Visual Evaluation for Attribute Differences in Graph Sampling

Yong Zhang
Hangzhou Dianzi University

Yuqi Zhou
Hangzhou Dianzi University

Jiajia Kou
Hangzhou Dianzi University

Yuhua Liu
Hangzhou Dianzi University

Yongheng Wang
Zhejiang Lab

Xiangyang Wu
Hangzhou Dianzi University

Zhiguang Zhou
Hangzhou Dianzi University

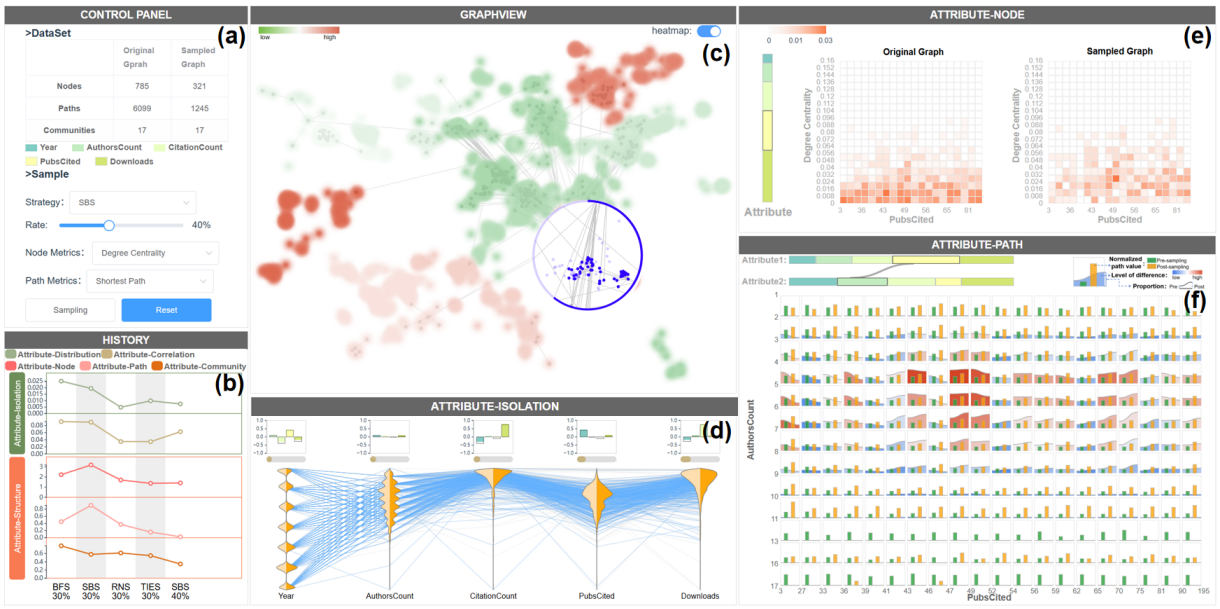

Figure 1: The interface of our visual system: (a) a control panel for adjusting sampling strategies, rates, and other parameters; (b) a history panel displaying differences between sampling results; (c) a graph view showing node-link topology with heatmap overlay for Attribute-Community difference distribution; (d) an Attribute-Isolation difference view presenting attribute changes via parallel coordinate plot and area maps; (e) an Attributes-Node panel visualizing correlations between node attributes and topological metrics through matrix heatmaps; (f) an Attribute-Path graph enabling visual exploration of attribute-path association changes.

Abstract

With the increasing scale and complexity of graph data, graph sampling has become a crucial dimensionality reduction technique, while the evaluation of its effectiveness has also garnered significant attention. However, traditional sampling evaluation methods primarily focus on preserving topological structures while neglecting the integrity of node attributes. To address this, we introduce an attribute-aware evaluation framework for assessing graph sampling differences. First, a novel taxonomy is proposed to categorize sampling differences into two major types: Attribute-Isolation and Attribute-Structure. The former is further divided into Attribute-Distribution and Attribute-Correlation, while the latter includes three subcategories: Attribute-Node, Attribute-Path and Attribute-Community. Based on this, we design a set of computable metrics to quantify each type of attribute-related differences. Furthermore, we develop an interactive visualization system that integrates multiview visual modules to visualize sampling impacts on attributes and structural correlations. Case studies and quantitative evaluations demonstrate the effectiveness of our method in characterizing the impact of different sampling strategies on attribute information and aiding users in network analysis and strategy selection.

Index Terms: Graph sampling, graph difference, visual evaluation.

1 Introduction

Attribute graphs can effectively model complex entity relationships while incorporating supplementary data through node attributes [40]. As these graphs gain prominence in social networks, bioinformatics, and finance [10, 51, 8], their increasing scale and complexity impose significant computational and storage challenges on traditional graph algorithms, hindering real-time performance and scalability for downstream applications. Under these conditions, various sampling strategies [20] have been developed to extract representative subgraphs or node subsets, thus alleviating computational overhead. These strategies provide an effective means of graph reduction, constituting a critical methodology for large-scale graph analysis and visualization [47]. However, sampling inherently omits subsets of nodes and edges, potentially distorting structural and attribute-related patterns. Such distortions may compromise both the representativeness of the sampled graphs and the reliability of subsequent analyses.

Numerous metrics, including degree distribution [17], shortest path [20], and clustering coefficient [48], have been proposed to evaluate structural differences between the sampled graphs and their original counterparts. However, conventional evaluation frameworks predominantly focus on topological preservation while overlooking alterations in node attribute information, which is a

critical oversight given the prevalence of attribute-rich nodes in real-world networks. In social networks, for example, the nodes typically contain demographic attributes (e.g., gender, age, education level, and income)[56], while the protein nodes in bioinformatics networks incorporate biological descriptors such as sequence data and structural identifiers [38]. These attributes not only define graph semantics but also exhibit intrinsic correlations with network topology that are essential for comprehensive network analysis [3]. The prevailing structural similarity paradigm fails to account for sampling-induced distortions in attribute distribution and its structural interdependence, potentially yielding biased analytical outcomes. For example, sampling strategies such as SB and FF preserving structural features may amplify specific attribute categories through selection bias, introducing systematic errors in downstream tasks like node classification. This loss of latent information remains undetectable by traditional structural metrics [23].

