GraphPrism: Compact Visualization of Network Structure

Sanjay Kairam, Diana MacLean, Manolis Savva, Jeffrey Heer
Stanford University Computer Science Department
{skairam, malcdi, msavva, jheer}@cs.stanford.edu

ABSTRACT

Visual methods for supporting the characterization, comparison, and classification of large networks remain an open challenge. Ideally, such techniques should surface useful structural features – such as effective diameter, small-world properties, and structural holes – not always apparent from either summary statistics or typical network visualizations. In this paper, we present GraphPrism, a technique for visually summarizing arbitrarily large graphs through combinations of 'facets', each corresponding to a single node- or edge-specific metric (e.g., transitivity). We describe a generalized approach for constructing facets by calculating distributions of graph metrics over increasingly large local neighborhoods and representing these as a stacked multi-scale histogram. Evaluation with paper prototypes shows that, with minimal training, static GraphPrism diagrams can aid network analysis experts in performing basic analysis tasks with network data. Finally, we contribute the design of an interactive system using linked selection between GraphPrism overviews and node-link detail views. Using a case study of data from a co-authorship network, we illustrate how GraphPrism facilitates interactive exploration of network data.

Categories and Subject Descriptors

I.6.9 [**Computing Methodologies**]: Visualization—*Visualization techniques and methodologies*

Keywords

Network analysis, graph visualization, scalability

1. INTRODUCTION

The size of available network datasets is growing considerably, already exceeding millions of nodes and edges [25]. Many existing graph visualization methods such as node-link diagrams and adjacency matrices were developed in the context of smaller graphs. For larger networks these techniques break down, making them unsuitable for supporting inference about higher-level network properties.

In this paper, we make two contributions towards the problem of scaling graph visualization to larger networks. First, we present GraphPrism, a visualization technique for compactly summarizing networks of arbitrary size. GraphPrism views combine multiple 'facets', each providing a statistical summary of the graph with respect to a single node- or edge-specific metric. By computing metrics over graph neighborhoods of increasing size, GraphPrism generalizes Bagrow et al.'s [3] *B-Matrix* technique. Paper prototype evaluations with network analysis experts demonstrate that analysts can use GraphPrism diagrams to assess and compare network structures.

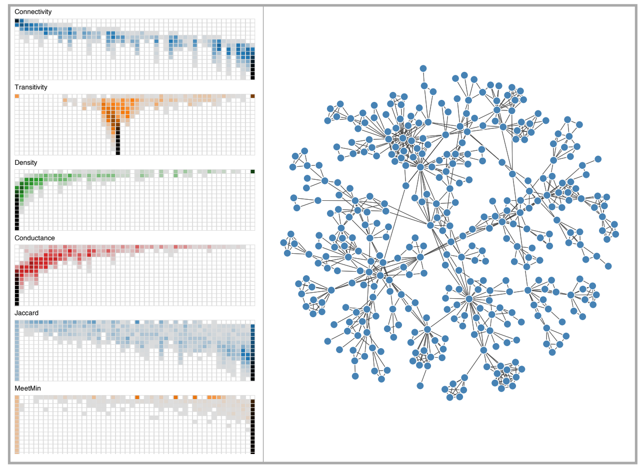

Figure 1: GraphPrism and node-link diagrams for the largest component of a co-authorship graph.

Second, we describe the design of an interactive system for inspecting graphs at multiple levels of detail using linked selection between a GraphPrism overview and a node-link detail view. Through a case study with a network science co-authorship graph (Figure 1), we show how GraphPrism facets function as both visual summaries and dynamic query selectors to investigate patterns ranging from the level of individual nodes and edges up to global network structure.

2. RELATED WORK

We first review relevant work in the area of network analysis and identify properties critical for understanding graph data. We then examine existing graph visualization approaches and discuss how well each conveys these properties.

2.1 Properties of Complex Networks

Our goal with GraphPrism is to surface structural properties of complex networks in a scalable, spatially compact manner. Here, we identify structural properties of networks which are critical in performing common analysis tasks, and thus should ideally be conveyed by network visualizations. Key definitions are summarized in Table 1.

AVI '12, May 21-25, 2012, Capri Island, Italy

Degree. Node degrees for many real networks have been shown to follow a power-law distribution [15]; such patterns can yield insight into mechanisms underlying system growth [4]. Deviations from expected distributions can reveal unpredicted system properties which may affect the structure of the network [25]. Thus, there are significant benefits to visualizing the overall shape of the degree distribution.

Diameter. The diameter of a graph intuitively expresses the graph's 'width'. The presence of specific structures, such as a long isolated chain, can change this value dramatically — even if such structures do not play a significant role in the overall system. A more robust version of this metric is the *effective diameter*, defined as the number of hops separating a given percentage of connected node pairs (commonly 90% [24]). Visualizing the relationship between this choice of threshold and the effective diameter may provide additional insight into patterns of graph connectivity.

Transitivity. Also named *clustering coefficient*, transitivity describes network interconnectedness. Local transitivity around individual nodes can aid reasoning about processes like clique formation [11]. Global transitivity in neural, infrastructure, and social networks (among others) [32] has been shown to be significantly higher than in random networks with equivalent degree distributions [24, 29].

