Uncertainty Visualization for Graph Coarsening

Fangfei Lan
University of Utah
Salt Lake City, USA
fangfei.lan@sci.utah.edu

Sourabh Palande
Michigan State University
East Lansing, USA
palandes@msu.edu

Michael Young
University of Utah
Salt Lake City, USA
m.c.young0@gmail.com

Bei Wang
University of Utah
Salt Lake City, USA
beiwang@sci.utah.edu

***Abstract*—The complexity of large real-world graphs makes their analyses prohibitively costly and their visualizations uninformative. The idea behind graph reduction is to reduce the size of a graph while preserving its properties of interest. To improve computational efficiency and to provide provable guarantees, many graph reduction techniques employ randomization. However, the uncertainty associated with randomized graph reduction and its subsequent interpretation has remained largely unexplored. In this paper, we present a framework to quantify and visualize the uncertainty associated with randomized graph reduction techniques. We focus on spectral clustering introduced by Ng, Jordan, and Weiss, a popular graph reduction technique that reduces the number of nodes by clustering the nodes of a graph into super-nodes. We introduce two uncertainty measures – local adjusted Rand indices and co-occurrences – to quantify and visualize uncertainty associated with an ensemble of reduced graphs. We demonstrate via experiments, that these measures complement each other in visualizing uncertainty and guiding the selection of optimal numbers of clusters.**

***Index Terms*—Summarization and visualization of large networks, uncertainty visualization, graph clustering and coarsening**

I. Introduction

Graphs are ubiquitous in modeling large and complex data in science and engineering. They are also increasingly relevant in deep learning [52]. In the age of big data, a real-world graph can become prohibitively large, thereby hampering the efficiency of its analysis and the interpretability of its visualization. These problems can be addressed by graph reduction, where the idea is to reduce the size of the graph, while preserving its properties of interest.

Graph sparsification and graph coarsening are the two most commonly used graph reduction techniques. Graph *sparsification* reduces the number of edges in a graph while maintaining the number of nodes. Graph *coarsening*, on the other hand, reduces the number of nodes, which implicitly reduces the number of edges. Graph coarsening appears in many applications, such as visualization [18], graph partitioning [42], dimensionality reduction [6], and convolutional neural networks [4]. In this paper, we focus on graph coarsening, which is typically achieved via node clustering, where nodes from the original graph are clustered together to form the *super-nodes* of the reduced graph. Specifically, we use a representative spectral clustering algorithm introduced by Ng, Jordan and Weiss [36] (referred to as the NJW algorithm); although our technique is applicable to any other graph coarsening algorithm.

Although graph reduction techniques have been used in analysis and visualization, the uncertainty associated with their outputs and the subsequent visual interpretation has remained largely unexplored. Many graph reduction techniques, including the NJW algorithm, employ randomization to provide theoretical guarantees and to improve computational efficiency. As a result, the same algorithm applied to the same input may produce different outputs across different runs. In this paper, we are interested in visualizing the uncertainty of an *ensemble* of reduced graphs as a result of such a randomized process. Even if different instances of the reduced graphs agree on the global level, variations may occur in the size and connectivity of individual communities. Understanding such variability will help us obtain deeper insights into the reduced graphs and gain more confidence in the analytical results.

Contribution. Randomization from the NJW algorithm introduces uncertainty among the reduced graphs and the induced insights from such graphs. We present a general and a flexible framework to quantify and visualize this uncertainty. Our contributions are as follows:

- We introduce two uncertainty measures - local adjusted Rand indices and co-occurrences – that not only provide an overall uncertainty score for the entire clustering, but also capture the uncertainty associated with each super-node of the reduced graph in an ensemble;
- We demonstrate via experiments, that these measures complement each other in visualizing uncertainty in the communities and guiding the selection of optimal coarsening parameters;
- Furthermore, we provide an open source demo of our framework [1] that allows the users to explore the structures of reduced graphs across multiple runs of the NJW spectral clustering algorithm.

It is important to note that although we use the NJW algorithm to illustrate our framework, our approach is applicable to other randomized graph reduction algorithms.

II. Related Work

Uncertainty visualization and graph visualization. Uncertainty visualization conveys uncertainty information through visualization; see [3], [38] for recent surveys on information and scientific visualization, [9] for uncertainty-aware visual

[1] https://github.com/tdavislab/uncertainty-graph-vis/

978-1-6654-8045-1/22/$31.00 2022 IEEE

analytics, and [1] for uncertain data mining. Uncertainty information is typically described by statistical quantities such as mean, median, and standard deviation, and visualized using color, opacity, texture, glyphs, and animation [38].

For graph visualization, there are excellent surveys for information visualization [20], scientific visualization [51], visual analysis of large graphs [47], dynamic graphs [2] and graph drawing [11]. In this paper, we employ a classic graph layout technique, the Fruchterman-Reingolds (FR) force-directed layout algorithm [13].

Previous works have studied uncertainty for special types of graphs such as trees [27], [53] and lattice graphs [8]. Innovative visual encodings have been proposed to visualize observational uncertainty, where the uncertainty of a community is given *a priori*. Vehlow et al. [46] visualized fuzzy overlapping network communities using a combination of color, geometry, brightness, position, edge color, edge thickness, and pie chart to encode uncertainty information. In contrast, the work of Schulz et al. [43] distinguished between uncertainty inherent to data (probabilities for observation uncertainty) and uncertainty introduced by their visualization technique (stress and distortion). In our formulation, we are interested in capturing the uncertainty that arises during a randomized graph reduction process.

Graph reduction. Graph reduction has been explored in graph visualization previously [19], [22], [50]. To the best of our knowledge, our paper is the first to explore and visualize uncertainty associated with randomized graph reduction techniques. Graph coarsening is the process of reducing the number of nodes by partitioning the graph and selecting a representative node for each partition, or merging nodes in a partition to form a *super-node*. Heuristics such as heavy edge matching [25], [26], [49] and node similarities or distances [7], [39], [42] are commonly used in coarsening algorithms. Coarsening is by far the most popular graph reduction technique in graph drawing [17], [18], [49]. However, apart from a few exceptions [5], [10], [29], the algorithms used in practice lack strong theoretical support.

Spectral clustering algorithms are a type of graph coarsening techniques that rely on the notion of spectral similarity between nodes. Spectral clustering was made popular by Shi and Malik [44], and Ng, Jordan and Weiss [36] (referred to as the NJW algorithm). The NJW algorithm, which we focus on, applies the standard k-means clustering to the first k nontrivial eigenvectors of the normalized graph Laplacian. While classic spectral clustering algorithms may not scale well with the size and density of the input graphs, they typically enjoy strong theoretical support (e.g., [28], [29], [36]). Loukas [28] recently proposed a general multi-level coarsening scheme such that the reduced graph preserves the eigenvalues and the eigenvectors of the original graph in a restricted setting.

When coarsening a graph, the correct choice for the number of clusters, k, is often not apparent from prior knowledge of the data. Several methods have been proposed in the literature to help determine the appropriate k [16], [37], [40], [45]. Our work aims to guide the choice of optimal k via visualization.

Consensus clustering. Finally, consensus clustering is used to represent the consensus across multiple clusterings from the same input data, across either multiple runs of the same algorithm [31] or different clustering algorithms. We use the notion of clustering-induced graphs to derive co-occurrences for uncertainty visualization. These ideas have previously been explored in the context of consensus clustering [12], [14].

