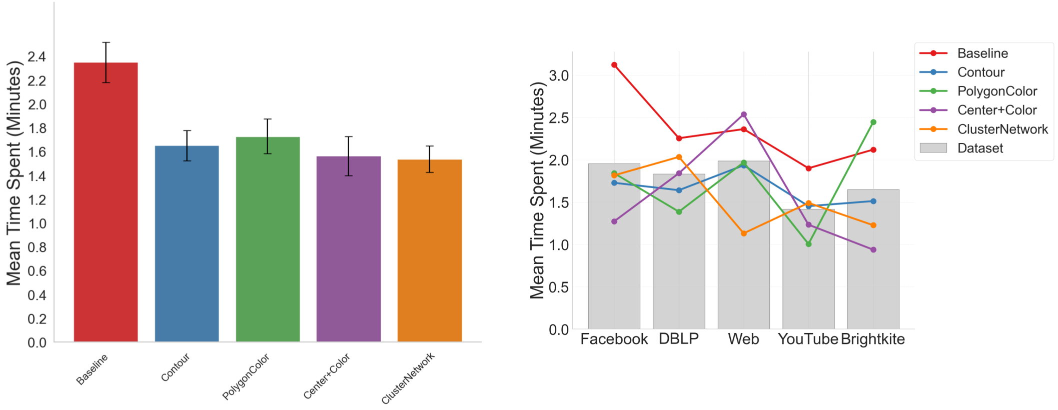


Figure 10: S1 Search Task: (Left) completion time ± SEM, by Design; (Right) completion time ± SEM, by Dataset and Design.

S1: Factor	DF (n,d)	F	p	η^2
Mismatch Count	1,38	1.75	0.19	–
Completion Time	1,38	1.53	0.22	–

Table 5: S1: Factorial Analysis of accuracy and completion time for factor ColorMap (search task)

the custom web-based system described above, which also recorded participants answers.

5.1.2 S2 Tasks and Datasets. Study 2 used two tasks from Study 1 (Ranking and Search), and added two additional tasks (see Table 1): the Degree-Estimation task required users to identify the cluster having links to largest number of other clusters (i.e., with the highest degree centrality in the cluster network), and the Subcluster-Estimation task required participants to identify the cluster with largest number of subclusters in it. These new tasks are somewhat more challenging, but are also common when interpreting a network visualization. The Subcluster-Estimation task did not include the `ClusterNetwork` design, while the remaining four tasks incorporated all five Designs across five different Datasets. As with Study

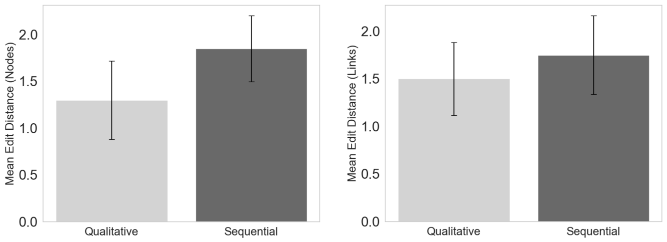


Figure 11: S1 Search: Mean mismatch ± SEM (Left) and mean time completion (Right), by ColorMap.

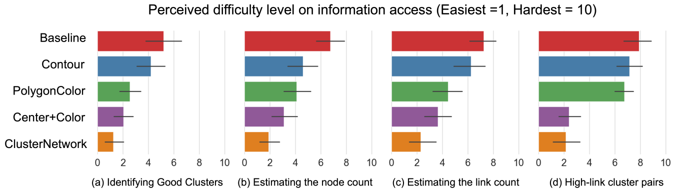


Figure 12: S1: Mean perceived difficulty for information access (± SEM), by Design. Shorter bars are better.

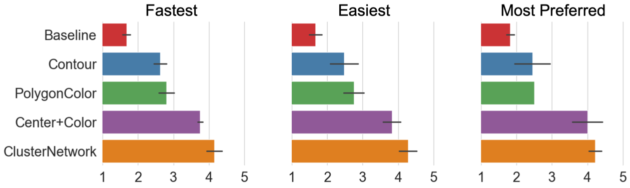


Figure 13: S1: Subjective Question Responses (Preference) (± SEM), by Design. Longer bars are better.

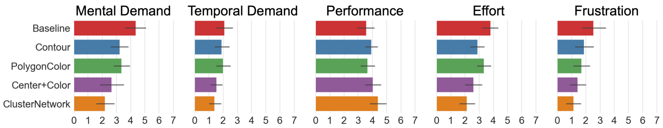


Figure 14: S1: Subjective Question Responses (Workload) (± SEM), by Design. Shorter bars are better (except for Performance).

1, each participant carried out their tasks with only one of the five datasets (Section 3.2); this meant that each participant completed 19 task instances (four Subcluster-Estimation instances, and five instances for each of the Ranking, Search, and Degree-Estimation tasks).

5.1.3 S2 Hypotheses. For Study 2, we hypothesized that all Designs would outperform Baseline in ranking tasks (*h4*). We expected Center+Color and ClusterNetwork, which provide a summary of between-cluster links, to achieve higher accuracy than other Designs (*h5*). Additionally, due to the difficulty of understanding subcluster structure, we hypothesized that the Designs would not have any effect on the mean agreement score for subcluster estimation (*h6*).