Therefore, this study aims to systematically assess and visually reveal information differences between the sampled and original graph from the perspective of node attributes. Through multiple rounds of interviews with graph domain experts and synthesis of their feedback, we concluded that this approach can help researchers identify the sources of attribute differences in graph sampling and inform strategy selection. We identify three core challenges: **CH1:** How to establish a systematic attribute-centric taxonomy that differentiates types while comprehensively covering both intrinsic attribute variations and multi-granularity attribute-structure consistency shifts, **CH2:** How to design quantifiable metrics to objectively assess sampling impacts on attribute information and attribute-structure relationships and **CH3:** How to construct visual mappings for diverse difference types to enable hierarchical tracing from global distributions to local associations.

To address these challenges, we propose a systematic taxonomy and quantification framework to evaluate differences between sampled and original graphs by analyzing node attributes and their relationships with topological structures. In addition, we develop a visual analytics system to enable fine-grained analysis of sampling-induced variations and facilitate source tracking of observed differences. Specifically, we categorize the impacts of graph sampling on attributes into two classes: Attribute-Isolation, which focuses on intrinsic attribute variations, and Attribute-Structure, which examines attribute-topology consistency. The Attribute-Isolation class comprises two subcategories: Attribute-Distribution and Attribute-Correlation. The Attribute-Structure class evaluates attribute-topology coherence across three structural levels: Attribute-Node, Attribute-Path, and Attribute-Community **(CH1)**. Building on this taxonomy, we design domain-specific quantification metrics to systematically measure differences across these categories, ensuring a comprehensive and reproducible analysis **(CH2)**. Furthermore, we implement an interactive visual analytics system that integrates multiview visual modules to holistically visualize sampling impacts on attributes and structural correlations. The system supports hierarchical exploration of differences, allowing users to progressively investigate sampling effects from global trends to localized anomalies while interactively tracing their root causes **(CH3)**. The framework's effectiveness is ultimately verified through case studies and quantitative evaluations. The contributions of our work are:

- We integrate node attributes into sampling difference analysis and propose two major categories with finer subcategories to establish a clear taxonomy for information comparison.
- We construct a suite of quantifiable metrics tailored to attribute differences across the defined categories.
- We develop a visual analytics system that supports multiview exploration to intuitively present hierarchical sampling differences and systematically facilitate fine-grained analysis.

2 Related Work

2.1 Graph Sampling Technique

Graph sampling aims to extract representative subgraphs from large-scale networks to reduce computational and storage costs while preserving essential structures and attributes [46, 20, 2]. Sampling methods are typically categorized into node-based, edge-based, and traversal-based approaches [20, 39]. Node-based sampling selects a subset of nodes and retains the edges between them. Random Node (RN) sampling chooses nodes uniformly at random [17, 45], while importance-aware variants such as Random Degree Node (RDN) [4] and Random PageRank Node (RPN) [27] prioritize nodes based on structural metrics. Edge-based sampling focuses on selecting edges and including their incident nodes to build subgraphs, with strategies like Random Edge (RE) [45, 31], Random Node Edge (RNE) [27, 22], and Random Edge Node (REN) [50, 36, 19] differing in sampling order and expansion logic. Traversal-based methods explore the graph using random walks or similar strategies, constructing subgraphs from visited nodes and edges [20]. Random Walk Sampling (RWS) [29, 9] may suffer from local trapping [26], leading to enhanced variants such as Random Walk with Jump (RJ) [44, 49] for improved global coverage. Other traversal strategies—including Breadth-First Sampling (BF) [24, 25], Depth-First Sampling (DF) [11, 32], Snowball Sampling (SB) [44, 18], and Forest Fire Sampling (FF) [28]—further enhance structural preservation and connectivity.

Despite advances, evaluating subgraph representativeness remains essential. We present an attribute-based framework to quantify sampling distortions, enhanced by multi-view interactive visualization for hierarchical exploration and strategy comparison.

2.2 Graph Sampling Evaluation

The diversity of graph properties necessitates varied evaluation metrics for graph sampling. From a structural perspective, evaluation metrics can be grouped into degree-based, path-based, and structure-based types. Degree-based metrics reflect changes in node importance by assessing neighborhood properties, including degree [46, 48], degree distribution [42], degree centrality [48], and eigenvector centrality [6]. Path-based metrics capture global positional significance via betweenness centrality [12], closeness centrality [15], and node connectivity [43]. Structure-based metrics quantify local structural patterns through clustering coefficient [42, 16], triangle count [5], and subgraph centrality [13].

Beyond topology, rich node attributes such as user profiles in social networks, render purely structural evaluations insufficient [37]. To address this, Seufert et al. [37] introduced joint metrics incorporating attribute and degree similarity. Wagner et al. [41] proposed combining attribute and structural evaluations using Top-k bias and Normalized Cumulative Group Relevance (nCGR). Lin et al. [30] developed AB Sampling and AB-RIS, measuring attribute preservation and structural deviation via attribute distribution and connectivity characteristics.

Graph sampling also influences visual representations. Some studies emphasize visual perception metrics over structural accuracy. Wu et al. [44] identified eight perceptual factors including area coverage, cluster quality, and visibility of high-degree nodes. Zhang et al. [48] introduced spatial and cluster-based visual criteria, while Nguyen et al. [35] proposed visual metrics tailored for proxy graphs to assess representational fidelity.

While most evaluations emphasize topology, attributes remain undervalued. Our work fills this gap by analyzing attribute distributions, their correlations, and alignment with structure in sampling.

2.3 Attribute Features of Graph

In attribute graphs, nodes and edges are associated with attribute, where each attribute represents a specific feature (e.g., user profiles, interests, or topics in social networks) [21]. This rich at-

tribute information facilitates applications in social network analysis [10], recommendation systems[14], and biological studies [51]. Researchers analyze attribute characteristics to optimize feature preservation in graph computations and improve downstream tasks.