III. Background

Let $G = (V, E, w)$ be a simple, weighted, undirected graph with n nodes, m edges, and positive edge weights $w : E \to \mathbb{R}^+$. Its $n \times n$ *weighted adjacency matrix* A is defined as $A_{ij} = A_{ji} = w_e := w(e)$ for all $i, j \in V$ forming an edge $e \in E$; otherwise, $A_{ij} = A_{ji} = 0$. Since G is a simple graph, $A_{ii} = 0$. Let D be an $n \times n$ diagonal matrix such that $D_{ii} = \sum_i A_{ij}$. The *graph Laplacian* of G is the matrix $L = D - A$. The *normalized adjacency matrix* and the *normalized graph Laplacian* are defined as $\tilde{A} = D^{-1/2}AD^{-1/2}$, and $\tilde{L} = D^{-1/2}LD^{-1/2}$.

Coarsening matrix. In graph coarsening, given a graph $G = (V, E, w)$, a reduced graph $H = (V', E', w')$ is constructed from G w.r.t. a set of k partitions $\mathbf{S} = \{s_1, \cdots, s_k\}$ of V. Each *super-node* of H is a *cluster*, denoted by c_i, and corresponds to a partition $s_i \subseteq V$ (for $1 \leq i \leq k$). The super-nodes c_i and c_j are connected via a *super-edge*. The weight of a super-edge is given by the sum of weights of edges connecting the nodes across the clusters. That is, $A'_{i,j} := \sum_{v \in s_i, u \in s_j} A(v, u)$, where A' is a $k \times k$ adjacency matrix for H.

We identify a *clustering* with a *coarsening matrix* $M \in \mathbb{R}^{k \times n}$, which is defined as [24, Section 3.1]: $M_{ij} = 1$ if $v_j \in s_i$; $M_{ij} = 0$, otherwise. We have $A' = MAM^T$.

Spectral clustering. Spectral clustering can be used in graph coarsening to group clusters of nodes into *super-nodes*, hence reducing the size of a graph. Many variants of spectral clustering are described in the literature, see [32] for a survey. In this paper, we employ the NJW algorithm; although our framework is easily generalizable to other graph coarsening algorithms.

Let k be the number of clusters. The NJW algorithm utilizes two key ingredients: the k largest eigenvalues of the normalized adjacency matrix and k-means clustering. Given a graph G with a weighted adjacency matrix A, we compute the eigenvectors $u_1, \ldots, u_k$ corresponding to the k largest eigenvalues of its normalized adjacency matrix $\tilde{A}$. The spectral embedding of G is the $n \times k$ matrix $U = [u_1, \ldots, u_k]$. The matrix U is then row-normalized and used as an input to the k-means clustering algorithm. The cluster assignment returned by k-means clustering is used to cluster the nodes of the graph, i.e., if i^{th} row of U is assigned to cluster j, then node i of G is assigned to cluster j. Then, all nodes assigned to cluster j are grouped into the super-node j to construct the reduced graph.

IV. Quantifying Node Uncertainty

In graph coarsening, node uncertainty may arise due to the initialization or the randomization inherent to the underlying algorithm. For instance, when coarsening a graph using the NJW algorithm, node uncertainty arises due to the initialization of k-means clustering internal to the algorithm. We introduce two complementary methods to quantify node uncertainty for visualization, namely, locally adjusted Rand indices and co-occurrences. The former captures the amount of contribution from one cluster to the global clustering; whereas the latter represents the co-occurrence probability of node pairs among all clusters.

A. Node Uncertainty via Local Adjusted Rand Indices

Given a graph $G = (V, E, w)$, let $\mathbf{S} = \{s_1, s_2, \cdots, s_k\}$ denote a *clustering* (coarsening) of the node set V into k clusters (super-nodes). Let s_i denote a cluster in $\mathbf{S}$, with size $|s_i|$; c_i is its corresponding super-node in the reduced graph $H = (V', E', w')$.

The Rand index is a commonly used measure of similarity between two clusterings. Consider a set V of n nodes in G and a pair of clusterings $\mathbf{S}^1$ and $\mathbf{S}^2$, the *Rand index* calculates the fraction of correctly classified pairs of nodes w.r.t. all pairs. It is defined as

$$\mathcal{R}(\mathbf{S}^1, \mathbf{S}^2) = \frac{n_{11} + n_{00}}{\binom{n}{2}}, \tag{1}$$

where n_{11} is the number of pairs in the same cluster under $\mathbf{S}^1$ and $\mathbf{S}^2$, and n_{00} the number of pairs in different clusters under $\mathbf{S}^1$ and $\mathbf{S}^2$ [48].

Given two clusterings $\mathbf{S}^1$ and $\mathbf{S}^2$, let M^1 and M^2 denote the corresponding coarsening matrices. Let $F = M^1(M^2)^T$ be the $k \times k$ *confusion matrix* for $\mathbf{S}^1$ and $\mathbf{S}^2$, where $F_{ij} = |s_i^1 \cap s_j^2|$ captures the size of overlap between clusters s_i^1 in $\mathbf{S}^1$ and s_j^2 in $\mathbf{S}^2$. The *adjusted Rand index* is defined as the normalized difference between the Rand index and its expected value [30], [48].

$$\begin{aligned} \mathcal{AR}(\mathbf{S}^1, \mathbf{S}^2) &= \frac{\mathcal{R}(\mathbf{S}^1, \mathbf{S}^2) - \mathbb{E}[\mathcal{R}(\mathbf{S}^1, \mathbf{S}^2)]}{1 - \mathbb{E}[\mathcal{R}(\mathbf{S}^1, \mathbf{S}^2)]} & (2) \\ &= \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} \binom{F_{ij}}{2} - r_3}{\frac{1}{2}(r_1 + r_2) - r_3}, & (3) \end{aligned}$$

where $r_1 = \sum_{i=1}^{k} \binom{|s_i^1|}{2}$, $r_2 = \sum_{j=1}^{k} \binom{|s_j^2|}{2}$, and $r_3 = r_1 r_2 / \binom{n}{2}$.

Given the adjusted Rand index that measures the *global* similarity between clusterings $\mathbf{S}^1$ and $\mathbf{S}^2$, we introduce a *local adjusted Rand index* ($\mathcal{LAR}$) that captures the amount of contribution from a cluster in $\mathbf{S}^1$ to the global measure with $\mathbf{S}^2$:

$$\mathcal{LAR}(s_i^1, \mathbf{S}^2) = \frac{\sum_{j=1}^{k} \binom{F_{ij}}{2} - r_3'}{\frac{1}{2}(r_1 + r_2) - r_3}, \tag{4}$$

where r_1, r_2, r_3 are the same as in (3), and $r_3' = \binom{|s_i^1|}{2} r_2 / \binom{n}{2}$. By definition, we have $\sum_{i=1}^{k} \mathcal{LAR}(s_i^1, \mathbf{S}^2) = \mathcal{AR}(\mathbf{S}^1, \mathbf{S}^2)$.

Suppose we are given an ensemble $\mathcal{S} = \{\mathbf{S}^0, \ldots, \mathbf{S}^l\}$ of $l + 1$ clusterings of nodes in G. $\mathcal{S}$ may be obtained by running a single randomized graph coarsening algorithm multiple times, or multiple graph coarsening algorithms. Each ensemble member gives rise to a reduced graph. We encode node uncertainty based on a *representative graph*, which arises from a representative clustering from the ensemble.

To find such a representative, we first define a distance between two clusterings in the ensemble $d(\mathbf{S}^1, \mathbf{S}^2)$, based on the adjusted Rand index:

$$d_{\mathcal{R}}(\mathbf{S}^1, \mathbf{S}^2) = 1 - \mathcal{AR}(\mathbf{S}^1, \mathbf{S}^2). \tag{5}$$

A representative clustering is the one in the ensemble that minimizes the sum of distances to other ensemble members, i.e., $\arg\min_{\mathbf{S} \in \mathcal{S}} \sum_{i=0}^{l} d(\mathbf{S}, \mathbf{S}^i)$; *w.l.o.g.*, let $\mathbf{S} := \mathbf{S}^0$ denote the representative clustering and $\mathbf{S}^1, \cdots, \mathbf{S}^l$ the remaining clusterings in the ensemble (we relabel the clusterings if necessary). Each $\mathbf{S}^t$ $(1 \leq t \leq l)$ gives rise to a coarsening matrix M^t and a reduced graph H^t.