5.1.4 S2 Procedure. Study 2's procedure was similar to Study 1. Participants completed informed consent and demographics forms before the study and were then given an explanation of each visualization and completed a training session which explained how to interpret each design. The study system then presented 19 task instances (five instances of the Ranking, Search, and Degree-Estimation tasks, and four instances of the Subcluster-Estimation task), and recorded all study data. After completing the tasks for each design, participants answered questions about the difficulty of retrieving information from the design. After all instances of a task were done, participants completed a NASA-TLX-style questionnaire, and rated the design in terms of speed, accuracy, and preference. To counterbalance learning effects, Datasets and Designs were paired using a Latin square structure. Participants were instructed to complete the tasks as quickly and accurately as possible. The mean completion time of the entire study was 61 minutes, and participants were compensated with $15 CAD for their participation. Ethical approval was received from the research ethics board at the University of Saskatchewan.

5.1.5 S2 Study Design. Study 2 used a mixed-factorial design similar to that of Study 1: Design was the primary (within-participants) factor, with levels (Baseline, Contour, PolygonColor, Center+Color, and ClusterNetwork), and Dataset was a secondary between-subjects factor (Facebook, DBLP, Web, YouTube, Brightkite). As with Study 1, the ordering of designs and the single dataset used by each participant was counterbalanced using a Latin square. We again carried out a separate investigation of the effects of ColorMap (Qualitative or Sequential), using data from designs PolygonColor and Center+Color.

Performance-based dependent measures were: accuracy and completion time for the Ranking and Search tasks (using the same metrics as for Study 1); accuracy and completion time for the Degree-Estimation task (based on distance from the correct answer); and degree of agreement for the Subcluster-Estimation tasks. (We used agreement for Subcluster-Estimation rather than accuracy because it is difficult for automated tools such as DBSCAN to identify the exact number of subclusters; due to the lack of a clear ground truth, we opted to assess participant agreement instead of accuracy). Subjective measures included participant ratings of how difficult it was to obtain information from each design, as well as ratings of each design in terms of effort and preference.

5.2 S2 Results

5.2.1 S2: Ranking Task. As in Study 1, we calculated the edit distance between a participant's rankings (one based on node count and one based on link count) and the corect ranking.

Accuracy of node-count ranking. Figure 15 summarizes edit distances, and Table 6 shows results of the Design × Dataset ANOVA. We found a significant main effect of Design – as can be seen in Figure 6, ClusterNetwork (mean edit distance 0.70) had lower edit distances across most datasets, and overall was significantly lower than Baseline (1.85) and Contour (2.45).

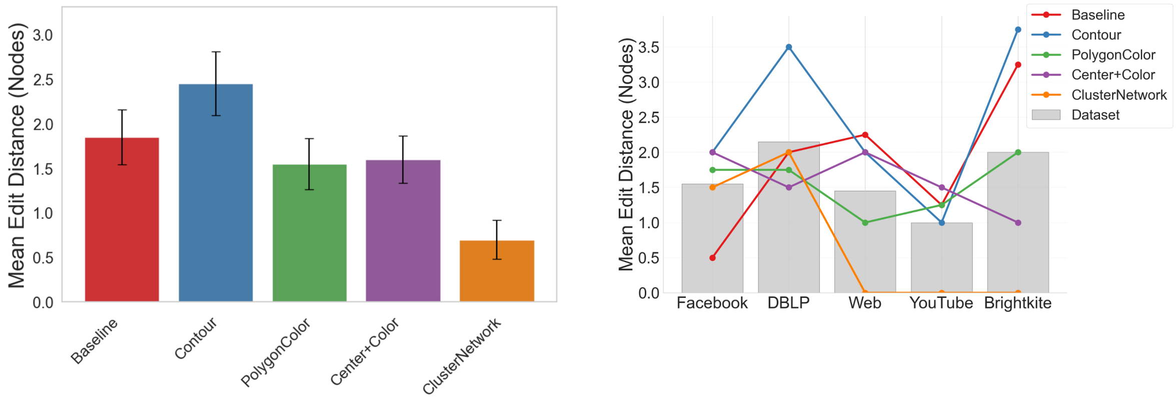


Figure 15: S2 Node ranking: (Left) mean edit distance ± SEM, by Design; (Right) edit distance by Design and Dataset.

We also found a significant Design × Dataset interaction (Table 6). To explore this interaction, we carried out one-way ANOVAs to look for effects of Design within each Dataset. We found significant effects of Design on edit distance in the Web Dataset ($F_{4,15} = 5.02, p < .05$) and the Brightkite dataset ($F_{4,15} = 5.29, p < .05$). Follow-up t-tests showed that for both datasets, `ClusterNetwork` (which had perfect accuracy in these datasets) had a lower edit distance than `Baseline` (edit distance 2.25 for Web, and 3.25 for Brightkite).

Accuracy of link-count ranking. For the link-ranking task, Figure 16 summarizes edit distances, and Table 6 shows the Design × Dataset ANOVA. We again found a significant effect of Design – as shown in Figure 16, `Baseline` (2.85) had higher edit distances than `PolygonColor` (1.55) and `Center+Color` (1.70). We did not find any significant interaction between Design and Dataset.

Completion time for ranking tasks. As in Study 1, a single completion-time measure included both the node-ranking and the link-ranking tasks. Table 6 summarizes the results of (Design × Dataset) ANOVA. We found no significant effect of Design on completion time, nor any interaction with Dataset.

Qualitative vs. Sequential colormaps. We again assessed whether Qualitative or Sequential colormaps led to greater accuracy or reduced completion time. Figure 17 summarizes the edit distance results by ColorMap, and Table 7 presents ANOVA results. No effects of ColorMap were found on either measure.