Without considering the possible relationship with the structure, the attribute characteristics of a graph can be reflected by the distribution of attributes and the associations between them. Attribute distributions provide statistical summaries of attribute sets. Kumar et al. [23] proposed Information Expansion Sampling (IXS), which prioritizes nodes with diverse attributes to preserve distributional features, evaluated using metrics like the Kolmogorov-Smirnov (KS) statistic. Lin et al. [30] introduced Attribute Deviation (AD) to measure how well sampling maintains original attribute proportions. Attribute correlations capture linear/nonlinear relationships between attributes, revealing structural and semantic patterns. Meng et al. [34] developed Coupled Node Similarity (CNS), quantifying node similarity via co-occurrence matrices and conditional probabilities to model complex attribute relationships.

Since attributes are inherently tied to topology, their interdependence necessitates examining attribute-structure relationships at node, edge, and community levels. The joint distribution of node attributes with topological metrics (e.g., degree, centrality) [37] reveals feature patterns (e.g., high-degree node attribute concentration implies hub association) [1], which also facilitate comparative analysis of pre- and post-sampling differences. At the edge level, node attributes exhibit connectivity dependencies [7], where edge changes can alter attribute-related shortest path lengths and connectivity. For community-level, local attribute characteristics are typically evaluated through attribute importance and distribution [33].

Given the importance of node attributes, this work quantifies sampling discrepancies via intrinsic attribute traits and their structural correlations, forming a hierarchical framework from global distributions to local patterns.

3 Task Analysis and System Overview

3.1 Task analysis

To assess the differences in node attributes due to graph sampling and develop a visualization system that intuitively presents and explores the impact of sampling on attribute information, we consulted with two domain experts, E1 and E2. E1 is an experienced scholar specializing in graph analysis and interactive visualization, with extensive experience in data visualization and leadership in multiple research projects involving large-scale network data. E2 is a senior data analyst from a renowned international IT company, with long-term involvement in complex network analysis, focusing on the impact of network sampling on structure and attributes, and with rich practical experience in social networks, financial risk control, and bioinformatics. Over two years of collaboration and their feedback helped us identify four core tasks:

T1: Establish a systematic framework for classifying attribute differences. Current graph sampling evaluations mainly emphasize topology while overlooking systematic analysis of node attributes. Sampling can cause attribute-level changes like distribution shifts and altered correlations, which are hard to capture from a single perspective. Designing separate methods for each case is inefficient and harms system readability. A unified framework is needed to classify and guide the analysis of attribute changes, enabling consistent quantification and visualization.

T2: Integrate attribute–topology associations. Node attributes are closely tied to graph topology, and many tasks like interest propagation or financial risk assessment rely on their interplay. Focusing only on attribute changes offers an incomplete view of sampling impacts. It's crucial to examine how attributes interact with topology and how this relationship may shift after sampling. Domain experts stress that analyzing attributes or structure in isolation risks information loss. Thus, joint analysis of both should be integrated into the evaluation to ensure a comprehensive assessment.

T3: Design quantitative evaluation metrics. Existing evaluations of attribute differences rely on qualitative descriptions, lacking measurable standards. To help users assess the impact of these differences on analytical tasks, numerical metrics are essential. While topological metrics are well-established, quantitative measures for node attributes remain underdeveloped. Enhancing sampling assessment rigor requires computable metrics that capture both attribute variations and their interactions with topology.

T4: Develop an interactive visual analytics system. Sampling's impact on node attributes is often complex and non-intuitive, making numerical statistics alone inadequate. An interactive visual analytics system is essential for multilevel exploration of how sampling affects attribute information and its structural relationships. Interaction should also support identifying the sources of differences.

3.2 System overview

In line with the aforementioned task, we propose a novel evaluation method to assess node attribute differences induced by graph sampling, aiming to investigate its impact on attribute information. The overall research framework is illustrated in Fig. 2.

First, we systematically analyze how graph sampling influences attribute information and establish a taxonomy of attribute differences. This framework categorizes sampling-induced changes into two primary classes: Attribute-Isolation and Attribute-Structure. The former examines intrinsic attribute alterations, including distribution shifts and correlation changes, while the latter evaluates attribute-structure consistency across three levels: Attribute-Node, Attribute-Path, and Attribute-Community (*R1*, *R2*). Building on this framework, we develop quantification methods tailored to different categories and data characteristics, measuring changes before and after sampling (*R3*). To enhance interpretability, we design an interactive visualization system (*R4*) featuring a multiview layout with heatmaps, parallel coordinate plot, network views, etc., enabling intuitive exploration of sampling effects. The system also supports user interaction for detailed analysis of attribute change sources. Finally, we validate its effectiveness through case studies and quantitative evaluations.

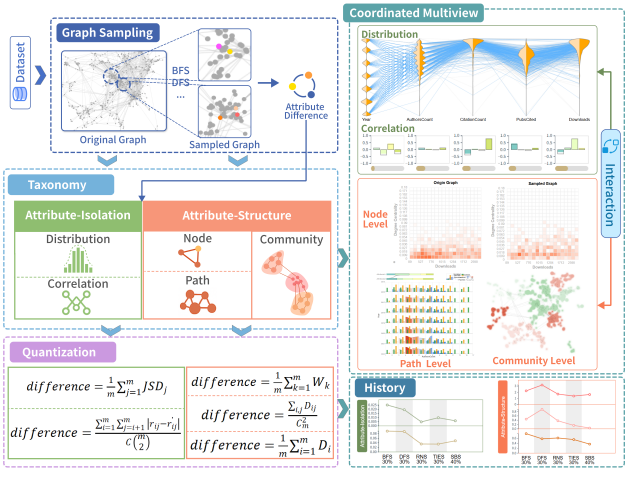

Figure 2: The pipeline of our visual analysis system, including classification of differences, construction of indicator systems, visual design and historical comparison.

4 Assessment of Attribute Difference in Graph Sampling

4.1 Taxonomy of attribute difference in graph sampling

After a comprehensive literature review on graph analysis, we summarized 21 key attribute information that users aim to preserve during graph analysis tasks, which are detailed in Appendix A. Through discussions with two domain experts, based on the frequently occurring node attribute information, we categorized the differences in graph sampling into two main categories: Attribute-Isolation and Attribute-Structure. Attribute-Isolation focuses on variations in node attributes themselves, further divided into Attribute-Distribution and Attribute-Correlation. Attribute-Structure differences examine the consistency of attribute-topology relationships before and after sampling, encompassing three levels: Attribute-Node, Attribute-Path, and Attribute-Community.