For each cluster s_i in $\mathbf{S}$, we compute its local contribution to the global similarity measure between $\mathbf{S}$ and each of the ensemble members $\{\mathbf{S}^1, \cdots, \mathbf{S}^l\}$ and obtain a distribution of local measurements. Formally, let α_i^t denote the $\mathcal{LAR}$ between s_i in $\mathbf{S}$ and $\mathbf{S}^t$ $(1 \leq t \leq l)$; then $\alpha_i^t = \mathcal{LAR}(s_i, \mathbf{S}^t)$. The uncertainty associated with a super-node c_i in the representative reduced graph H is therefore described by the mean and standard deviation of the distribution $\{\alpha_i^1, \cdots, \alpha_i^l\}$.

To establish the relationship between the number of clusters (super-nodes) and the uncertainty of the clusterings (ensembles of reduced graphs), we compute a global uncertainty measure for each ensemble based on $\mathcal{LAR}$ of its representative graph, referred to as the *aggregated Rand index* ($\mathcal{ARI}$). For each ensemble, its $\mathcal{ARI}$ is defined to be the sum of the standard deviations of $\mathcal{LAR}$ of each super-node in the representative graph. In other words, $\mathcal{LAR}$ captures the local uncertainty for a single super-node in a reduced graph while $\mathcal{ARI}$ captures the global uncertainty for an entire reduced graph.

Given a pair of clusterings $\mathbf{S}^1$ and $\mathbf{S}^2$, computing $\mathcal{LAR}(s_i^1, \mathbf{S}^2)$ for all i reduces to constructing F (by checking set memberships of n nodes) and computing $\sum_{j=1}^{k} \binom{F_{ij}}{2}$, which takes $O(n)$ and $O(k^2)$ time, respectively. For $l+1$ clusterings in the ensemble, computing $\mathcal{LAR}$ among all pairs in the ensemble takes $O(l^2(n + k^2))$ time.

B. Node Uncertainty via Co-occurrences

Consider a graph $G(V, E, w)$, with a set V of n nodes. Let $\mathcal{S} = \{\mathbf{S}^0, \ldots, \mathbf{S}^l\}$ be an ensemble of clusterings, where each $\mathbf{S}^t$ $(0 \leq t \leq l)$ is a clustering of nodes of G into k clusters. For each clustering $\mathbf{S}^t \in \mathcal{S}$, we define an $n \times n$ cluster-induced adjacency matrix A^t such that

$$A_{jk}^t = \begin{cases} 1 & \text{if } v_j, v_k \in s_i^t \text{ for some } s_i^t \in \mathbf{S}^t, \\ 0 & \text{otherwise.} \end{cases}$$

A^t is the same as the *cluster-induced element graph* proposed by Gates et al. [14], and it captures the co-occurrence of node pairs among the clusters of $\mathbf{S}^t$. By definition, $A^t = (M^t)^T M^t$.

Let A^* denote the element-wise average of matrices A^t, i.e., $A^* = \frac{1}{l+1} \sum_{t=0}^{l} A^t$. A_{jk}^* is the empirical estimate of the co-occurrence probability of nodes v_j and v_k across all clusterings in the ensemble. We have $0 \leq A_{jk}^* \leq 1$, where a value close to 1 indicates that its end points have a high probability of belonging to the same cluster.

Treating the binarized A^* as an adjacency matrix, we construct a graph $G^* = (V^*, E^*, w^*)$, referred to as the A^* *graph*, with edge weights $w^*_e = A^*_{jk}(1 - A^*_{jk})$. Since probabilities A^*_{jk} close to 0 or 1 are both considered stable, w^*_e captures the uncertainty of an edge; its largest value (i.e., 0.25) corresponds to the highest uncertainty w.r.t. co-occurrences of its end points. Finally, for any given reduced graph H^t with k clusters (super-nodes), with the corresponding clustering $\mathbf{S}^t = \{s^t_1, \ldots, s^t_k\}$, and the coarsening matrix M^t, the $k \times k$ matrix $Q^t = M^t A^* (M^t)^T$ captures the co-occurrence probabilities among pairs of clusters in H^t.

Computing A^t for a fixed t takes matrix multiplication time of $O(n^2 k)$; computing A^t for all t therefore takes $O(ln^2 k)$ time. Given all A^t, computing A^* takes $O(ln^2)$. Computing Q^t for a fixed t takes matrix multiplication time of $O(n^2 k)$; thus Computing Q^t for all t takes $O(ln^2 k)$ time.

V. Results

We now apply the NJW algorithm to four datasets and obtain ensembles of reduced graphs. We visualize node uncertainty of these ensembles via local adjusted Rand indices ($\mathcal{LAR}$) and co-occurrences. A key takeaway is that these two uncertainty measures complement each other to visualize uncertainty in randomized spectral clustering and empirically guide the choices of the optimal numbers of clusters (super-nodes). This is particularly useful for understanding uncertainty associated with community detection based on graph coarsening.

Visual encoding of uncertainty measures. We first introduce our visual encoding. For each dataset, we visualize the reduced graphs with varying number of clusters. In terms of $\mathcal{LAR}$ uncertainty measure in each reduced graph, the size and color of a super-node double encode the mean $\mathcal{LAR}$ measure, which reflects the contribution of the super-node to the global uncertainty. The orange ring around each super-node encodes the 1st standard derivation of $\mathcal{LAR}$, which reflects node uncertainty. Communities with large mean $\mathcal{LAR}$ measures are typically considered to be more important in our uncertainty framework. In other words, a large super-node with a thick orange ring is more important than a small super-node with an orange ring of the same width. The thickness and color of an edge in a reduced graph reflect the mean edge weight.

For co-occurrence, we visualize the A^* graphs, where edge uncertainty is encoded by a color map. A stable clustering structure is indicated by well-separated clusters with high edge weights among nodes within a cluster and low edge weights between clusters.

A. *Les Miserables Dataset*

We begin with the well-known *Les Miserables* dataset. Each node represents a character in Victor Hugo's novel "Les Miserables", and an edge connects two nodes if the corresponding characters co-appear in a scene within the novel. This leads to a graph of 77 nodes and 254 edges. This dataset is commonly used to benchmark the performance of community detection algorithms, where the goal is to distinguish groups of characters based on their social interactions and to detect key players in a storyline.

Recently, Hu et al. [21] presented a modified spectral clustering algorithm that incorporated a Bayesian analysis model to improve the selection of the number of clusters. For the *Les Miserables* test case, they concluded that the best clustering should consist of 11 communities. We apply our uncertainty visualization framework to demonstrate that we obtain interpretable community structures when the number of clusters k is set to 8 or 11.

Global uncertainty with $\mathcal{ARI}$. To study the global uncertainty trend, we apply the NJW algorithm to the dataset for $2 \leq k \leq 76$ (since there are 77 nodes in the graph). We plot the global uncertainty measure – aggregated Rand index ($\mathcal{ARI}$) – w.r.t. the number of clusters, as shown in Fig. 1a. We observe two local minima at 8 and 11 clusters, respectively.

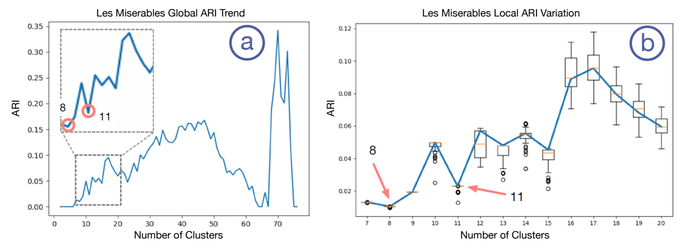

Fig. 1. *Les Miserables* dataset. Left: $\mathcal{ARI}$ with an increasing number of clusters. Right: the variability of $\mathcal{ARI}$ with 7 to 20 clusters.