5.2.2 *S2: Search Task.*

As in Study 1, participants identified the top three cluster pairs with the highest between-cluster link counts; we measured mismatch count by the number of selected pairs that were not in the solution set.

Accuracy of search. Figure 18 summarizes the mismatch results, and Table 8 presents the Design × Dataset ANOVA. We found a significant effect of Design - as shown in Figure 18, `Baseline` (1.55) had significantly higher mismatch counts than `Center+Color` (0.75) and `ClusterNetwork` (0.80). We did not find a significant interaction between Design and Dataset.

Completion time for search. Figure 19 summarizes completion time measures, and Table 8 presents findings from the Design × Dataset ANOVA. We found a significant main effects of Design – as seen in Figure 19, `Baseline` (1.48) took significantly more time than `ClusterNetwork` (2.82). We did not find a Design × Dataset interaction.

Qualitative vs. Sequential colormaps. We compared the Qualitative and Sequential colormaps in terms of accuracy and completion time in the Search tasks. Figure 20 summarizes the mismatch result and completion time by ColorMap, and Table 9 presents ANOVA results. We did not find effects of ColorMap on mismatch count, but there was a significant effect of ColorMap on completion time – as shown in Figure 20, tasks with the Qualitative colormap (2.12 minutes) took more time than with the Sequential colormap (1.85 minutes).

5.2.3 *S2: Degree-Estimation Task.*

Participants identified the cluster with the highest degree of centrality (i.e., the cluster connecting to the largest number of other clusters). We scored accuracy based on the distance from the correct answer in a ranked list (e.g., if a participant correctly chose the cluster with the highest degree, their answer was scored as 5; if they chose the cluster with the second-highest degree, their answer was scored as 4, and so on, with minimum score 0).

Accuracy of Degree-Estimation. Figure 21 summarizes the accuracy results (distance from correct answer) for the Degree-Estimation task. Table 10 presents findings from the Design × Dataset ANOVA. We found a significant effect of Design – as shown in Figure 21, `Center+Color` (4.0) had a significantly better score than `Contour` (2.8). We did not find a significant interaction between Design and Dataset.

Completion time for Degree-Estimation task. Table 10 shows the Design × Dataset ANOVA; we did not find an effect of Design or a Design × Dataset interaction.

Qualitative vs. Sequential colormaps. We again compared the Qualitative and Sequential colormaps; Table 11 presents ANOVA results for factor ColorMap on accuracy and completion time (no effects were found).

5.2.4 *S2: Subcluster-Estimation Task.*

In this task, participants were asked to choose the cluster with the most subclusters. As described above, we chose to assess participant agreement in this task due to the difficulty of calculating true accuracy (see Section 5.1.2).

S2: Factor	**DF (n,d)**	**F**	***p***	η^2	**Pairwise Contrasts (mean), t-test result**
			Node Count		
Design	4,75	5.84	**p<.001**	0.165	`ClusterNetwork` (0.70) < `Baseline` (1.85), **p<.01** `ClusterNetwork` (0.70) < `Contour` (2.45), **p<.01**
Design × Dataset	16,75	1.94	**p < .05**	0.219	Contrasts reported in text
			Link Count		
Design	4,75	6.12	**p < .001**	0.186	`PolygonColor` (1.55) < `Baseline` (2.85), **p < .05** `Center+Color` (1.70) < `Baseline` (2.85), **p < .05**
Design × Dataset	16,75	1.32	0.21	-	
			Completion Time - Combined Node Count and Link Count		
Design	4,75	1.90	0.11	-	
Design × Dataset	16,75	0.68	0.80	-	

Table 6: S2: Factorial Analysis of Accuracy and Completion Time, Node and Link Ranking Tasks

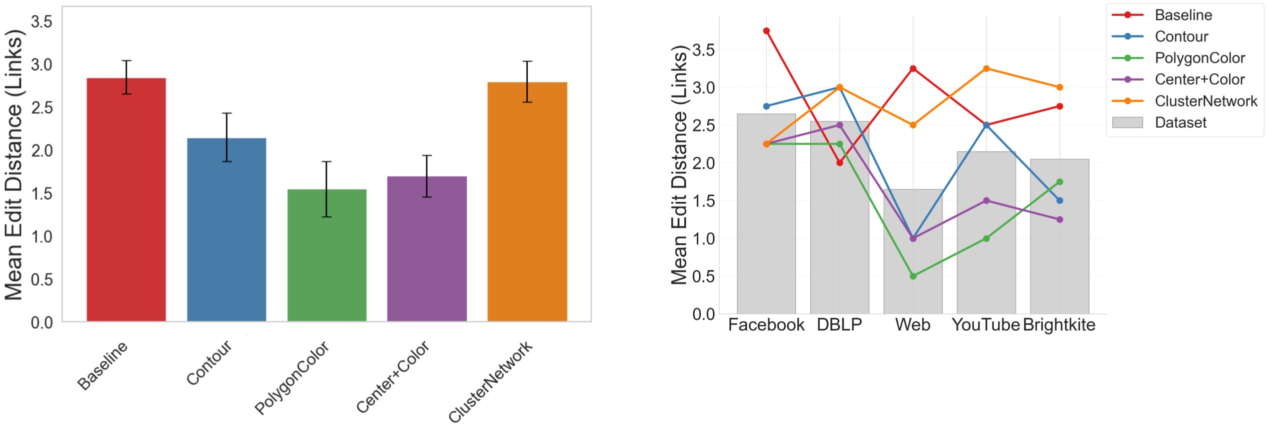


Figure 16: S2 Link ranking: (Left) mean edit distance ± SEM, by Design; (Right) mean edit distance ± SEM, by Dataset and Design.