(1) Attribute-Isolation: This evaluates inherent changes in node attributes, specifically whether sampling significantly distorts attribute distributions and their correlations. Since many applications depend on node attribute information, sampling may introduce bias or loss of critical attributes, potentially undermining subsequent analyses. We categorize Attribute-Isolation into two aspects: 1) Attribute-Distribution: Sampling may alter the overall distribution of attribute values, including shifts in mean, median, or variance. For example, in social networks, user activity levels may originally follow a specific distribution but could skew toward highly active users after sampling. 2) Attribute-Correlation: Sampling may disrupt relationships between attributes, thereby affecting their interdependencies. For instance, in academic networks, the original strong correlation between a researcher's influence and collaboration count may weaken or disappear after sampling.

(2) Attribute-Structure: Beyond intrinsic attribute changes, sampling also affects the relationship between node attributes and topological structure. Key concerns include whether high-centrality nodes preserve their attribute characteristics and whether nodes sharing attributes maintain close connectivity. We systematically classify these impacts into three types: 1)Attribute-Node: This assesses the preservation of correlations between node attributes and topological features post-sampling. For example, in kinship networks, the typically higher centrality of elder nodes may be disrupted by sampling. 2)Attribute-Path: This evaluates alterations in the association between attribute values and path characteristics. For instance, in academic collaboration networks, the naturally stronger connectivity among scholars with shared research interests may diminish after sampling. 3)Attribute-Community: Communities, as dense node clusters, often represent social groups, research domains, or functional modules, and are key to evaluating representation and clustering methods. This category focuses on the maintenance of attribute distribution within the community. Sampling may retain global attribute distributions but obscure distinct preferences within specific user communities.

4.2 Quantification of attribute differences in graph sampling

In this section, we introduce quantitative metrics to systematically assess changes in attribute information and attribute–structure associations induced by sampling. The datasets used in this study are undirected, unweighted graphs, with node attributes that are either numerical or categorical. Categorical attributes are numerically encoded to ensure the generality and comparability of metric calculations. Given an original graph G with n nodes, and a sampled graph G' with n' nodes, each node is associated with m attributes.

4.2.1 Calculation of Attribute-Isolation Difference

Attribute distribution: For each node attribute, we apply kernel density estimation (KDE) to continuous attributes to partition the value range, capturing distributional patterns while avoiding information loss from fixed binning. As a non-parametric method, KDE smoothly approximates data distributions and provides reliable estimates even with limited samples. For categorical attributes, the categories themselves define the groups, ensuring effective preservation during sampling. Let X denote a given attribute, with its kernel density estimate defined as below:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \tag{1}$$

Here, $K(\cdot)$ is the kernel function, h is the bandwidth parameter, and n is the sample size. We adopt the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$. The optimal bandwidth is selected via cross-validation to obtain the best-fit distribution. After partitioning the attribute into R intervals, we compute the frequency of data points in each interval to construct the attribute probability distribution of the original graph G as $P = (p_1, p_2, \ldots, p_R)$. Similarly, we partition the sampled graph G' in the same way and compute its attribute probability distribution $P' = (p'_1, p'_2, \ldots, p'_R)$. To quantify distributional deviations caused by sampling, we use Jensen–Shannon (JS) divergence, a symmetric and interpretable metric based on Kullback–Leibler (KL) divergence. JS divergence offers greater stability and is widely used for comparing one-dimensional distributions:

$$JSD(P\|P') = \frac{1}{2}KL(P\|M) + \frac{1}{2}KL(P'\|M) \tag{2}$$

Among them, M is the average distribution of P and P'.

After computing the distributional difference for each attribute, we aggregate these values by taking their sum and average to obtain the final attribute distribution difference metric:

$$difference = \frac{1}{m}\sum_{j=1}^{m} JSD_j \tag{3}$$

Attribute-Correlation: Attribute-Correlation measures whether the relationship between two attribute variables remains consistent before and after sampling. To quantify this difference, we employ the Pearson Correlation Coefficient to calculate the correlation between attribute pairs and further evaluate the impact of sampling on their correlation. Given two attributes X and Y in the original graph G, their Pearson Correlation Coefficient is defined as follows:

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i-\overline{X})(y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{N}(x_i-\overline{X})^2}\sqrt{\sum_{i=1}^{N}(y_i-\overline{Y})^2}} \tag{4}$$

We compute the Pearson Correlation Coefficients pairwise between the i-th attribute and the j-th attribute in both G and G', denoted as r_{ij} and r'_{ij}, respectively. By comparing the difference between r_{ij} and r'_{ij}, we derive the Attribute-Correlation metric:

$$difference = \frac{\sum_{i=1}^{m}\sum_{j=i+1}^{m}|r_{ij}-r'_{ij}|}{C\binom{m}{2}} \tag{5}$$

Here, $C\binom{m}{2} = \frac{m(m-1)}{2}$ represents the total number of all possible attribute pairs.

4.2.2 Calculation of Attribute-Structure Difference

Attribute-Node: We denote a node attribute as X and a corresponding topological metric of the node (e.g., degree[46] or eigenvector centrality[6]) as Y. To analyze their relationship, kernel density estimation is applied by partitioning the range of X into a intervals and Y into b intervals, forming a joint probability distribution $a \times b$,

denoted as P. In the original graph, each element of this distribution indicates the proportion of nodes whose X and Y values fall within the i-th and j-th intervals respectively.

Similarly, joint probability distribution P' is computed for the sampled graph. To quantify distributional changes between the original and sampled graphs, we use the Wasserstein distance (Earth Mover's Distance, EMD), a metric from optimal transport theory that considers both distributional differences and transport cost. It is well-suited for comparing multidimensional distributions:

$$W(P,P') = \inf_{\gamma \in \Gamma(P,P')} \sum_{i,j,i',j'} \gamma(i,j,i',j') d((X_i,Y_j),(X_{i'},Y_{j'})) \quad (6)$$

Here, $\Gamma(P,P')$ represents the set of all possible transport plans, that is, the ways to transform the original distribution P into the sampled distribution P'. $\gamma(i,j,i',j')$ denotes the quality of transmission from (X_i,Y_j) to $(X_{i'},Y_{j'})$. $d\big((X_i,Y_j),(X_{i'},Y_{j'})\big)$ is the Euclidean distance between two distribution elements.