Small-Worlds. Networks with short average path lengths and local clustering of nodes exhibit so-called *small-world effects* [32]. First identified in social networks [28], further study has revealed this phenomenon in many other types of systems, such as biological and information networks [25, 32]. The presence of small-world properties in communication and information networks indicates that these systems can be leveraged for rapid information passing and search [1, 23]. Making such properties evident in a visualization can aid reasoning about global system behavior.

Bridges. We define a *local bridge* as an edge connecting two nodes which otherwise share no immediate neighbors, and a *global bridge* as connecting two otherwise disconnected nodes. Such bridges can represent critical pathways through which resources in a system might travel [12]. *Structural holes* created around these bridges have been shown to foster creativity and productivity in professional information networks by preventing 'echo chambers' which might occur in more heavily clustered networks [13, 14].

Communities. In network analysis, communities are generally defined as collections of nodes which tend to link more with each other than with nodes outside. Though this concept is clearly derived from social networks, meaningful analogues have been discovered in a variety of systems, including biological, information, and citation networks [18].

2.2 Network Visualization

We review prior visualization approaches to see how well they accomplish the described analysis goals.

Node-Link Diagrams and Adjacency Matrices. For small, sparse networks, node-link diagrams provide an effective means of visualizing connectivity. Producing a perceptually effective layout becomes difficult, however, as graph complexity increases [9]. Edge crossings, cluttering, and occlusion problems can hinder inference. Filtering techniques [2] can mitigate scaling problems by limiting the information shown; however, this solution does not translate to static visualization and does not facilitate perception of structural properties at scales beyond local node neighborhoods.

Degree	For a node v in G, the count of edges in G incident on v.
Diameter	The length of the maximum shortest path between any pair of connected nodes u, v in G.
Triad	A triplet of connected nodes. A triad is *closed* if edges connect all three nodes to each other.
Local Transitivity	For some v in G, the fraction of triads in G which are closed for the subgraph defined by v and its immediate neighbors.
Global Transitivity	The fraction of all triads in G which are closed.

Table 1: Common Graph-Theoretic Measures

Adjacency matrices avoid edge-crossing and occlusion problems but are still subject to layout effects; the perception of larger structures such as communities depends on a proper permutation of rows and columns [6]. Algorithms exist for sorting adjacency matrices [16], but the choice of which to use is dependent on task context and prior insight into the expected structure [20]. Like node-link diagrams, the scalability of adjacency matrices is limited due to the bounded resolution of both pixel displays and human visual acuity.

Recent work has attempted to overcome limitations of each of these approaches by combining them [20, 21, 22]. To aid path-following, MatLink [21] augments an adjacency matrix with arcs along row and column headers. NodeTrix [22] starts with a traditional node-link view, but reduces occlusion by representing dense clusters as small adjacency matrices. While these approaches marry benefits of the two visualization types, they remain limited by layout and scaling issues as network size increases and show only a subset of the topological features we hope to convey.

Abstracting Network Properties. To facilitate analysis of larger graphs, successful visualization approaches must abstract the data. PivotGraph [31] and Honeycomb [19] both aggregate graph data based on node attributes. PivotGraph collapses nodes with particular attribute values into a single super-node, allowing viewers to easily see how edges connect nodes of different types. Honeycomb aggregates on the basis of node hierarchy and visualizes customizable metrics to surface interesting network properties at multiple levels of resolution. The requirement of node metadata limit these approaches, however, making them unsuitable for many instances of general graph data.

ManyNets [17] enables visual comparison of multiple networks through a tabular display. Table cells contain either individual descriptive statistics or visualizations of distributions. This approach is scalable and can surface many of the properties described in §2.1. Though useful for comparisons of network data sets, simple descriptive statistics still fail to reveal much of the structural insight needed for rich characterization of graph data. ManyNets adapts to this challenge by allowing analysts to connect back to a node-link diagram, an approach that we also adopt in the design of our interactive system. Our aim with GraphPrism is to improve on prior work by crafting a visualization technique which scales effectively while giving a richer view of network structure.

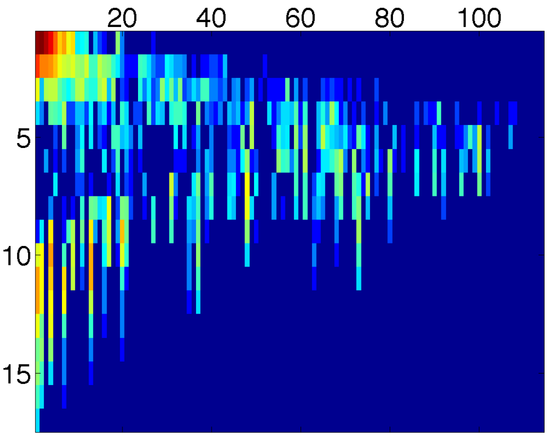

Figure 2: A B-Matrix of the co-author network in Fig. 1. Path length ℓ increases from top to bottom and number of nodes k increases from left to right.

3. DESIGNING GRAPHPRISM

GraphPrism extends the *B-Matrix* approach to visualizing graph connectivity [3]. We first modify B-Matrices to accommodate additional metrics and then describe visual encodings intended to improve graphical perception.

3.1 B-Matrices

The GraphPrism technique was inspired by Bagrow's B-Matrix, which visually presents graph connectivity patterns. The B-Matrix (Figure 2) is a two-dimensional matrix where each cell $B_{\ell,k}$ represents the number of nodes which can reach k other nodes in exactly ℓ hops (k increases from left to right, ℓ from top to bottom). Using color to encode cell values, the resulting image forms an abstract 'portrait' of the network that reveals connectivity patterns.