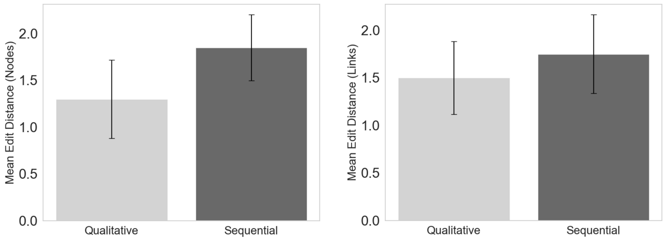


Figure 17: S2 Ranking: Mean edit distances for nodes ± SEM (Left) and links (Right), by ColorMap.

S2: Measure	**DF (n,d)**	**F**	***p***	η^2
Edit Distance – Node Count	1,38	2.10	0.16	–
Edit Distance – Link Count	1,38	0.39	0.54	–
Completion Time	1,38	1.66	0.21	–

Table 7: S2: Factorial analysis of accuracy and completion time for factor ColorMap (ranking tasks)

Therefore, we do not carry out a factorial analysis, but rather report the degree of agreement (similar to the approach used for the Identification task of Study 1).

We calculated agreement rate by comparing the clusters chosen by each pair of participants within each design and dataset. We counted the number of times the same cluster was chosen, and divided by the number of pairs to get an agreement rate. We then calculated the average for each Design (collapsing across Datasets); these results are shown in Figure 22. As can be seen in the figure, the agreement rate for `Contour` (36.7%) was substantially lower than that of `Baseline` and `PolygonColor` (both 56.7%), with `Center+Color` in between at (46.7%).

5.2.5 S2 Subjective Measures. As in Study 1, we gathered two types of subjective measures. First, participants rated each Design based on how well it supported four analyses: identifying good clusters, estimating cluster size based on node count, estimating size based on link count, and estimating between-cluster links. Figure 23 summarizes participants' ratings. As shown in the figure, participants again gave better ratings to designs with a higher degree of summarization (the `ClusterNetwork` or `Center+Color` visualizations were consistently ranked best).

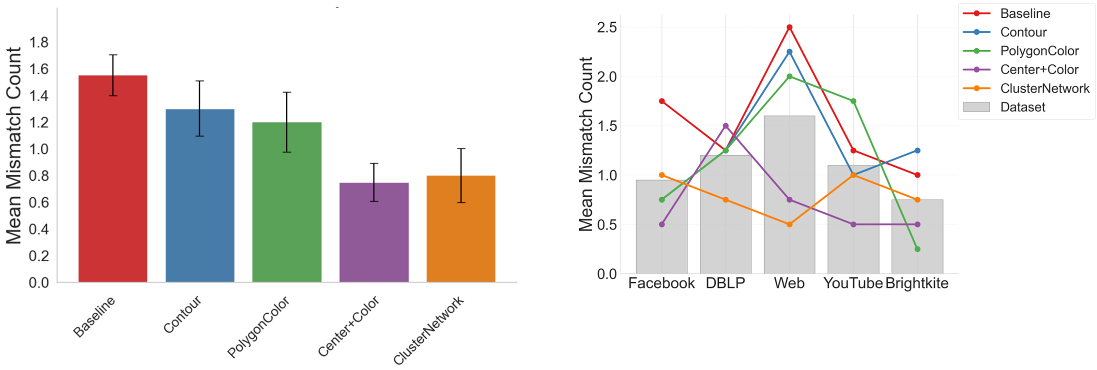


Figure 18: S2 Accuracy for Search task: Mean mismatch ± SEM, by Design (left) and by Dataset and Design (right).

S2: Factor	DF (n,d)	F	p	η^2	Pairwise Contrasts (mean), t-test result
			Mismatch Count		
Design	4,75	3.90	**p < .01**	0.121	Center+Color (0.75) < Baseline (1.55), **p < .05** ClusterNetwork (0.80) < Baseline (1.55), **p < .05**
Design × Dataset	16,75	1.68	0.07	-	
			Completion Time		
Design	4,75	4.22	**p<.01**	0.146	ClusterNetwork (1.48) < Baseline (2.82)
Design × Dataset	16,75	1.00	0.46	-	

Table 8: S2: Factorial Analysis of Accuracy and Completion Time, Search Task.

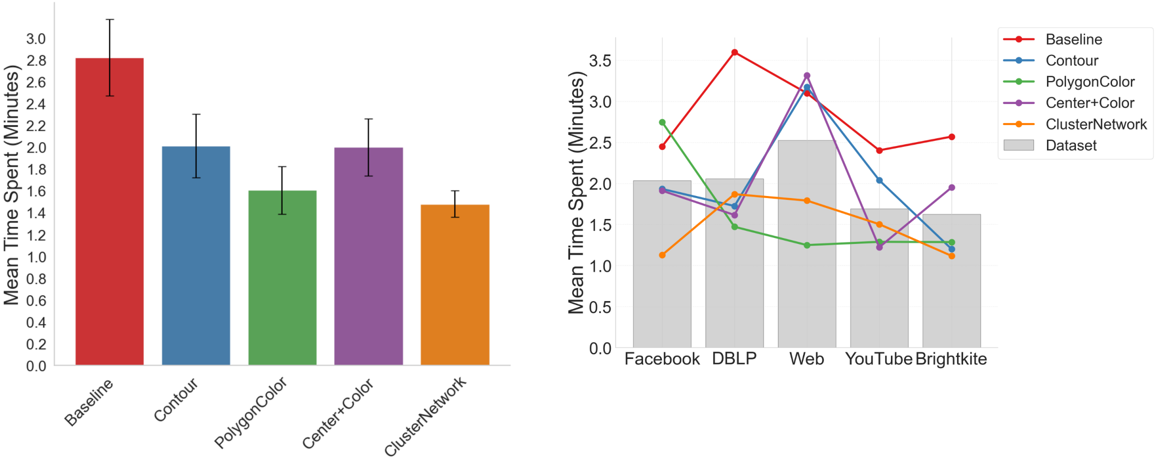


Figure 19: S2 Search Task: Completion time ± SEM, by Design (left) and by Dataset and Design (right).