For a fixed k, we apply the NJW algorithm to the dataset 100 times and generate an ensemble of 100 members. To study the variability of such an ensemble, we treat each ensemble member (a reduced graph) as the representative graph and compute its $\mathcal{ARI}$. We then study the distribution of this global uncertainty measure by plotting the set of 100 $\mathcal{ARI}$ as a box plot (see Fig. 1b). As we vary the cluster sizes from 7 to 20, we observe low ensemble variability surrounding the two local minima (with 8 and 11 clusters). This means that the randomized process during graph coarsening produced similar reduced graphs across the 100 runs.

Local uncertainty with $\mathcal{LAR}$. In Fig. 2, using $\mathcal{LAR}$, we highlight the node uncertainties for representative graphs with 8, 9, 10, 11, and 12 clusters. Recall a representative graph is the one in the ensemble that minimizes the sum of distances to the others in the ensemble. For each representative graph, the size of the super-node reflects the average $\mathcal{LAR}$ score of the corresponding cluster. The standard deviation of the $\mathcal{LAR}$ scores is represented by the width of the orange ring around each super-node, indicating the associated uncertainty. In Fig. 2, the representative graphs with 8 and 11 clusters are shown to contain one and four clusters with thin, but visible, orange rings, respectively, indicating a low level of uncertainty among these super-nodes. The number of clusters with visible orange rings and the widths of the orange rings increase when we deviate from 8 and 11 clusters. This observation confirms an increase in overall uncertainty as we move away from the two local minima. In addition, the representative graph at 11

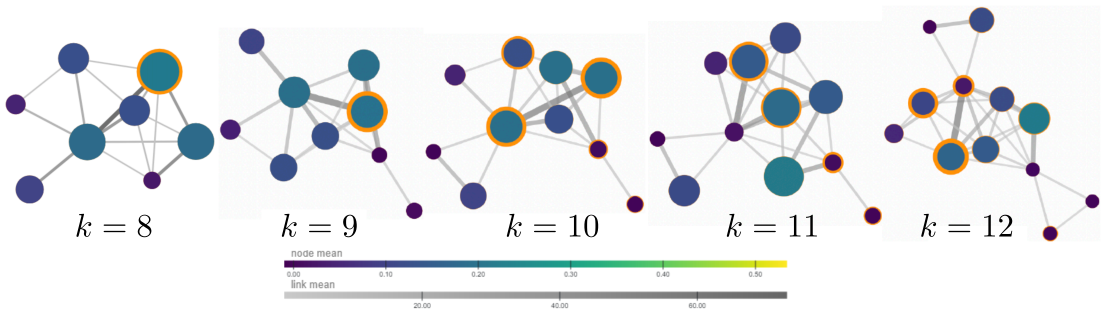

Fig. 2. *Les Miserables* dataset: visualizing node uncertainty of representative graphs with 8, 9, 10, 11, and 12 clusters, respectively.

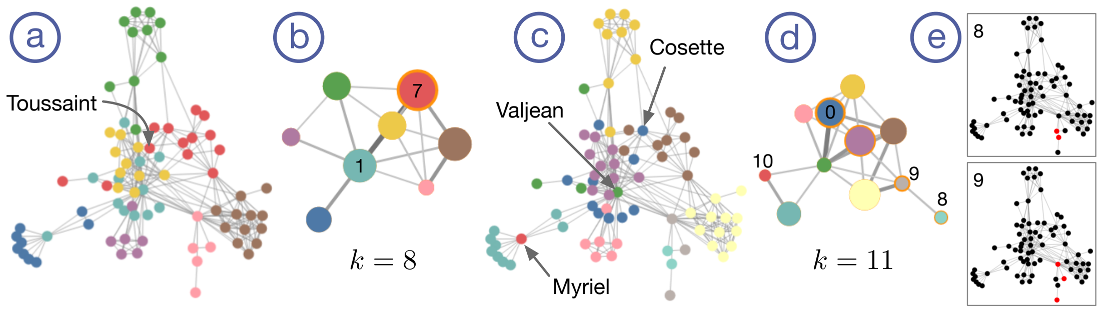

Fig. 3. *Les Miserables* dataset: the original and the reduced representative graphs with 8 (a-b) and 11 (c-d) clusters. Nodes are colored by cluster memberships.

clusters is well aligned with the results of Hu et al., where the cluster memberships are almost identical to the ones in [21].

Comparing cluster memberships. We compare cluster memberships at the two local minima ($k = 8$ and 11) by computing a matching of clusters based on the largest pairwise intersections. We find that the representative graphs for $k = 8$ vs. $k = 11$ have close to identical cluster memberships.

A detailed analysis reveals that clusters in Fig. 3d typically merge to form clusters in Fig. 3b. Such merging behaviors are explainable based on the storyline. For example, clusters 0 and 10 in Fig. 3d merge to form cluster 1 in Fig. 3b. Cluster 0 in Fig. 3d reflects the close social circle around the main character Valjean, and cluster 10 contains a single character Myriel. Myriel shared a scene with Valjean early on in the story, years before the main storyline for Valjean began. However, Myriel did not appear in Valjean's main storyline. Therefore, Myriel is in a cluster by himself in Fig. 3d, but he can also be considered part of Valjean's social circle in Fig. 3b. This example further illustrates that the selection of the optimal number of clusters guided by our visualization framework is reasonable and interpretable.

We further investigate the possible cause of uncertainty in the super-nodes (clusters) of Fig. 3b and Fig. 3d. Fig. 3b shows that super-node 7 has the largest uncertainty (indicated by its thick orange ring). A close examination of the cluster reveals that this uncertainty is largely due to the node Toussaint in Fig. 3a. Toussaint is a servant for both Cosette and Valjean, two of the main characters in the novel and each with their own social circles. Across multiple runs, Toussaint is frequently clustered into Cosette's social circle, which corresponds to super-node 7; it is also occasionally considered to be part of Valjean's circle, which corresponds to super-node 1.

In Fig. 3d, super-node 0 appears to have the highest uncertainty (with the thickest orange ring), due to the character Cosette (see Fig. 3c). Cosette has complicated relationships with characters from different social groups in the novel, making her a difficult character to cluster. The uncertainties associated with super-nodes 8 and 9 in Fig. 3d can also be explained using a frequency analysis. Specifically, the five red nodes in Fig. 3e correspond to one parent, three children, and a housekeeper of the Gorbeau household. Graph coarsening frequently splits up the five characters from the same household, resulting in the high uncertainties for super-nodes 8 and 9 in Fig. 3d.

Co-occurrences with A^* graphs. As a secondary guidance, we visualize the A^* graphs computed based on co-occurrences for $k = 8, 9, 10, 11$, and 12, see Fig. 4 top. Fig. 4 bottom shows the histograms of non-zero edge weights of these A^* graphs. The histograms for $k = 8$ and $k = 11$ are shown to have fewer and shorter bars in the highly uncertain region, i.e., with values close to 0.25, compared to their neighbors. Meanwhile, the A^* graphs at $k = 8$ and $k = 11$ display well-separated clusters; each of such clusters corresponds to a low uncertainty super-node in the representative graph of Fig. 3b and Fig. 3d, respectively. Such observations based on the A^* graphs further guide our selection of optimal number(s) of clusters.