Finally, we quantify the overall difference by computing the mean Wasserstein distance between the distributions of all attributes and node topological metrics before and after sampling:

$$difference = \frac{1}{m}\sum_{k=1}^{m} W_k \quad (7)$$

Attribute-Path: A path-based metric between nodes x and y, denoted as $path_{xy}$ (e.g., shortest path length or connectivity[43]), is used. The average shortest path lengths of graphs G and G' are denoted as p and p', respectively. Kernel density estimation is applied to partition the attributes of the nodes A and B into intervals a and b, forming a distribution matrix of attributes. Each element (i,j) in matrix P represents the normalized path length between the pair of nodes whose attributes fall into the intervals A_i and B_j. The definition is given by:

$$P_{ij} = \frac{\sum_{x,y \in S_{ij}} \frac{path_{xy}}{p}}{k} \quad (8)$$

Where S_{ij} denotes the set of node pairs whose attributes fall into intervals A_i and B_j, with k representing the number of such pairs. To account for differences in graph size, path lengths are normalized by the average path length p. The matrix P' is computed similarly for the sampled graph G'. If $n = n' = 0$, indicating no node pairs in this category, both P_{ij} and P'_{ij} are set to 0. If $n \neq 0$ but $n' = 0$, meaning the attribute pair is lost after sampling, P'_{ij} is set to the normalized maximum path length, $\frac{path'_{\max}}{p'}$. The difference between P and P' is then computed using a weighted Euclidean distance, where a weight W_{ij} adjusts the contribution of each attribute pair:

$$D = \sum_{i,j} W_{ij}\sqrt{(P_{ij} - P'_{ij})^2} \quad (9)$$

Here, $W_{ij} = \frac{|S_{ij}|}{|S|}$ represents the proportion of node pairs in the attribute region (A_i,B_j) relative to the total number of path pairs. By introducing W_{ij} as a weight factor, we can reduce the influence of intervals that have a small proportion but exhibit drastic changes on the overall difference, while enhancing the contribution of intervals with a relatively larger proportion. Finally, we compute the average of weighted Euclidean distances over all attribute combinations to obtain the final difference, as shown below:

$$difference = \frac{\sum_{i,j} D_{ij}}{C_m^2} \quad (10)$$

Attribute–Community: We first partition the original graph G into c communities $C_1, C_2, \ldots, C_c$ using the Louvain algorithm. For each community C_i, we estimate the node attribute X distribution via KDE, dividing it into a intervals to obtain the probability distribution $P_i = (p_1, p_2, \ldots, p_a)$. After sampling, we similarly compute the distribution P'_i from the remaining nodes in C_i, using a uniform distribution when the community is empty. The JSD is then used to quantify the difference between P_i and P'_i as $JS_i = JSD(P_i, P'_i)$

Considering the sizes of different communities, we introduce the proportion of community nodes as a weighting factor to ensure that changes from smaller communities do not disproportionately impact the global results. Let R_i denote the proportion of nodes in community C_i relative to the total number of nodes in the original graph. The overall difference of attribute X across all communities is then calculated as the weighted Jensen-Shannon Divergence:

$$D_X = \sum_{i=1}^{c} R_i \cdot JS_i \quad (11)$$

The overall difference is computed by averaging the weighted JSD values across all attributes, yielding a quantitative metric of sampling-induced attribute-community differences, as shown:

$$difference = \frac{1}{m}\sum_{i=1}^{m} D_i \quad (12)$$

5 Visualization of Sampling Differences in Attribute Graphs

To visually present the differences between the original and sampled graphs, we design an attribute-aware visual analytics system for sampling analysis, as shown in Fig. 1.

5.1 Attribute-Isolation Visualization

Parallel coordinates plots are commonly used to visualize multidimensional data and relationships between dimensions. In our Attribute-Isolation visualization design, we integrate parallel coordinates with statistical distribution plots and correlation analysis views to intuitively compare Attribute-Distribution and Attribute-Correlation between original and sampled graphs. As shown in Fig. 1 (d), each vertical axis in the parallel coordinates plot represents a node attribute. Blue polylines denote nodes retained after sampling, connecting their attribute values across dimensions to illustrate their positions in the multidimensional space. Gray polylines represent discarded nodes, providing a reference for comparison and highlighting sampling-induced coverage changes. Adjacent to each attribute axis, area charts visualize distribution shifts before (left) and after (right) sampling. At the top of each axis, stacked bar charts quantify changes in Pearson correlation coefficients between attributes. The filled bars indicate pre-sampling correlation values, while the outlined bars represent post-sampling values. This side-by-side comparison enables clear identification of sampling effects on attribute relationships.

5.2 Attribute-Structure Visualization

Attribute-Node: The attribute-node metrics examine whether the relationships between node attributes and topological metrics (e.g., degree centrality, eigenvector centrality) remain stable after sampling. We employ a matrix heatmap to visualize the joint distribution of attributes and topological metrics in both original and sampled graphs. The stacked histograms on the left side depict the proportion of differences of each attribute. Users can click to select specific attributes and expand their corresponding heatmap views. Dual-view comparison design displays the original joint distribution (left) alongside the sampled distribution (right), facilitating direct comparison of sampling-induced variations. In the heatmaps:

the x-axis represents node attributes, the y-axis represents topological metrics. The color intensity indicates the proportion of nodes that are simultaneously distributed within the specified attribute range and the corresponding structural indicator range, with darker shades indicating higher proportion. Additionally, users can click on any heatmap cell to filter corresponding nodes in both the Node-Link View and parallel coordinates plot for detailed inspection.

Attribute-Path: The Attribute-Path metric captures variations in path characteristics across different node attribute combinations, while also accounting for their proportions among all paths. It helps users assess the consistency of path structures and identify attribute combinations significantly impacted by sampling. To visualize these differences, we designed a glyph view (Fig. 3), where each glyph represents an attribute combination. The internal bar chart shows path metric changes. Height differences reflect path variations. Waveform graph illustrate proportion changes: the left and right heights represent the proportion of paths before and after sampling, respectively. Color intensity encodes the magnitude of Attribute-Path differences. The stacked bar allows users to compare the differences brought by different attributes and the Attribute-Path differences between attributes under a single attribute. By clicking, users can flexibly select attribute combinations for detailed analysis, while glyph interactions highlight corresponding nodes and parallel coordinate segments for focused exploration.