S2: Factor	DF (n,d)	F	p	η^2
Mismatch Count	1,38	0.84	0.37	–
Completion Time	1,38	4.22	**p < .05**	0.10

Table 9: S2: Factorial Analysis of accuracy and completion time for factor ColorMap (search task)

Second, after each task, participants ranked the Designs in terms of speed, ease of use, and preference; they also completed a workload assessment using a TLX-style survey [13]. Figures 24 and 25 summarize these results.

S2: Factor	DF (n,d)	F	p	η^2	Pairwise Contrasts (mean), t-test result
Score					
Design	4,75	4.28	**p < .01**	0.160	`Center+Color` (4.0) > `Contour` (2.8), **p < .05**
Design × Dataset	16,75	0.95	0.51	-	
Completion Time					
Design	4,75	1.14	0.24	-	
Design × Dataset	16,75	1.36	0.19	-	

Table 10: S2: Factorial Analysis of Accuracy and Completion Time, Degree-Estimation Task.

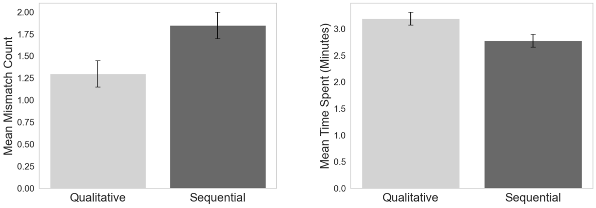


Figure 20: S2 Search: (Left) mean mismatch ± SEM, by ColorMap; (Right) mean time completion ± SEM, by ColorMap.

S2: Factor	DF (n,d)	F	p	η^2
Score	1,38	0.06	0.81	–
Completion Time	1,38	0.04	0.84	-

Table 11: S2: Factorial Analysis of accuracy and completion time for factor ColorMap (Degree-Estimation task).

6 OVERALL DISCUSSION

Here we compare and consider the main results from Study 1 and Study 2, provide design guidelines based on our findings, and outline limitations of the studies that lead to future work.

6.1 Summary of Main Study Results

Performance of* `Baseline` *compared with summary designs. Several of the tasks and measures in both studies showed significant effects of Design, with `Baseline` often performing worse than the summary representations (particularly the most granular designs `ClusterNetwork` and `Center+Color`). For example, `ClusterNetwork` outperformed `Baseline` and `Contour` in ranking clusters by node count (S1), and had better accuracy and completion time than `Baseline` in the Search task (S2). However, other tasks did not show any differences between designs (e.g., no effect of Design on accuracy in either the link-count ranking of S1 or the Search task in either S1 or S2). In addition, other tasks showed effects of specific designs – e.g., the two color-based designs (`PolygonColor` and `Center+Color`) were more accurate than `Baseline` for link-count ranking in Study 2. We also observed that `PolygonColor` and `Center+Color` had lower edit distances in Study 2 than in Study 1 for both node and link counting, which suggests that the design refinements identified after Study 1 (see Section 4.4), such as the importance of a careful binning strategy using a geometric series, were effective.

Does between-cluster link summarization achieve better performances in the Search task? Although Study 1 only found significant main effect of Design on task completion time for the Search task, Study 2 showed significance on both accuracy and task completion time with `Baseline` being significantly less accurate than `Center+Color` and `ClusterNetwork`. This strengthens the support for ($h5$) where we expected link summarization to be effective in searching for cluster pairs (we also believe this to be an outcome of the design refinements implemented for Study 2).

Does the effectiveness of visual summaries depend on how visible the clusters are in the* `Baseline` *representation? When ranking by node count, we observed a significant interaction between Design and Datasets in both Studies 1 and 2. Moreover, Study 2 showed the `Center+Color` to be significantly more accurate than `Baseline` in Web Dataset for the pair-search task. For the degree-estimation task, we observed `Center+Color` to be more accurate than `Contour` in the Facebook Dataset.

Performance differences across two colormaps. We did not find accuracy differences for colormaps in either Study 1 or Study 2. We believe the improved color binning strategy offered enough variation such that the two maps performed similarly. While Study 1 did not show any significant differences when considering completion time for both ranking and search tasks, Study 2 showed colormaps to have a significant effect on the search task (with sequential color being faster). This further suggests that increased granularity in color bins (variation in color) contributes to faster task completion.

Did user performance differ across the Designs for more detailed tasks? The Subcluster-Estimation task was a fine-grained task where participants had to assess and count small cluster structures within a bigger cluster. Although we used a large 75-inch display, this task appeared to be difficult. We already observed in Study 1 that participants may have low agreement in cluster identification if the clusters are not clearly visible in the node-link visualization. This issue becomes even more prominent when identifying subclusters, as they are often less distinct within a cluster, supporting ($h6$).