As this representation does not depend on a sorting or labeling of individual nodes, it is invariant for all isomorphs of a graph. Each row ℓ of the B-Matrix is a histogram of paths whose length is exactly ℓ. The first row ($\ell = 1$) is the degree distribution. Together, these distributions concisely capture many structural properties of the graph, such as diameter and dimensionality, and to some extent, global properties such as small-world behavior.

Bagrow et al. describe the utility of the B-Matrix for characterizing and categorizing networks using both empirical and simulated data. However, their representation fails to convey some of the properties that we identified as highly relevant to network analysis, such as transitivity and the presence of bridges. Moreover, aspects of the visual design could be improved to facilitate graphical perception.

3.2 Generalization to Other Metrics

We have generalized the B-Matrix approach to depict other node- and edge-specific metrics. We chose our current set of metrics to create GraphPrisms for analyzing unweighted directed and undirected graphs; additional metrics may be suitable for other types of graphs.

3.2.1 Node-Specific Metrics

For any node-specific metric, we aim to surface changes as we shift from a local (immediate node neighborhood) to a global (entire network) perspective. To characterize intermediate levels, we define the ℓ-level neighborhood for a node $v \in V(G)$ (where V is the set of nodes in G) as: $N_\ell(v) = \{U : u \in V, distance(u, v) \leq \ell\}$. We compute each node-specific metric over ℓ-neighborhood levels starting with each individual node and expanding outwards. At some level ℓ_n ($n \leq diameter(G)$), the node neighborhood will comprise the entire graph and the metric distribution will converge such that $\forall \ell \geq \ell_n$, the values will remain the same.

Connectivity. This first metric is simply a variation of that used in the B-Matrix. Rather than representing the number of nodes exactly ℓ hops away from a given node v (as in the B-Matrix), we use $|N_\ell(v)|$, or the number of nodes $\leq \ell$ hops away. In other words, each row depicts the cumulative distribution thus far. This metric is suited to undirected networks; we can use alternative versions for directed networks. We define *in-connectivity* to be the number of nodes which can reach v within ℓ hops using directed edges and *out-connectivity* to be the number of nodes which v can reach within ℓ hops.

Transitivity. At $\ell = 1$, transitivity can vary from 0, indicating the presence of chains or structural bridges, to 1, indicating that all local triads are closed. We extend this notion to larger neighborhood levels ($\ell \geq 1$) by computing the global transitivity of the subgraph defined by $N_\ell(v)$. For a fully-connected graph, this distribution will converge to a single value for all nodes at some $\ell \leq diameter(G)$. For graphs with multiple components, the distributions will converge to a single value for each connected component.

Conductance. The conductance of a subgraph $H \in G$ measures how 'community-like' H is. It is defined as:

$$conductance(H) = 1 - \frac{2 * |E(H)|}{\sum_{v \in H} k_v}$$

where $|E(H)|$ represents the number of edges in the subgraph H and k_v represents the degree of node v in the original graph G. Thus, if H consists of a set of nodes which link to many other nodes, but not each other, the conductance of H will be 1. If H is an isolated clique, then its conductance will be 0. At each level ℓ, we compute this measure on the subgraphs defined by the ℓ-neighborhoods of each node. If a node is roughly at the center of a well-defined community, we may see a sharp decrease in the conductance when the ℓ-neighborhood roughly overlaps. Though not an exact method for finding communities, this method requires calculating conductance only for a limited set of node combinations. The 'ideal' approach of calculating conductance on all possible node subsets and choosing the minimum will find better communities but is computationally intractable for large graphs [24].

Density. The density of a graph is simply the ratio of the number of edges to the total number possible. We compute this metric in a manner similar to the transitivity, where we calculate the density over the subgraph defined by the ℓ-neighborhood of a particular node. This gives us an alternative view into local clustering around nodes.

3.2.2 Edge-Specific Metrics

GraphPrism can be similarly used to describe graph edges. To evaluate the potential utility of visualizing edge metrics, we construct two closely-related versions of metrics designed to identify how 'redundant' a particular edge is in the structure of the surrounding graph. We calculate this generally by measuring the overlap between the neighborhoods of the two nodes incident to the edge when the edge is removed.

Jaccard and **MeetMin.** The Jaccard similarity of two node neighborhoods is the ratio of the size of the intersection over the size of the union. Nodes which share many common neighbors will have Jaccard close to 1, while those which are disconnected aside from their connecting edge will have Jaccard close to 0. Melançon et al. [27] describe a method for extending this metric to neighborhoods of arbitrary size which operates analogously to the approach we use for node-specific metrics. MeetMin is defined similarly as the ratio of the size of the intersection of the two neighborhoods over the size of the smaller neighborhood. Cases in which edges connect nodes with neighborhoods of dissimilar size may have higher MeetMin values than Jaccard, and thus this metric may surface different patterns.

3.3 Visual Encoding Choices

In designing GraphPrism, we evaluated the visual encoding choices made for the B-Matrix and modified them in order to better support perceptual inference.