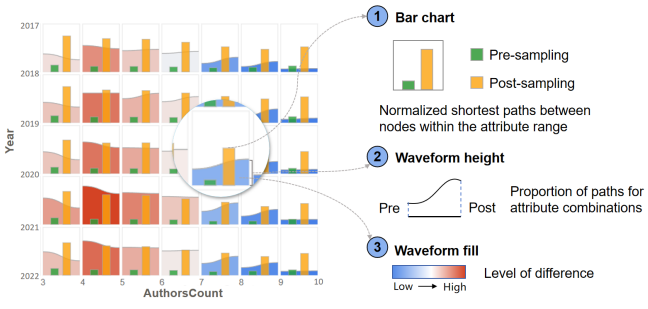

Figure 3: Glyph design and visualization of Attribution-Path differences.

Attribute-Community: A community is inherently a local structure, representing a tightly connected group of nodes within the network. We augment the node-link diagram with heatmap and provide interactive features for inspecting local details. The color gradient from green to red indicates the magnitude of differences, allowing users to quickly identify which communities exhibit significant shifts in attribute distribution during sampling. When a user clicks on a specific community, the system dynamically expands to display its sampling status. Here, a donut chart illustrates the proportion of retained nodes, blue representing preserved nodes, while light blue denotes filtered-out nodes. Simultaneously, the parallel coordinates update to reflect the node retention status of the selected community.

5.3 History Visualization

To facilitate intuitive comparison of how different sampling strategies and rates affect graph quality, we designed a history panel (Fig. 1 (b)). This panel presents line charts that display the values and trends of various metrics across sampling iterations, helping users track the stability and distinctions among sampling results. Each line chart corresponds to a specific metric category, with the x-axis representing different sampling methods and their respective rates, and the y-axis indicating metric values. Users can select a specific sampling iteration by clicking, triggering synchronized updates in other visualization views (e.g., Attribute-Isolation and Attribute-Structure correlation). The history visualization evaluates method stability across conditions, revealing consistent performance or advantages at specific sampling rates and enabling anomaly investigation when needed.

6 Evaluation

6.1 Case Study

To validate the effectiveness of our proposed graph sampling evaluation method and its visualization system, we engaged two domain experts (E1 and E2) in system evaluation and analysis. The experimental dataset comprises the academic network of IEEE VIS conference papers (2015–2021), consisting of 786 nodes and 6,101 edges. Each node represents a paper published at IEEE VIS, and an undirected edge connects two papers if they share at least one common author. Each paper includes five key attributes for analyzing sampling differences: publication year, download count, reference count, citation count, and author count.

6.1.1 Exploring differences under a single sampling method

In this case, we invited E1 to evaluate our visualization system. After loading the graph dataset, E1 selected Degree Centrality as the node topology metric and Shortest Path as the path topology metric. The sampling strategy was configured as Simple Random Walk (SRW) with a 30% sampling rate. The results are shown in Fig. 4 (2). Based on the output, E1 began by examining attribute changes using the Attribute-Isolation panel (Fig. 4 (4)). The parallel coordinates plot revealed distinct distributional differences between sampled and retained nodes across various attributes. The accompanying streamgraph highlighted notable shifts in the distributions of AuthorsCount and Downloads, while the remaining attributes exhibited relatively stable patterns. These observations were further supported by the bar charts above the attribute axes. E1 then explored changes in attribute correlations. The hybrid bar charts revealed an unexpected pattern: while CitationCount and Downloads had weak correlations with Year in the original dataset, they appeared strongly negatively correlated after sampling. E1 noted that this could mislead users into believing that recently published papers tend to have fewer citations and downloads, potentially resulting in inaccurate interpretations.

Next, E1 analyzed the Attribute-Node visualization panel (Fig. 4 (5)). The stacked bar chart indicated significant differences in the relationship between Downloads and Degree Centrality before and after sampling. To investigate further, E1 interactively unfolded the joint distributions of these two attributes. The matrix heatmap revealed a pronounced shift in their correlation. For instance, in the original data, papers with 500–1000 Downloads generally had Degree Centrality values between 0.005 and 0.01, a pattern that did not persist in the sampled data. This disruption decouples attribute information from the graph's structural properties, thereby distorting interpretation of key features.

In the Attribute-Path visualization panel (Fig. 4 (7)), E1 observed via stacked bar charts that the shortest paths between Year and AuthorsCount changed most significantly. A broader examination of Attribute-Path difference revealed that, in nearly all intervals, the ratio of shortest path to diameter increased after sampling. E1 attributed this to the inherent bias of SRW sampling, which favors highly connected nodes, thus disrupting certain path structures and resulting in longer shortest paths in the sampled graph. To identify the source of these differences, E1 examined the heatmap and found that nodes with AuthorsCount = 5 exhibited the most pronounced changes in shortest path length across the Year dimension. A follow-up inspection using the streamgraph indicated that this was due to the dominance of 5-author nodes in the original graph.

In the node-link diagram, E1 used the heatmap (Fig. 4 (3)) to analyze Attribute-Community differences. While most communities showed minimal deviation, community C in the upper-right corner appeared distinctly dark red, indicating significant differences. Clicking this community prompted the system to dynam-