Did user preferences vary across the Designs? In both studies, users strongly preferred more coarse-grained visual summaries for all information types, with `ClusterNetwork` rated as the most efficient and `Baseline` as the least. This does not always align with participants' performance, however: for example, `ClusterNetwork` in Study 2 had low mean accuracy for ranking links. We believe this is a result of not applying the improved color binning strategy for `ClusterNetwork`.

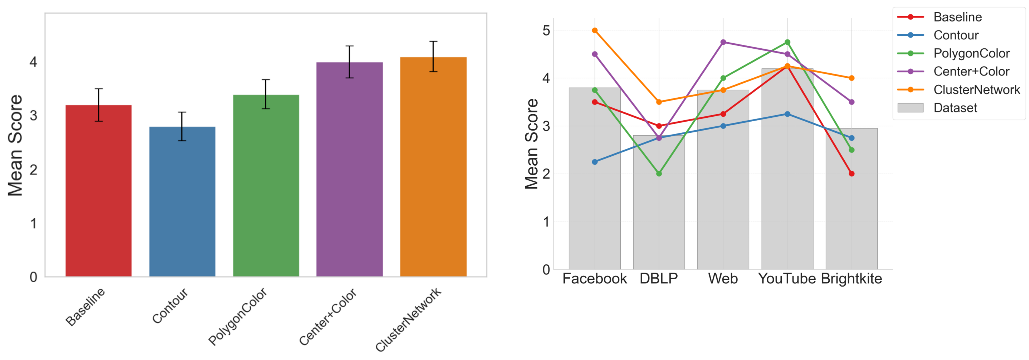


Figure 21: Accuracy for Degree-Estimation task. S2: (Left) mean score ± SEM, by Design; (Right) mean score ± SEM, by Design and Dataset.

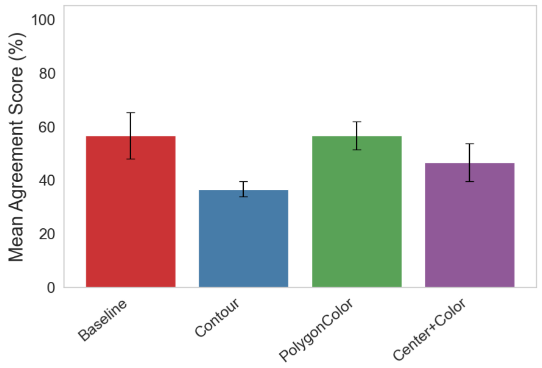


Figure 22: S2 Subcluster Estimation: Agreement Rate, by Design. Note that ClusterNetwork was not used in this task.

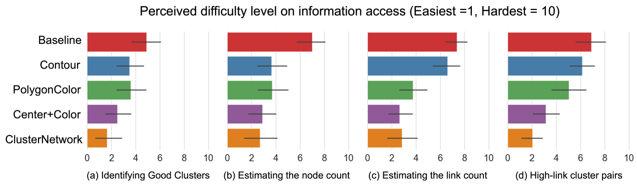


Figure 23: S2: Mean perceived difficulty for information access (± SEM), by Design. Shorter bars are better.

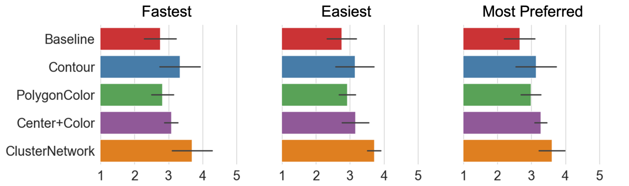


Figure 24: S2: Subjective Question Responses (Preference) (± SEM), by Design. Longer bars are better.

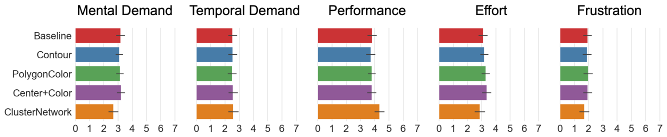


Figure 25: S2: Subjective Question Responses (Workload) (± SEM), by Design. Shorter bars are better (except for Performance).

6.2 Design Guidelines

Our studies reveal some key factors that should be carefully considered for designing node-link representations of large networks.

Providing Visual Summaries: To improve the readability of the FA layout, one should consider providing supplementary summaries. These summaries can help users identify clusters more easily, estimate their sizes, and understand the pairwise relationships between them. The summaries should be sufficiently different to complement each other. The summaries may be created by identifying clusters algorithmically, where a spatial clustering algorithm could be used to find clusters from an FA layout. Since edges play an important role in how people perceive a tangled structure, such a clustering algorithm should consider both the node and edge locations to achieve a better agreement between the algorithmically detected and visually identifiable clusters.

Designing ColorMaps to Capture Cluster Variation: Proper attention should be paid to designing colormaps based on data distributions. For example, if the largest clusters are chosen to create a summary, these clusters will naturally have a large number of nodes and edges. Using uniform size thresholds for coloring may result in many clusters being rendered in the same color. Instead, uniform percentile-based thresholds or geometric-series-based percentiles can be more effective in differentiating the sizes of clusters.

Choosing Granularity in Color Matrices: When using a color matrix to map cluster sizes to colors, selecting a more granular option (e.g., 5 × 5) can enhance the ability to distinguish between different cluster sizes and accomplish size estimation tasks more efficiently compared to a less granular option (e.g., 3 × 3). For a more granular matrix, a sequential colormap is preferable to a qualitative colormap, as it helps reduce the number of different hues while naturally maps darker shades to larger cluster sizes.