Stacked Histograms. Preserved from the B-Matrix approach is the use of stacked histograms. Values for each level ℓ are binned into histograms and stacked to form a heatmap that facilitates both row comparisons (distributions for a single level) and column comparisons (changes between levels). This design is similar in this regard to Bertin's matrices [7] and aligns with Mackinlay's recommendations for composing data along a single axis [26].

Color Encoding. As shown in Figure 2, B-Matrices encode histogram values using hue, which may not be optimal for encoding values on a continuous scale [8, 26]. We instead use color intensity to represent histogram values, more accurately enabling quantitative inferences about the values of bins. To enable comparisons between networks of different sizes, we map histogram values to the interval $[0, 1]$ and use a logarithmic scale to map color intensity to normalized value (with black representing 1).

3.4 Example: Zachary's Karate Club

Through a simple example familiar to the network analysis community, we now show how GraphPrism can illustrate features of real networks. Zachary's 'Karate Club' describes friendships among 34 members of a karate club at US university in the 1970s [34]; it is a common example of community fission, as the club later split into two separate groups. For simplicity, we discuss just three facets: node-centric *Connectivity* and *Transitivity*, and edge-centric *Jaccard Overlap* (abbreviated as *Jaccard*). Figure 3 shows the node-link and GraphPrism diagrams for this network, through which we can observe the following properties:

Degree. The white top-left square of the *Connectivity* facet (1) indicates that the group has no isolates, while the two leftmost blocks in the top row (2) show that most members ($\sim$ 70%) are friends with less than 15% of the group. Similarly, the rightmost block in that row (3) shows that $<$ 5% of members are friends with at least half of the group. In contrast to the low connectivity in $\ell = 1$, looking at $\ell = 2$ (4) shows us that most nodes can reach at least half of the other nodes in the graph within 2 hops (friends of friends).

Diameter. The *Connectivity* facet shows that most nodes can reach each other within 4 hops, thus conveying the effective diameter (5). By $\ell = 5$, the degree distribution converges to the exact diameter. The facet converges to a single value, showing that the graph consists of a single connected component. If, for instance, the largest connected component comprised 80% of the nodes, the *Connectivity* facet would converge at multiple values, the largest being 0.8.

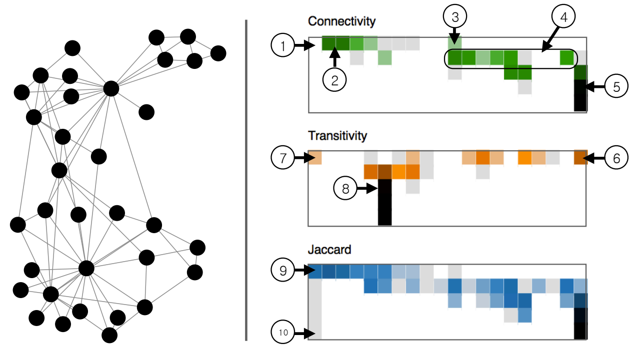

Figure 3: Node-Link and GraphPrism Diagrams for the Zachary's Karate Club Network.

Transitivity. The rightmost square of the *Transitivity* facet (6) indicates that $\sim$ 30% of nodes have transitivity close to 1 (i.e., these members' friends typically are also friends of each other). The leftmost square (7) shows that $<$ 5% of nodes have transitivity at or near 0. This distribution suggests the presence of multiple natural clusters (adding the *Conductance* facet could help to verify this). As ℓ increases and neighborhoods overlap more, the spread of calculated ℓ-transitivity values shrinks. By $\ell = 4$, values converge to the global transitivity for the graph, ~ 0.25. Again, if the graph had multiple connected components, the facet would converge to multiple values showing the global transitivities for those different components.

Bridges. Focusing on the leftmost column of the *Jaccard* facet, we find that for $\ell = 1$, about ($\sim$ 30%) of edges connect nodes which have no neighbors in common, making them local bridges (9). Values in the bottom left corner (10) indicate the presence of *global* bridges; in the Karate network, we see that there is only one (a leaf node). The wide spread in this metric in the first level ($\ell = 1$) also reveals diverse local structure in the graph.

4. PAPER PROTOTYPE EVALUATION

To assess whether the abstract representation provided by GraphPrism is legible and useful to analysts, we engaged in a paper prototype evaluation with 7 network analysis experts. We had participants complete two tasks involving comparison and classification of network data and asked open-ended feedback questions about how the visualization might be best integrated into interactive analysis systems. Diagrams used in both study tasks contained the three facets discussed above: *Connectivity*, *Transitivity*, and *Jaccard Overlap*.

4.1 Participants

Seven academic researchers (6 Male, 1 Female), all of whom had actively published in the field of network analysis (5 CS, 2 Sociology), each participated in a 1-hour evaluation and interview. The first 15 minutes consisted of a structured tutorial, with a walkthrough of metrics used, explanation of how facets were constructed, and examples of how the diagram could be used to characterize sample graphs. Tasks were administered over the next 30 minutes, and the interview was given in the final 15 minutes of the session.