B. *College Football Dataset*

The *College Football* network collected by Girvan and Newman [15] represents the schedule of Division I games for the 2000 season. The nodes represent the 115 teams. There is an edge between two nodes if a game was played between the corresponding teams. The teams were divided into 11 conferences, plus a group of independent teams. Recently, Newman and Reinert [35] used a maximum-likelihood method to estimate that the correct number of communities for the

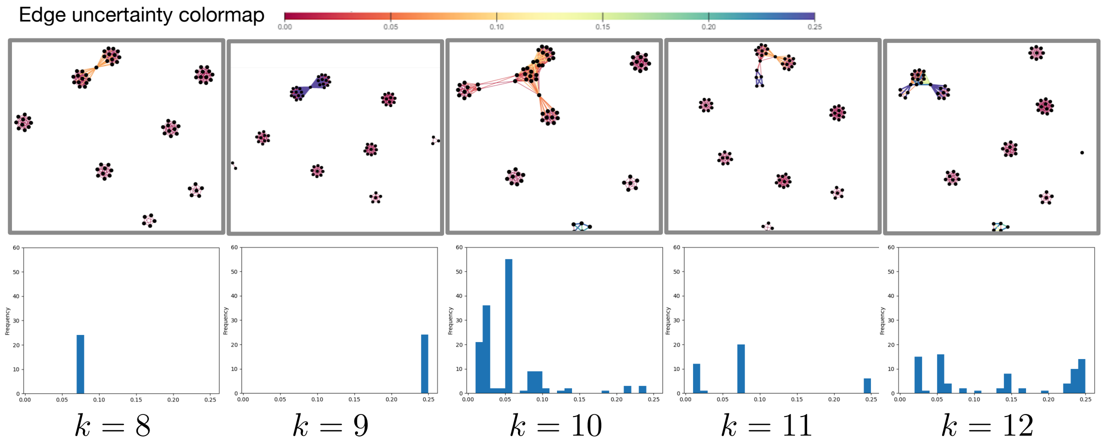

Fig. 4. *Les Miserables* dataset. Top: A^* graphs for $k = 8, 9, 10, 11$, and 12. Bottom: histograms for non-zero edge weights in the A^* graphs.

College Football dataset is 11. We again would like to validate previous results using our uncertainty visualization framework.

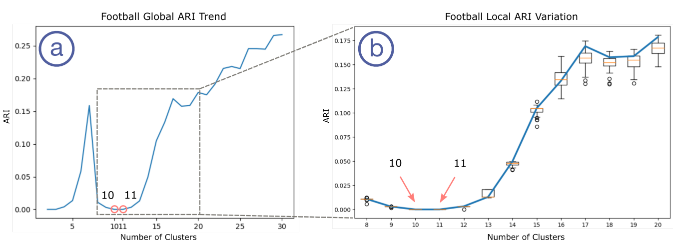

Fig. 5. *College Football* dataset. (a): $\mathcal{ARI}$ with an increasing number of clusters. (b): the variability of $\mathcal{ARI}$ with 8 to 20 clusters.

Global uncertainty with $\mathcal{ARI}$. We apply the NJW algorithm and show the $\mathcal{ARI}$ global trend and the ensemble variability across 100 runs in Fig. 5, varying the number of clusters from 2 to 30. The local minimum of $\mathcal{ARI}$ appears at 11 clusters, with 10 clusters following closely. Both clusterings also display little to none local variability across the ensembles (Fig. 5b).

Local uncertainty with $\mathcal{LAR}$. The reduced graph with 11 clusters in Fig. 6 identifies the 11 conferences mostly accurately, with the independent teams spread out in various clusters. The independent teams do not form a tight cluster because they played few games against each other, whereas the teams from the same conference played mostly among themselves. For example, teams in the Big Twelve played a total of 82 games and 48 of them were against other teams in the same conference. However, the independent teams played a total of 45 games, with only 1 played between two independent teams. In Fig. 6, we visualize the representative graphs with node uncertainty for $k = 9, 10, 11$, and 12. The graphs with 10 and 11 super-nodes show close to zero node uncertainty, whereas the graphs with 9 and 12 clusters display visible orange rings (pointed by orange arrows), indicating an increase in uncertainty.

Cluster memberships. Comparing the membership of the reduced graphs with 10 and 11 clusters, we conclude that the latter captures the community structure better than the former. In the graph with 10 super-nodes (Fig. 7a), the Sun Belt teams (highlighted by black circles) are spread across multiple clusters. Although this result is justifiable since the

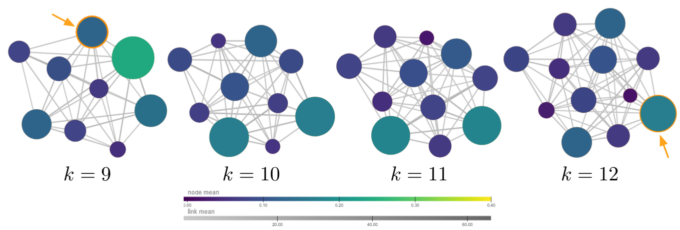

Fig. 6. *College Football* dataset: visualizing node uncertainty of representative graphs with 9, 10, 11, and 12 clusters, respectively.

Sun Belt teams played more games outside the conference (45 games) than among the conference teams (10 games), the clustering with 11 super-nodes separate the 11 conferences with higher accuracy. In particular, Fig. 7c contains a new red cluster that includes 4 out of the 7 Sun Belt teams, cf., Fig. 7a.

Co-occurrences with A^* graphs. The A^* graphs provide additional guidance to the selection of k. The A^* graphs with 10 and 11 clusters display well-separated clusters with almost no uncertain edges, indicating stable community structures, see Fig. 8 top. Furthermore, the A^* histograms at $k = 10$ and $k = 11$ contain almost no uncertain bars, in comparison with those at $k = 9$ and 12; see Fig. 8 bottom. These findings align well with the Newman's results [35], that 11 is the appropriate number of communities for this dataset.

C. Co-authorship Dataset

The *Co-authorship* dataset, compiled by Newman in [33], is a graph of scientists conducting research on network related topics. We take the largest connected component of the graph consisting of 379 scientists, where an edge is drawn between two nodes if the corresponding scientists co-authored a publication; the edge weight captures the strength of the collaboration (computed by combining the number of papers coauthored and the number of coauthors on the paper, see [33, section V] for details).

Global uncertainty with $\mathcal{ARI}$. Using our framework, we apply the NJW algorithm to this dataset, varying the number of clusters from 2 to 50. The global $\mathcal{ARI}$ trend shown in Fig. 9a displays two local minima far apart from each other, at

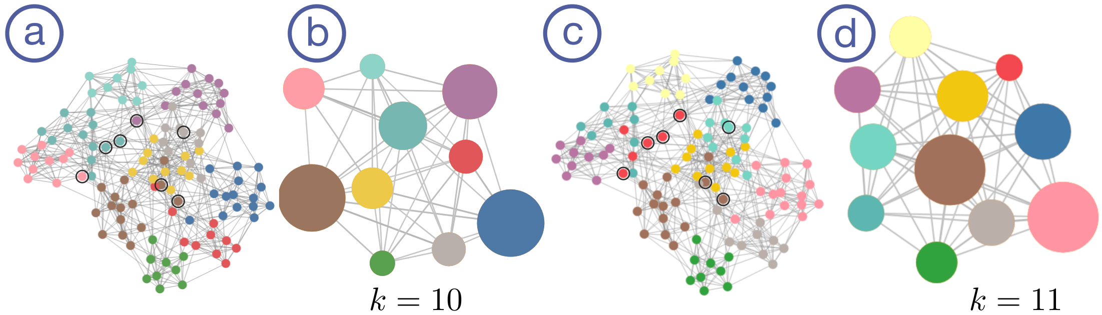

Fig. 7. *College Football* dataset: the original and the reduced representative graphs with 10 (a-b) and 11(c-d) clusters. Nodes are colored by cluster memberships.