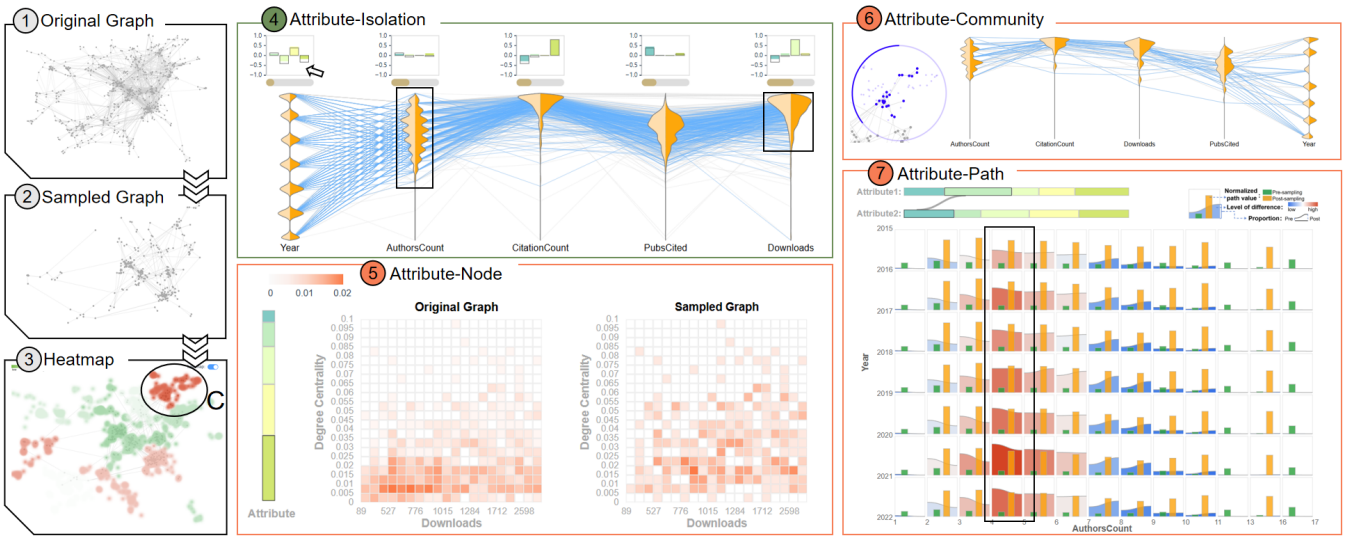

Figure 4: (1)-(3) display the original graph network, sampled graph, and heatmap visualizing Attribute-Community differences respectively; (4) compares Attribute-Distribution and Attribute-Correlation before and after sampling; (5) presents joint distributions of attributes with degree centrality across sampling conditions; (6) provides detailed visualization for in-depth examination of Community C; (7) visualizes Attribute-Path differences in the sampled graph.

ically display its sampling characteristics via a donut chart and a parallel coordinates plot (Fig. 4 (6)). The donut chart showed that approximately 70% of nodes in this community were removed during sampling. The parallel coordinates plot further revealed that this substantial node loss caused marked distributional shifts across all attributes. These findings suggest that although sampling may preserve global attribute distributions, it can significantly distort local community characteristics, potentially leading to misinterpretation.

After completing the analysis, E1 expressed strong approval of our system, stating: "The system's synchronized multi-view visualizations and interactive features effectively reveal hierarchical data differences and enable rapid identification of key variations, supported by evidence explaining their causes." He emphasized that the system facilitates detailed exploration of micro-level attribute differences, enhances understanding of data integrity post-sampling, and improves the reliability of graph-based analysis.

6.1.2 Comparison across different sampling strategies

In this case, E2 evaluated the performance of various sampling techniques in preserving graph attributes using our analytical system. E2 applied four sampling methods: BFS, DFS, RNS, and TIES, each with a fixed 30% sampling rate. The system automatically recorded the results and generated visualizations, as shown in Fig. 5 (1). Comparative analysis revealed that RNS performed best in preserving attribute distributions and showed strong results in maintaining attribute correlations, though it was average on other metrics. E2 attributed this to RNS's uniform node selection probability, which promotes balanced attribute retention. This was further supported by parallel coordinates visualization in Fig. 5 (2).

In contrast, BFS exhibited significant differences in both Attribute-Community and Attribute-Distribution metrics. The heatmap (Fig. 5 (3)) showed that while some communities were largely preserved, others experienced near-complete node loss (marked in dark red). E2 explained that BFS's local exploration from seed nodes tends to retain nearby communities while neglecting those beyond its limited reach, leading to distorted attribute distributions in peripheral regions. DFS, on the other hand, showed the greatest deviation in Attribute-Path relationships. The Attribute-Path view (Fig. 5 (4)) revealed a marked increase in the ratio of shortest path to diameter across nearly all attribute pairs post-sampling, consistent with DFS's deep-path traversal pattern. TIES, however, achieved the most consistent performance across all metrics, particularly minimizing deviations in Attribute-Path. E2 attributed this to TIES's approach of preserving all nodes connected to sampled edges, which effectively maintains the graph's structural integrity and reduces path fragmentation.

After the evaluation, E2 praised the system for efficiently assessing sampling strategies in preserving attribute properties and attribute-topology relationships. He particularly appreciated the history panel's interactive design for intuitive, side-by-side comparisons and in-depth analysis.

6.2 Quantitative analysis

To assess the effectiveness of our proposed metrics in capturing attribute differences between sampled and original graphs, we conducted a quantitative study using a patent citation dataset comprising 4,533 nodes and 7,382 edges. Each node was annotated with five attributes: publication year, country, category, group, and score. In consultation with domain experts, we selected eight of the most used graph sampling strategies in their daily work: Breadth First Sampling(BFS), Depth First Sampling(DFS), Forest Fire(FF), Snowball Sampling(SBS), Random Walk(RW), Random Node Edge(RNS), Random Edge Sampling(RES) and Topology-Induced Edge Sampling(TIES), each evaluated under four sampling rates (5%, 10%, 20%, 30%). Metric values were averaged across rates for comparison. Detailed results are shown in Tab. 1.