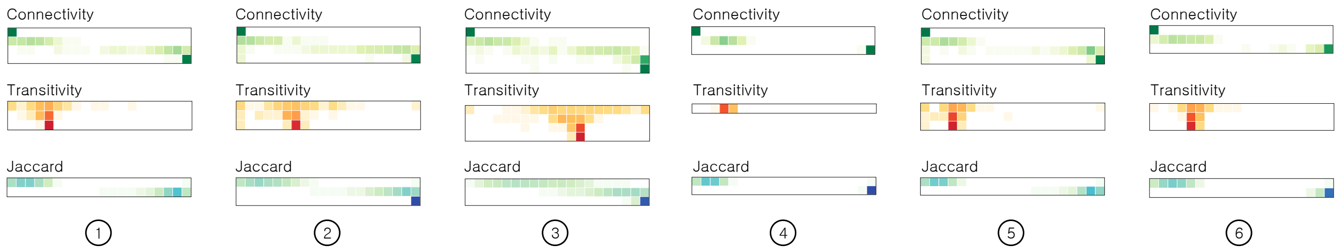

Figure 4: An example of a line-up task (Task 1) completed by participants. One of the diagrams above corresponds to the *Jazz* network and the others to random models (see footnote for answer).

Network	$\lvert V\rvert$	$\lvert E\rvert$	Comps.	Diam.	Trans.
Airlines	332	2126	1	6	0.75
Jazz	198	5484	1	6	0.63
Networks	1589	2742	396	17	0.87
Yeast	2361	7182	101	11	0.2

Table 2: Summary statistics given to participants for Task 1: network characterization.

Network	Airlines	Jazz	NetSci	Yeast	All
N	4	5	5	4	18
Accuracy	0.5	0.8	1.0	0.75	0.78
Confidence	3.75	4.6	4.8	4.0	4.33
Ease	4.0	4.2	4.0	4.0	4.05

Table 3: Task 1 network characterization results.

4.2 Tasks

We designed Task 1 to assess how well analysts could characterize network data sets using GraphPrism. We employed a methodology similar to Wickham et al.'s *line-up* [33] for testing the "inferential validity" of a visualization technique. Specifically, we asked analysts to choose a real network data set from a line up with synthetic data using the GraphPrism diagrams. We utilized four data sets from prior work [5, 11, 30], summarized in Table 2. In each trial, participants viewed ten GraphPrism diagrams, one corresponding to the real-world data, and nine constructed from synthetically generated data. Synthetic data sets were sampled using a variety of random graph models in order to increase task difficulty. Trials were presented in a random order.

As we did not expect participants to be familiar with the data sets, they were given summary statistics about the real network shown in Table 2. They were asked to choose the diagram which corresponded to the real data, selecting a second choice as well if desired. An example of the task for the *Jazz* network (using 6 diagrams) is shown in Figure 4[1]

Task 2 was designed to test whether GraphPrism could facilitate rapid comparison and classification of multiple network data sets. Analysts viewed 12 GraphPrism diagrams, each generated from synthetic networks constructed using one of the following random graph models: Erdös-Renyí, Barabási-Albert, Growth Model, and Preferential Attachment. For each model, parameters were varied in order to generate three qualitatively different GraphPrism diagrams. Each participant was briefed on how the 12 diagrams were constructed and asked to sort them into four groups of three.

[1]The real *Jazz* network is (3). Notice the larger diameter, higher global transitivity, and absence of local bridges resulting from the many clique structures formed by bands.

Subject	Pairwise Accuracy	Confidence	Ease
1	12	3	3
2	12	3	4
3	12	5	5
4	12	4	4
5	8	4	4
6	8	3.5	4
7	5	4	3
Average	9.9	3.8	3.9

Table 4: Task 2 network classification results.

4.3 Results

Table 3 summarizes the results for Task 1. Overall, classification accuracy was high ($\mu = 0.78$), though it did vary across data sets. Participants generally engaged in one of two distinct strategies. The first was matching summary statistics, such as the global transitivity or number of components, to distinctive patterns in the GraphPrism diagram, as illustrated in the example in the prior section. The second strategy was to extrapolate what the diagrams *should* look like based on inferred structural features of the network. For instance, some participants inferred that the *Jazz* network would include a number of cliques and a low number of structural bridges, allowing them to identify the real GraphPrism. Analysts reported their subjective *confidence* in the selection and the *ease* of the task using a 5-point Likert scale; higher ratings indicated more favorable scores. Overall, the reported confidence was high ($\mu = 4.33$), with higher ratings for tasks performed more accurately. Judgments of ease ($\mu = 4.05$) were relatively uniform across tasks.

Correctness in Task 2 was evaluated using a pairwise accuracy metric: we compared the number of correctly classified pairs to the total number of pairings in the data set. Random guessing yields an expected value of 2.18 correct pairs out of 12, providing a baseline for our results. As shown in Table 4, 4 of the analysts performed the task perfectly, with all 7 performing better than chance. Moreover, all participants completed the task easily in under 5 minutes. All of the analysts made reference to the "overall patterns" evident in each GraphPrism diagram, and three of the four who performed the task perfectly reported starting with a "pattern recognition" approach. Despite the high performance, the average confidence ($\mu = 3.8$) and ease ($\mu = 3.9$) were lower than in Task 1, perhaps due to the novelty of the task.