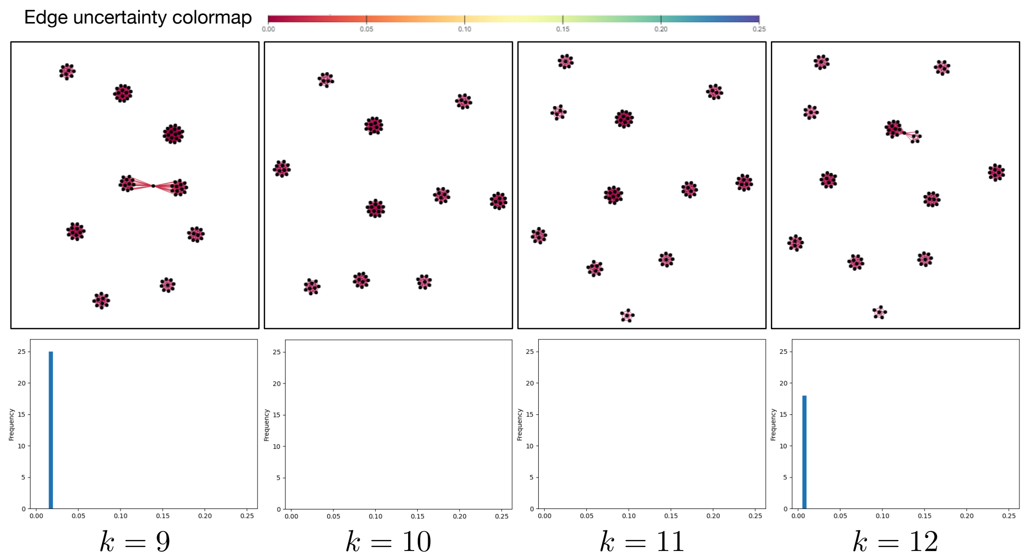

Fig. 8. *College Football* dataset. Top: A^* graphs for $k = 9, 10, 11$, and 12. Bottom: histograms for non-zero edge weights.

10 and 29 clusters. Both clusterings also have low ensemble variability shown in Fig. 9b and Fig. 9c, respectively.

Local uncertainty with $\mathcal{LAR}$. We investigate this global trend more closely by visualizing the representative graphs with 9, 10, 11, 28, 29, and 30 clusters in Fig. 10a. We observe a similar behavior as the previous datasets: as we deviate from 10 and 29 clusters, the orange rings around some clusters thicken, i.e., the local uncertainty for these corresponding clusters increases.

Cluster memberships. The two clusterings at $k = 10$ and $k = 29$ are closely related, see Fig. 10b-c. The nodes in the two graphs are assigned the same color when a cluster in Fig. 10c is completely contained in a cluster in Fig. 10b. This means that all clusters but one (i.e., the black super-node) from the representative graph at $k = 29$ merge to form clusters in the representative graph at $k = 10$. In other words, almost all clusters in Fig. 10c are refinements of clusters in Fig. 10b. An in-depth analysis reveals that clusters in Fig. 10b are often composed of multiple related research topics, while clusters in Fig. 10c provide further refinement of these topics. For instance, when $k = 10$, cluster 1 consists of scientists working on biological networks. The corresponding clusters when $k = 29$ each focus on one of the following topics: ecological networks, bacterial and neuronal networks, human and animal population networks, and social network analysis using statistical physics.

Co-occurrences with A^* graphs. Finally, A^* graphs for $k = 10$ and $k = 29$ show more separation among clusters, indicating more stable cluster memberships, compared to their neighbors; see Fig. 11 top. Histograms with 10 and 29 clusters (Fig. 11 bottom) mainly display one bar at a low uncertainty value, indicating stability w.r.t. co-occurrences.

D. LastFM Asia Dataset

To demonstrate the effectiveness of our framework on larger datasets, we work with the *LastFM Asia* dataset [41] collected in March 2020. It is a social network of the music streaming service LastFM users in Asia. The nodes represent the 7,624 users and an edge is drawn if there is a mutual follower relationship between pairs of users. There are 18 ground-truth communities, labeled by the users' home countries (although the exact country names are not provided).

Global uncertainty with $\mathcal{ARI}$. We apply the NJW algorithm by varying the number of clusters from 2 to 30. Similar to the result of the *Co-authorship* dataset, we observe more than one local minima, at $k = 12$, 15, and 18, see Fig. 12a. All three local minima also display low ensemble variability in Fig. 12b.

Local uncertainty with $\mathcal{LAR}$. We visualize the representative graphs and their associated uncertainty with the number

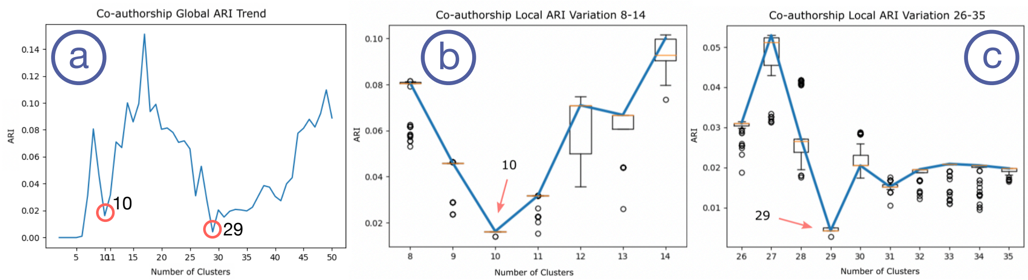

Fig. 9. *Co-authorship* dataset: (a) $\mathcal{ARI}$ with an increasing number of clusters; the variability of $\mathcal{ARI}$ with (b) 8 to 14 and (c) 26 to 35 clusters.

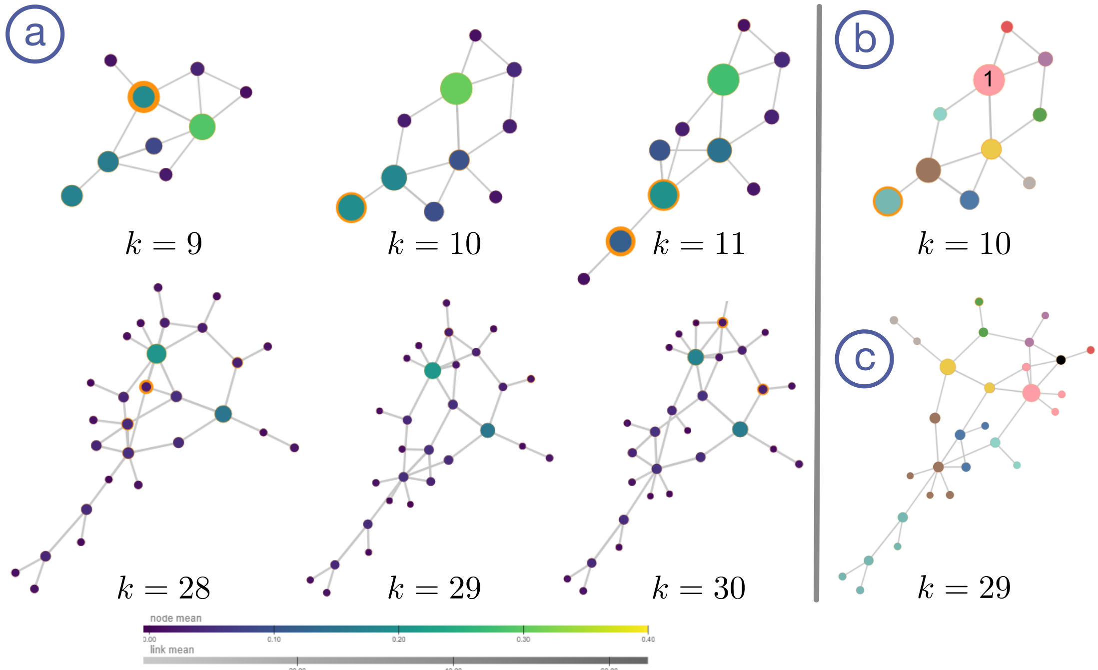

Fig. 10. *Co-authorship* dataset: (a) visualizing node uncertainty of representative graphs with 9 – 11, 28 – 30 clusters, respectively; representative graphs with 10 (b) and 29 (c) clusters, where nodes are colored by cluster memberships.