Considering Differences in Datasets: Depending on the network size and edge density within clusters, the same opacity for rendering may not be equally effective for all networks. This requires tuning by the visualization designer so that nodes and edges are rendered with appropriate opacity to ensure better visual exposure of the clusters. In datasets where node distributions are mostly uniform and distinct clusters are challenging to identify visually, careful consideration is needed when using a spatial clustering algorithm to automatically detect clusters. For example, the algorithm's parameters should be chosen carefully by experimenting with various configurations.

Choosing Summaries Based on Information Needs: The desired summaries may vary depending on user needs. For example, if understanding node distribution is the priority, then `Contour` could be the preferred option. Alternatively, if estimating cluster sizes is more important, then `ClusterNetwork` is more suitable. The `Center+Color` design serves as a middle ground between the two, revealing pairwise relationships between clusters while providing some visibility of node distributions.

Improving Readability through Consistent Labeling: When visualizing one or more summaries, using clear cluster labels for the FA layout and its summaries, and a consistent color scheme across summaries can reduce the risk of misinterpretation. In interactive settings, all the visualization views can be information linked such that hovering over a cluster in one view highlights the corresponding cluster in all other views.

6.3 Limitations and Future Work

There are several limitations to our studies, each of which provides an opportunity for further research.

First, although our studies involved real-world datasets and realistic analytics tasks, the analysis scenario was artificial. In future studies, we will identify datasets and tasks that have been used in existing real-world presentations, and determine whether our findings are consistent in these scenarios.

Second, our in-person studies involved only 20 participants per study, and the participants were all university students; therefore, a crowdsourced experiment could be useful in gaining insights into the perception of a more general audience. We also limited our focus to node-link visualizations, whereas edge bundling is frequently employed to reduce visual clutter; therefore, conducting an in-depth exploration of edge bundling presents a promising direction for future research.

Third, our study uses FA visualizations which are often generated by parameter tuning. Hence the exposure of the clusters depends on the choice of parameters being used, which may vary depending on the dataset. However, we believe that our results will generalize to other datasets because our study included a diverse set of node-link visualizations where clusters appeared in many different forms.

Fourth, while our study demonstrated the effectiveness of visual summaries in aiding cluster interpretation, the subcluster-estimation task was particularly tedious for participants. This observation suggests that future approaches might benefit from rethinking how such complex tasks are supported—possibly through more interactive or adaptive techniques that guide users progressively, rather than expecting full-detail interpretation upfront. Incorporating interaction, such as zoomable details or on-demand expansion of visual elements, could improve both usability and engagement in similar tasks.

Fifth, the participants in our study did not report any color vision deficiencies, but color perception varies widely across individuals, and we plan to further investigate how color perception might affect user performance with complex visualizations. For example, individuals have different abilities in differentiating colors, which could affect the choice of color-binning strategy. In addition, overly-complex or high-contrast color schemes may alter the user's cognitive load and make it harder for some users to focus on the task. Future studies could benefit from incorporating accessible palettes, or calibrating user perceptual abilities and then adapting color palette to the capabilities of the individual.

7 CONCLUSION

Network datasets generated from real-world contexts often contain millions of nodes and edges, and it can be difficult to interpret the overall structure of these networks. To explore ways of improving interpretation of clusters with large networks, we designed new visualizations and tested them in two user studies. We showed that user interpretation can vary substantially when reading large node-link visualization, especially when cluster structures are difficult to discern; these results show the potential value of providing a visual summary alongside the baseline visualization. We designed four visual summaries with varying levels of granularity, and tested the designes in two user studies. The studies showed strong user preference for more coarse-grained summaries; we also observed that the effectiveness of different summaries varied by task type, and performance often improved as the summaries increased the degree of summarization compared to the baseline node-link representation. Our findings suggest that large node-link visualizations can improve interpretation by providing a summary representation that complements the original information.

8 ACKNOWLEDGMENTS

The work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, Vol. 3. 361–362.

[2] Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media, Inc. https://www.nltk.org/

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[4] Govert G Brinkmann, Kristian FD Rietveld, and Frank W Takes. 2017. Exploiting gpus for fast force-directed visualization of large-scale networks. In *2017 46th International Conference on Parallel Processing (ICPP)*. IEEE, 382–391.

[5] Ariyawat Chonbodeechalermroong and Rattikorn Hewett. 2017. Towards visualizing big data with large-scale edge constraint graph drawing. *Big Data Research* 10 (2017), 21–32.

[6] Gabor Csardi and Tamas Nepusz. 2006. The igraph software. *Complex syst* 1695 (2006), 1–9.

[7] Herbert Edelsbrunner. 2011. Alpha shapes-a survey. In *Tessellations in the sciences: Virtues, techniques and applications of geometric tilings.*

[8] Niklas Elmqvist, Thanh-Nghi Do, Howard Goodell, Nathalie Henry, and Jean-Daniel Fekete. 2008. ZAME: Interactive large-scale graph visualization. In *2008 IEEE Pacific visualization symposium*. IEEE, 215–222.

[9] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck. 2010. ManyNets: an interface for multiple network analysis and visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 213–222.

[10] Emden R Gansner, Yifan Hu, and Stephen G Kobourov. 2010. Gmap: Drawing graphs as maps. In *Graph Drawing: 17th International Symposium, GD 2009, Chicago, IL, USA, September 22-25, 2009. Revised Papers 17*. Springer, 405–407.

[11] Emden R Gansner, Yehuda Koren, and Stephen C North. 2005. Topological fisheye views for visualizing large graphs. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 457–468.