While this evaluation does not compare GraphPrism to other techniques such as node-link diagrams or adjacency matrices, it provides some evidence that GraphPrism enables effective characterization, comparison, and classification of network data sets. The fact that analysts were able to

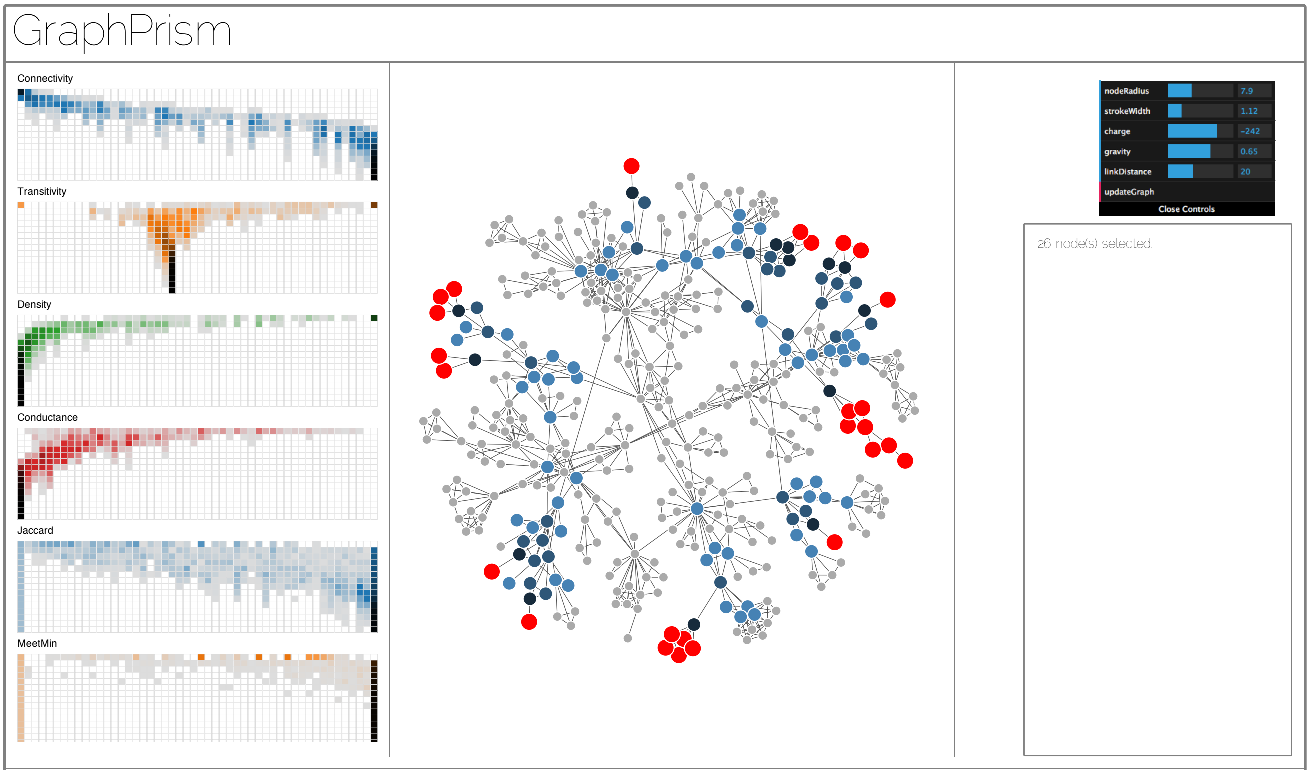

Figure 5: The Network Science collaboration network. Using the *Connectivity* facet, we have selected nodes which can reach the fewest other nodes within 3 hops, representing 'outsiders'.

perform tasks quickly and with relatively high accuracy after minimal training indicates that GraphPrism shows promise in providing an overview of network structure.

4.4 Design Feedback

After completing the evaluation tasks and gaining familiarity with the diagrams, analysts provided feedback on the GraphPrism design. One area in which experts believed GraphPrism excelled was in representing multiple levels of structure. Analysts praised the immediate visibility of the degree and transitivity distributions and the availability of information about the existence of particular types of nodes or edges via focusing attention on a single row or cell of a facet. Analysts appreciated the seamless transition to making inferences about properties of larger structures in the graph, such as the size of the largest component.

Some of the analysts indicated that a node-link diagram would have complemented insights gained from the GraphPrism diagrams. Two of the analysts, for instance, speculated about the existence of core-periphery structures in some networks, something not available in the GraphPrism diagram but potentially visible in a node-link view. Two analysts also pointed out that the GraphPrism diagram obscured the existence of singletons. The primary takeaway regarding the design of the interactive system was that GraphPrism was best coupled with a node-link diagram such that analysts could drill down from the GraphPrism overview to details about node and edges available in a node-link layout.

5. INTERACTIVE SYSTEM

Informed by the results of our paper prototype study, we have developed an interactive system that combines GraphPrism with standard node-link diagrams. Below, we describe the system and demonstrate how it can be used to perform relevant network analysis tasks. We illustrate how GraphPrism and node-link views can complement each other in two ways: (1) GraphPrism provides an overview of important properties not visible in the node-link view and (2) GraphPrism can serve as a selection mechanism for identifying and inspecting nodes and edges of interest.

Our interactive system can be used to accomplish relevant network analysis tasks, ranging from the summarization of important global properties, to the discovery of communities and structures of interest, to the identification of individual nodes and edges which may be important. As an example, we utilize data from the largest connected component of the network science co-authorship graph described earlier; this component contains 379 nodes (authors) and 914 edges (co-authorship ties). The network is large enough to begin to pose a challenge for traditional node-link diagrams and adjacency matrices, but small enough to discuss thoroughly in the remainder of this article.