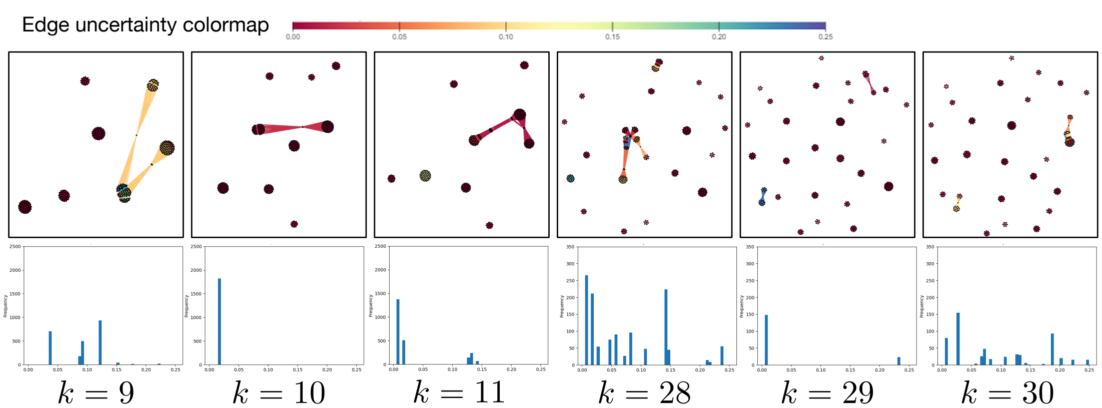

Fig. 11. *Co-authorship* dataset. Top: A^* graphs for $k = 9, 10, 11, 28, 29,$ and 30. Bottom: histograms for non-zero edge weights.

of clusters ranging from 11 to 19 in Fig. 13a. We focus our attention on the cluster (labeled cluster 1) in each clustering that contributes the most to the global uncertainty, that is, the cluster with the largest radius which reflects the largest mean $\mathcal{LAR}$. These clusters contain almost identical members across all representative graphs with $k = 11$ to 19. Compared to their immediate neighbors, clusterings with 12, 15, and 18 clusters display thinner orange rings, especially around cluster 1, indicating more stable community structures. This observation aligns with the observations from previous datasets.

Cluster memberships. We discover that the the cluster members of the representative graph with $k = 18$ correspond well with the ground-truth communities. From the ground-truth, 13 out of the 18 communities have an average of 80% overlap with a cluster in the representative graph. These 13 clusters

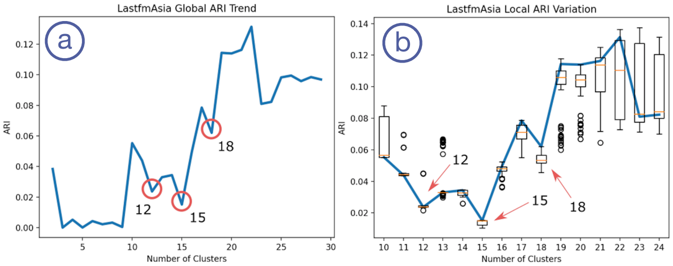

Fig. 12. *LastFM Asia* dataset. (a) $\mathcal{ARI}$ with an increasing number of clusters. (b) The variability of $\mathcal{ARI}$ with 10 to 24 clusters.

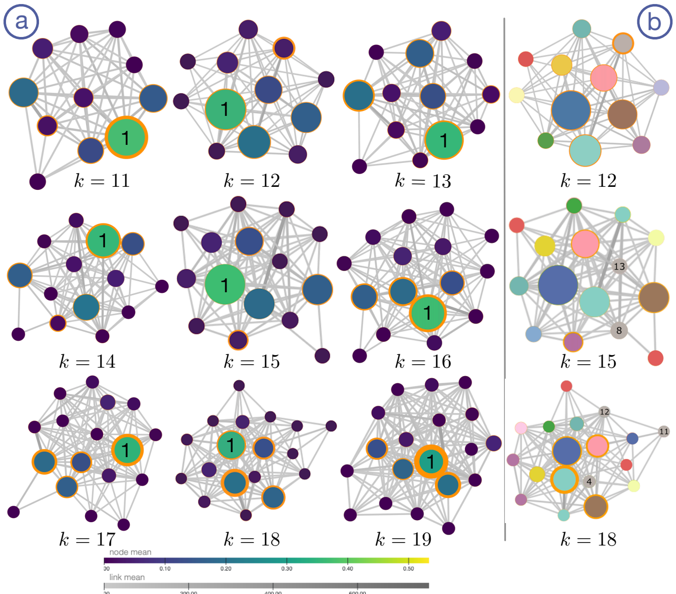

Fig. 13. *LastFM Asia* dataset. (a) Visualizing node uncertainty of representative graphs with 11 – 13, 14 – 16, 17 – 19 clusters, respectively. (b) Representative graphs with 12, 15, 18 clusters, where nodes are colored by cluster memberships.

consist of 89% of the total 7,624 nodes.

We further investigate the relationship among the three clusterings with $k = 12$, 15, and 18 in Fig. 13b. When $k = 18$, 6 small clusters merge to form 3 clusters in the representative graph with $k = 15$. For instance, clusters 11 and 12 (for $k = 18$) form cluster 13 (for $k = 15$); cluster 4 (for $k = 18$) maps into cluster 8 (for $k = 15$). Similarly, the representative graph with $k = 15$ contains 6 clusters that merge into 3 cluster when $k = 12$. For instance, clusters 8 and 13 (for $k = 15$) merge and form cluster 10 (for $k = 12$). Fig. 13b illustrates this relationship by encoding the corresponding clusters among the three representative graphs using the same color.

VI. Conclusion and Discussion

In this paper, we present two complementary uncertainty measures based on local adjusted Rand indices and co-occurrences to quantify and visualize uncertainty associated with a randomized spectral clustering algorithm. We demonstrate that these two uncertainty measures complement each other to serve as an empirical guide for the selection of the appropriate number of clusters. These measures can also be generalized to any randomized graph coarsening algorithms.

We have considered using a user study to evaluate our framework in addition to our quantitative evaluation. However, we concluded that a traditional user study approach is infeasible for a specialized tool that we provide. The targeted users of our framework are data scientists with expertise in network reduction algorithms such as spectral clustering and community detection. A user study with the general population will not provide insights relevant to our technical contribution. On the other hand, a user satisfaction survey with a small sample of specialized participants could suffer from significant bias because of participants' prior knowledge of the datasets and the desired outcome. Therefore, we chose to demonstrate the effectiveness of our framework by applying it to datasets from different domain areas. We argue that the presented quantitative evaluation is a good initial approach to establish the credibility of the proposed framework. Nevertheless, a recent work by Jefferson et al. [23] presented a technique that uses conjoint analysis to quickly conduct expert elicitation. Applying the method to our framework could be an interesting direction in the future.

Our framework serves as a proof of concept that uncertainty visualization can be used to improve interpretability and confidence in graph reduction tasks. Although we have used small to medium size datasets in this paper, our framework is applicable to larger datasets. However, one obstacle is scalable computation, as our current framework requires multiple runs of the same randomized algorithm. Another obstacle we encountered is that many of the existing large networks often lack ground truth communities, making it hard to evaluate the resulting visualization. Developing benchmark large networks with curated ground truth communities is something that would benefit the whole network analysis community.