[12] Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. 2004. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE symposium on information visualization*. IEEE, 17–24.

[13] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[14] Yifan Hu and Lei Shi. 2015. Visualizing large graphs. *Wiley Interdisciplinary Reviews: Computational Statistics* 7, 2 (2015), 115–136.

[15] Weidong Huang. 2007. Using eye tracking to investigate graph layout effects. In *2007 6th International Asia-Pacific Symposium on Visualization*. IEEE, 97–100.

[16] Zhenhua Huang, Junxian Wu, Wentao Zhu, Zhenyu Wang, Sharad Mehrotra, and Yangyang Zhao. 2021. Visualizing complex networks by leveraging community structures. *Physica A: Statistical Mechanics and its Applications* 565 (2021), 125506.

[17] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One* 9, 6 (2014), e98679.

[18] Radu Jianu, Adrian Rusu, Yifan Hu, and Douglas Taggart. 2014. How to display group information on node-link diagrams: An evaluation. *IEEE Transactions on Visualization and Computer Graphics* 20, 11 (2014), 1530–1541.

[19] Sanjay Kairam, Diana MacLean, Manolis Savva, and Jeffrey Heer. 2012. Graphprism: compact visualization of network structure. In *Proceedings of the international working conference on advanced visual interfaces*. 498–505.

[20] Ethan Kerzner, Alexander Lex, Crystal Lynn Sigulinsky, Timothy Urness, Bryan W Jones, Robert E Marc, and Miriah Meyer. 2017. Graffinity: Visualizing connectivity in large graphs. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 251–260.

[21] Bongshin Lee, Catherine Plaisant, Cynthia Sims Parr, Jean-Daniel Fekete, and Nathalie Henry. 2006. Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. 1–5.

[22] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

[23] Timothy R. Levine and Craig R. Hullett. 2002. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research* 28, 4 (2002), 612–625.

[24] Sungsu Lim, Junghoon Kim, and Jae-Gil Lee. 2016. BlackHole: Robust community detection inspired by graph drawing. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 25–36.

[25] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.

[26] Debajyoti Mondal and Lev Nachmanson. 2018. A New Approach to GraphMaps, a System Browsing Large Graphs as Interactive Maps. In *VISIGRAPP - Volume 3: IVAPP*, Alexandru C. Telea, Andreas Kerren, and José Braz (Eds.). SciTePress, 108–119.

[27] Ehsan Moradi and Debajyoti Mondal. 2023. BigGraphVis: Visualizing Communities in Big Graphs Leveraging GPU-Accelerated Streaming Algorithms.. In *VISIGRAPP (3: IVAPP)*. 195–202.

[28] Lev Nachmanson, Roman Prutkin, Bongshin Lee, Nathalie Henry Riche, Alexander E Holroyd, and Xiaoji Chen. 2015. Graphmaps: Browsing large graphs as interactive maps. In *Graph Drawing and Network Visualization: 23rd International Symposium, GD 2015, Los Angeles, CA, USA, September 24-26, 2015, Revised Selected Papers 23*. Springer, 3–15.

[29] Mershack Okoe, Radu Jianu, and Stephen Kobourov. 2018. Node-link or adjacency matrices: Old question, new insights. *IEEE transactions on visualization and computer graphics* 25, 10 (2018), 2940–2952.

[30] Stephen Olejnik and James Algina. 2003. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 8, 4 (2003), 434–447.

[31] Mathias Pohl, Markus Schmitt, and Stephan Diehl. 2009. Comparing the Readability of Graph Layouts using Eyetracking and Task-oriented Analysis.. In *CAe*. 49–56.

[32] Aaron Quigley and Peter Eades. 2000. Fade: Graph drawing, clustering, and visual abstraction. In *International Symposium on Graph Drawing*. Springer, 197–210.

[33] Bahador Saket, Paolo Simonetto, Stephen Kobourov, and Katy Börner. 2014. Node, node-link, and node-link-group diagrams: An evaluation. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2231–2240.

[34] SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[35] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13, 11 (2003), 2498–2504.

[36] Natalie Stanley, Roland Kwitt, Marc Niethammer, and Peter J Mucha. 2018. Compressing networks with super nodes. *Scientific reports* 8, 1 (2018), 10892.

[37] Georgianna Strode, John Derek Morgan, Benjamin Thornton, Victor Mesev, Evan Rau, Sean Shortes, and Nathan Johnson. 2019. Operationalizing Trumbo's principles of bivariate choropleth map design. *Cartographic Perspectives* 94 (2019), 5–24.

[38] Martin Wattenberg. 2006. Visual exploration of multivariate graphs. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 811–819.

[39] Yanhong Wu, Wenbin Wu, Sixiao Yang, Youliang Yan, and Huamin Qu. 2015. Interactive visual summary of major communities in a large network. In *2015 IEEE pacific visualization symposium (PacificVis)*. IEEE, 47–54.

[40] Vahan Yoghourdjian, Tim Dwyer, Karsten Klein, Kim Marriott, and Michael Wybrow. 2018. Graph thumbnails: Identifying and comparing multiple graphs at a glance. *IEEE Transactions on Visualization and Computer Graphics* 24, 12 (2018), 3081–3095.

[41] Gazi Md Hasnat Zahan, Debajyoti Mondal, and Carl Gutwin. 2021. Contour line stylization to visualize multivariate information. In *Graphics Interface 2021*.

[42] Michael Zinsmaier, Ulrik Brandes, Oliver Deussen, and Hendrik Strobelt. 2012. Interactive level-of-detail rendering of large graphs. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2486–2495.