5.1 Complementing the Node-Link Diagram

The node-link diagram is rendered using a standard force-directed layout. Users can adjust a gravitational parameter which controls how strongly nodes are attracted to the center and a charge parameter which controls how strongly nodes repel one another. Attributes of individual nodes or edges are available via mouseover or selection.

As shown in Figure 5, the GraphPrism for this data set is rendered with 6 facets: *Connectivity*, *Transitivity*, *Density*, *Conductance*, *Jaccard*, and *MeetMin*. Swapping in additional metrics is simple; for instance, one can add *In-*

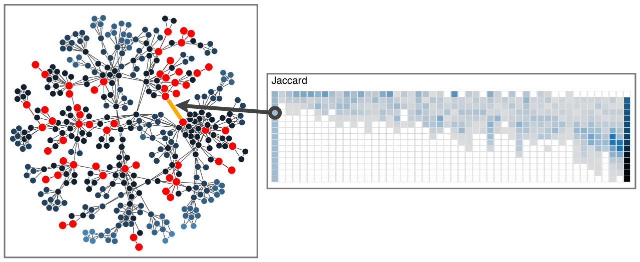

Figure 6: Using GraphPrism and node-link views in tandem, we identify an interesting bridge which spans a distance of 8 to 10 hops.

Connectivity and *Out-Connectivity* for directed graphs. Both the node-link diagram and GraphPrism were implemented using the Data-Driven Documents (D3) framework [10].

Within each facet, values are encoded by mapping color intensity approximately to a logarithmic scale. Each facet is assigned a unique color hue. We ensure that cells containing non-zero values are visually distinguishable from those with zero values. The color then ramps from grey, through darker hues, to black. Black indicates that (nearly) all the entities in the graph are represented in that cell. As in the Karate example in Section §3.4, the static image alone allows us to quickly infer many properties about the network.

The first row of the *Connectivity* facet shows that node degrees appear to follow roughly a power-law distribution: there are many nodes with low degree and few with large degree. At $\ell = 10$, we see that most nodes can reach one another, giving us a sense of the effective diameter, even though the true diameter is much greater. In the *Transitivity* facet, we see that global transitivity is relatively high (roughly 0.45), and from the first row we note a high number of nodes with clustering near 1 (to be expected given that papers with multiple authors form cliques).

5.2 Interaction via Linked Selection

Our system supports linked selection (brushing & linking) between GraphPrism facets and the node-link diagram. Selecting the fifth cell in the first row of *Connectivity* highlights the node with the highest number of collaborators. We can quickly identify this node as Albert-László Barabási — a prominent author in the community. While high-degree nodes are visible in the node-link view, it is difficult to identify the most-connected node as quickly without an aid.

In Figure 5, the leftmost colored cell in the third row of *Connectivity* has been selected, showing nodes which can reach the fewest nodes ($< 5\%$) within 3 hops. Selected elements (nodes or edges) are highlighted in red in the node-link view, allowing us to quickly identify these peripheral nodes. The cell selected is in row $\ell = 3$; accordingly the 3-neighborhoods of the selected nodes are highlighted as well, with color ranging from black to blue to represent the shortest distance to any selected nodes. If an edge is selected, the neighborhoods of incident nodes are highlighted. All passive elements are colored gray to reduce saliency.

Often the GraphPrism and node-link views work in tandem to surface interesting elements. For example, when looking for bridges, we might start with the *Jaccard* facet, shown in more detail in Figure 6. Selecting the circled cell (leftmost column, 4th row) highlights 31 bridges in the graph. The node-link view then reveals one edge which appears to span quite a distance. When the cell below it is se-

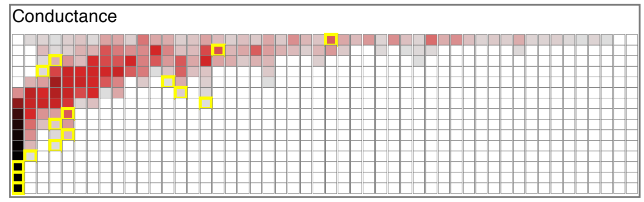

Figure 7: Viewing the conductance facet for a selected node, we see a possible natural community boundary at $\ell = 3$ or 4.

lected the edge disappears, indicating that the edge bridges a path of length 8 to 10. This signifies a potentially important collaboration linking otherwise distant elements.[2]

Users can iterate through a set of selected elements (nodes or edges) using the left and right arrow keys. As each element is selected, the corresponding neighborhood is highlighted as above. Within each row of each facet, the cell corresponding to the selected element is highlighted using a yellow outline stroke, providing a quick visual summary of how metric values for that element vary with ℓ. Pictured in Figure 7 is the *Conductance* facet with cells highlighted for a single node. Here we see that conductance values dip appreciably at $\ell = 3$ and 4, signaling the possible presence of a meaningful community boundary around the 3- and/or 4-neighborhood of the selected node.[3] Cells corresponding to the selected node are highlighted in other facets, as well, enabling inspection of other structural features.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have described a novel visualization technique called GraphPrism for the analysis of large, complex networks. By calculating distributions of metrics over neighborhoods of increasing size, we create a set of multi-scale histograms that we call 'facets'. The composition of these facets yields a compact, scalable visualization which surfaces a number of interesting structural features of a graph. Through paper prototype studies with network analysis experts, we found that GraphPrism provides a useful visual overview of graph datasets. We also introduced an interactive network analysis system that links GraphPrism facets with a node-link diagram. Our case study illustrates how this combination of features can aid analysts in exploring graph data.