Our approach is currently developed for ensembles generated from multi-runs of a single graph coarsening algorithm. It may be generalized to multiple algorithm scenarios. Finally, the matrix Q^t defined in Sect. IV is closely related to the matrix used in modularity computations [34]; establishing a formal connection between co-occurrences and modularity, and using such a connection for uncertainty visualization would be interesting.

Acknowledgment

This work was partially supported by DOE DE-SC0021015, NSF IIS-1910733, NSF IIS-2145499, and NSF IIS-1513616.

References

[1] Charu C. Aggarwal and Philip S. Yu. A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623, 2009.

[2] Fabian Beck, Michael Burch, Stephan Diehl, and Daniel Weiskopf. The state of the art in visualizing dynamic graphs. In *EuroVis - STARs*, 2014.

[3] Ken Brodlie, Rodolfo Osorio Allendes, and Adriano Lopes. A review of uncertainty in data visualization. *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109, 2012.

[4] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *International Conference on Learning Representations*, 2014.

[5] M. Charikar, T. Leighton, S. Li, and A. Moitra. Vertex sparsifiers and abstract rounding algorithms. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 265–274, 2010.

[6] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. HARP: hierarchical representation learning for networks. *AAAI Conference on Artificial Intelligence*, 2018.

[7] Jie Chen and Ilya Safro. Algebraic distance on graphs. *SIAM Journal on Scientific Computing*, 33(6):3468–3490, 2011.

[8] Christopher Collins, Sheelagh Carpendale, and Gerald Penn. Visualization of uncertainty in lattices to support decision-making. In *Proceedings of the 9th Joint Eurographics / IEEE VGTC Conference on Visualization*, pages 51–58, 2007.

[9] Carlos D. Correa, Yu-Hsuan Chan, and Kwan-Liu Ma. A framework for uncertainty-aware visual analytics. *IEEE Symposium on Visual Analytics Science and Technology*, 2009.

[10] F. Dorfler and F. Bullo. Kron reduction of graphs with applications to electrical networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 60(1):150–163, 2013.

[11] Peter Eades and Roberto Tamassia. Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry: Theory and Applications*, 4(5):235–282, 1994.

[12] Sonia Fiol-González, Cássio de Almeida, Ariane Rodrigues, Simone Barbosa, and Hélio Lopes. Visual exploration tools for ensemble clustering analysis. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 3, pages 259–266, 2019.

[13] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[14] Alexander J Gates, Ian B Wood, William P Hetrick, and Yong-Yeol Ahn. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, 9(1):8574, 2019.

[15] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[16] Cyril Goutte, Lars Kai Hansen, Matthew G. Liptrot, and Egill Rostrup. Feature-space clustering for fmri meta-analysis. *Human Brain Mapping*, 13(3):165–183, 2001.

[17] Ronny Hadany and David Harel. A multi-scale algorithm for drawing graphs nicely. In *Graph-Theoretic Concepts in Computer Science*, pages 262–277. Springer Berlin Heidelberg, 1999.

[18] David Harel and Yehuda Koren. A fast multi-scale method for drawing large graphs. In *Graph Drawing*, pages 183–196, 2001.

[19] David Harel and Yehuda Koren. A fast multi-scale method for drawing large graphs. *Journal of Graph Algorithms and Applications*, 6(3):179–202, 2002.

[20] Ivan Herman, Guy Melancon, and M. Scott Marshall. Graph visualization and navigation in information visualization: a survey. *IEEE Transactions of Visualization and Computer Graphics*, 6(1):24–43, 2000.

[21] Fang Hu, Yanhui Zhu, Jia Liu, and Yalin Jia. Computing communities in complex networks using the dirichlet processing gaussian mixture model with spectral clustering. *Physics Letters A*, 383(9):813–824, 2019.

[22] Yifan Hu. Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, 10(1):37–71, 2005.

[23] Brett A. Jefferson, Natalie C. Heller, Joseph A. Cottam, Nhuy Van, and George Chin. The utility in conjoint analysis as a fast expert elicitation technique. In *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6, 2021.

[24] Yu Jin, Andreas Loukas, and Joseph F. JaJa. Graph coarsening with preserved spectral properties. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 4452–4462, 2020.

[25] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

[26] B. Kulis, I. S. Dhillon, and Y. Guan. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

[27] Bongshin Lee, George G Robertson, Mary Czerwinski, and Cynthia Sims Parr. CandidTree: visualizing structural uncertainty in similar hierarchies. *Information Visualization*, 6(3):233–246, 2007.

[28] Andreas Loukas. Graph reduction with spectral and cut guarantees. *Journal of Machine Learning Research*, 20(116):1–42, 2019.

[29] Andreas Loukas and Pierre Vandergheynst. Spectrally approximating large graphs with smaller graphs. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3237–3246, 2018.

[30] Marina Meilă. Comparing clusterings — an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[31] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based methodfor class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118, 2003.

[32] Mariá C.V. Nascimento and André C.P.L.F. de Carvalho. Spectral methods for graph clustering – a survey. *European Journal of Operational Research*, 211(2):221–231, 2011.

[33] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

[34] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[35] M. E. J. Newman and Gesine Reinert. Estimating the number of communities in a network. *Phys. Rev. Lett.*, 117:078301, 2016.

[36] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 849–856, 2001.

[37] Dan Pelleg and Andrew Moore. X-means: extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning*, pages 727–734, 2000.

[38] K. Potter, P. Rosen, and C.R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. *Uncertainty Quantification in Scientific Computing*, 377:226–249, 2012.

[39] Dorit Ron, Ilya Safro, and Achi Brandt. Relaxation-based coarsening and multiscale graph organization. *Multiscale Modeling & Simulation*, 9(1):407–423, 2011.

[40] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[41] Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 13251334, New York, NY, USA, 2020. Association for Computing Machinery.

[42] Ilya Safro, Peter Sanders, and Christian Schulz. Advanced coarsening schemes for graph partitioning. *Journal of Experimental Algorithmics*, 19(1):1–24, 2015.

[43] Christoph Schulz, Arlind Nocaj, Jochen Goertler, Oliver Deussen, Ulrik Brandes, and Daniel Weiskopf. Probabilistic graph layout for uncertain network visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):531–540, 2017.

[44] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[45] Catherine A Sugar and Gareth M James. Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463):750–763, 2003.

[46] Corinna Vehlow, Thomas Reinhardt, and Daniel Weiskopf. Visualizing fuzzy overlapping communities in networks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2486–2495, 2013.

[47] Tatiana Von Landesberger, Arjan Kuijper, Tobias Schreck, Jörn Kohlhammer, Jarke J van Wijk, J-D Fekete, and Dieter W Fellner. Visual analysis of large graphs: state-of-the-art and future research challenges. *Computer Graphics Forum*, 30(6):1719–1749, 2011.

[48] Silke Wagner and Dorothea Wagner. Comparing clusterings - an overview. *Technical Report 2006-04*, 2007.

[49] C Walshaw. A multilevel algorithm for force-directed graph drawing. In *Graph Drawing*, pages 171–182. Springer Berlin Heidelberg, 2001.

[50] Chris Walshaw. A multilevel algorithm for force-directed graph-drawing. *Journal of Graph Algorithms and Applications*, 7(3):253–285, 2003.

[51] Chaoli Wang and Jun Tao. Graphs in scientific visualization: a survey. *Computer Graphic Forum*, 36(1):263–287, 2017.

[52] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 43(1):4–24, 2021.

[53] Lin Yan, Yusu Wang, Elizabeth Munch, Ellen Gasparovic, and Bei Wang. A structural average of labeled merge trees for uncertainty visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):832–842, 2020.