One avenue for future work is exploring algorithms for efficient calculation of graph metrics across multiple network neighborhoods. Some parallelization can be trivially achieved by separately starting calculations at each individual node or edge. More sophisticated approaches might eliminate unnecessary computation via dynamic programming or other methods. Also, while the individual metrics used in this paper are relatively cheap to compute, one might wish to apply the GraphPrism approach to more computationally complex metrics such as betweenness centrality. In such cases, approximation methods may be needed.

[2] Additional investigation reveals that this edge represents a paper by G. Bianconi and A. Capocci entitled "Number of Loops of Size h in Growing Scale-Free Networks", which interestingly creates its own sizable loop in this network.

[3] This node is connected peripherally to the network. After 3 hops, a paper (clique) with 5 authors is added, and after 4, another 3-author collaboration, explaining the decrease in conductance. After this, the network branches out again.

Another future step will be to improve our interactive visualization through continued evaluation. We intend to evaluate our interactive prototype with network analysis experts in order to solicit feedback on our design. One specific goal is identifying useful data to surface in a detail panel, thereby facilitating analysis of both aggregate and individual selections. In addition, through a comparison with other visualization methods for large networks, we hope to gain more insight into the strengths and weaknesses of our approach. Future research might also identify alternative interaction techniques well-suited to GraphPrism views. For instance, when dealing with large graphs where complete node-link or matrix views are ill-advised, how might GraphPrism be used to select subgraphs for closer inspection?

Finally, while our current metrics provide insight into simple directed and undirected graphs, an important motivation underlying the modular approach to constructing GraphPrism is the ability to adapt this technique to multiple graph types, including hypergraphs and bi-partite networks. A corresponding area for future work is to identify metrics which may be more suitable for analyzing these types of graphs which can be added as new GraphPrism facets.

7. REFERENCES

[1] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman. Search in power-law networks. *Physical Review E*, 64(4), Sep 2001.

[2] D. Auber, Y. Chiricota, F. Jourdan, and G. Melancon. Multiscale visualization of small world networks. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 75–81. IEEE Computer Society, 2003.

[3] J. Bagrow, E. Bollt, and J. Skufca. Portraits of complex networks. *EPL (Europhysics Letters)*, 81:68004, 2008.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[5] V. Batagelj and A. Mrvar. Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/, 2006.

[6] R. Becker, S. Eick, and A. Wilks. Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):16–28, 2002.

[7] J. Bertin. *Graphics and Graphic Information Processing*. Walter de Gruyter, Berlin, 1981.

[8] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, Madison, WI, USA, 1983.

[9] J. Blythe, C. McGrath, and D. Krackhardt. The effect of graph layout on inference from social network data. In *Proceedings of the Symposium on Graph Drawing*, 1995.

[10] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, dec 2011.

[11] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9), 2003.

[12] R. Burt. Structural holes versus network closure as social capital. *Social capital: Theory and research*, pages 31–56, 2001.

[13] R. S. Burt. *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press, New York, NY, USA, 2005.

[14] R. S. Burt. *Neighbor Networks: Competitive Advantage Local and Personal*. Oxford University Press, New York, NY, USA, 2010.

[15] A. Clauset, C. R. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2007.

[16] E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172. ACM, 1969.

[17] M. Freire, C. Plaisant, B. Shneiderman, and J. Golbeck. ManyNets: an interface for multiple network analysis and visualization. In *Proceedings of ACM Human Factors in Computing Systems (CHI)*, pages 213–222. ACM, 2010.

[18] M. Girvan and M. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, June 2002.

[19] F. Ham, H.-J. Schulz, and J. M. Dimicco. Honeycomb: Visual analysis of large scale social networks. In *Proceedings of the 12th IFIP TC 13 International Conference on Human Computer Interaction: Part II*, INTERACT '09, pages 429–442, Berlin, Heidelberg, 2009. Springer-Verlag.

[20] N. Henry and J. Fekete. Matrixexplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics*, pages 677–684, 2006.

[21] N. Henry and J. Fekete. Matlink: Enhanced matrix visualization for analyzing social networks. *Human-Computer Interaction–INTERACT 2007*, pages 288–302, 2010.

[22] N. Henry, J. Fekete, and M. McGuffin. NodeTrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007.

[23] J. Kleinberg. Complex networks and decentralized search algorithms. In *In Proceedings of the International Congress of Mathematicians (ICM)*, 2006.

[24] J. Leskovec. *Dynamics of Large Networks*. PhD thesis, Carnegie Mellon University, Sep 2008.

[25] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM.

[26] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5:110–141, 1986.

[27] G. Melancon and A. Sallaberry. Edge metrics for visual graph analytics: A comparative study. In *Proceedings of IEEE Information Visualization (InfoVis)*, pages 610–615, 2008.

[28] S. Milgram. The small-world problem. *Psychology Today*, May 1967.

[29] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[30] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(036104), 2006.

[31] M. Wattenberg. Visual exploration of multivariate graphs. In *Proceedings of ACM Human Factors in Computing Systems (CHI)*, CHI '06, pages 811–819, New York, NY, USA, 2006. ACM.

[32] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[33] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010.

[34] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.