The experimental results demonstrate significant variations in performance across different sampling methods when evaluated against various difference metrics. Notably, BFS exhibited the poorest performance across multiple categories, particularly in Attribute-Distribution, Attribute-Node, and Attribute-Community metrics. In contrast, TIES sampling consistently outperformed other methods, demonstrating exceptional stability in maintaining Attribute-Node, Attribute-Path, and Attribute-Community relationships. A detailed analysis reveals that RNS achieved optimal preservation of Attribute-Distributions, while BFS induced the most substantial distributional deviations. This suggests that random node selection better maintains global attribute distributions, whereas BFS's breadth-first expansion leads to uneven sampling by over-representing certain regions and severely neglecting

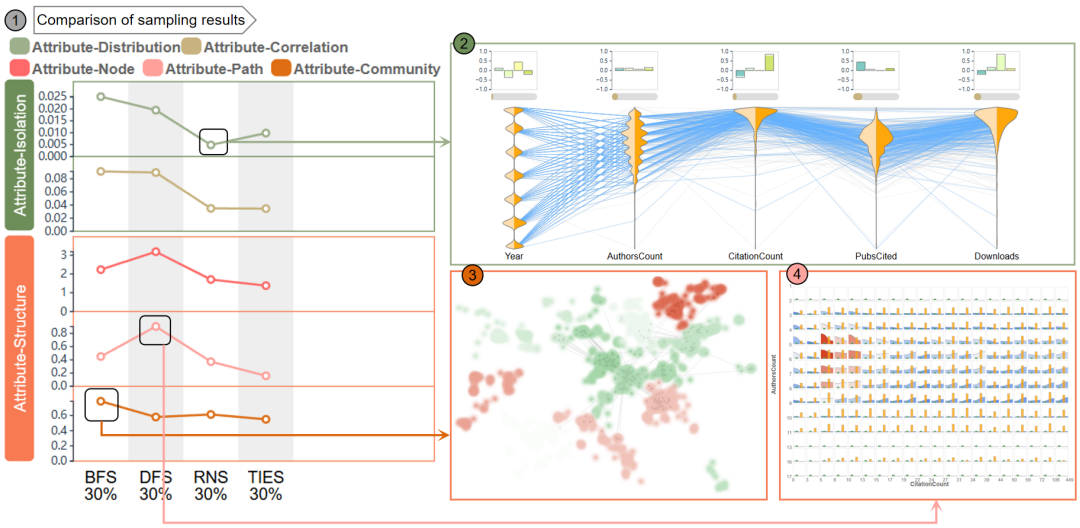

Figure 5: (1) Shows the metric values of differences for each category under the four sampling methods; (2)(3)(4) Display the details of the differences.

Table 1: The difference values across sampling strategy, with bold black text indicating the best-performing strategy and the underline text highlighting the worst-performing strategy for each difference category.

Category	Strategies / Metrics	BFS	DFS	FF	SBS	RW	RNS	RES	TIES
Attribute-Isolation	Attribute-Distribution	0.0197	0.0093	0.0066	0.0055	0.0079	**0.0017**	0.0045	0.0031
	Attribute-Correlation	0.0261	0.0303	0.0215	0.0263	**0.0193**	0.0229	0.0321	0.0278
Attribute-Structure	Attribute-Node	1.4403	0.9569	0.9090	0.6309	0.3567	0.3868	0.3882	**0.3415**
	Attribute-Path	0.2371	0.8266	0.1451	0.0951	0.0427	0.1857	0.0984	**0.0084**
	Attribute-Community	0.0562	0.0469	0.0432	0.0349	0.0342	0.0421	0.0431	**0.0301**

others, thereby disrupting distributional equilibrium. Regarding Attribute-Correlation, DFS caused the most significant alterations due to its depth-first traversal pattern, which follows extended paths and consequently distorts correlation structures. Conversely, RW sampling preserved original attribute correlations most effectively through its uniform graph coverage. In Attribute-Node, BFS also performed poorly, while TIES excelled by best preserving associations between attributes and node topological metrics. This advantage stems from TIES's edge-induced sampling mechanism, which maintains complete node sets and thus better conserves Attribute-Structure relationships. For Attribute-Path metrics, DFS produced the most severe distortions, whereas TIES most effectively maintained original relationships between attribute and path length. Finally, in Attribute-Community evaluations, BFS most dramatically affected community structures, disproportionately preserving some communities while severely losing others.

Additional observations revealed that FF and SBS, as variants of BFS and DFS respectively, exhibited similar sampling characteristics to their parent algorithms while demonstrating modest overall improvements. RES showed intermediate performance across all difference metrics, without particularly notable strengths or weaknesses. The above experimental results have been recognized by experts and are consistent with our prior knowledge, further demonstrating the effectiveness of our work.

6.3 Discussion

Attributes and topology jointly define graph information, and integrating both enhances sampling evaluation. We developed a visual analytics system to explore attribute-related differences from sampling, with case studies and quantitative analyses demonstrating its effectiveness in tracking differences and comparing strategies. However, several challenges remain for future work: (1) Scalability: The Attribute-Isolation view which combines parallel coordinates with area and bar charts, effectively conveys attribute distribution and correlation differences. However, as the number of attributes increases, visual clutter may arise. Furthermore, although the system performs well on small and medium-sized graphs, handling the complexity of large-scale and multi-attribute graphs remains a challenge, which may be addressed through grouped attribute visualization and progressive rendering techniques. (2) Weakness of optimization support: Although the system helps identify differences and analyze sampling behavior, it does not guide algorithm refinement. Leveraging difference patterns to optimize sampling strategies is an important direction. (3) Attribute completeness: Our current evaluation focuses on node attributes. To achieve more comprehensive sampling assessment, future work should incorporate edge attributes, which often carry important semantic information. This requires designing metrics and visual encodings that capture differences on both nodes and edges.

7 Conclusion

In this paper, we present an attribute-aware framework for evaluating differences in graph sampling by integrating node attributes with topological structures. Based on this, we propose a taxonomy comprising Attribute-Isolation and Attribute-Structure correlation, further detailed into five difference types. For each, we design customized quantitative metrics and a multiview visual analytics system combining heatmaps, parallel coordinate plot, and topology-based views. Case studies and quantitative results confirm the system's effectiveness in revealing sampling-induced attribute changes and supporting informed analysis and method selection.

REFERENCES

[1] H. Ahmed, T. Howton, Y. Sun, N. Weinberger, Y. Belkhadir, and M. S. Mukhtar. Network biology discovers pathogen contact points in host protein-protein interactomes. *Nature communications*, 9(1):2312, 2018. doi: 10.1038/s41467-018-04632-8 3

[2] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):1–56, 2013. doi: 10.1145/2601438 2

[3] A. Badalyan, N. Ruggeri, and C. De Bacco. Structure and inference in hypergraphs with node attributes. *Nature Communications*, 15(1):7073, 2024. doi: 10.1038/s41467-024-51388-5 2

[4] A.-L. Barabási. The new science of networks. *Cambridge MA. Perseus*, 2002. doi: 10.1119/1.1538577 2

[5] M. Bloznelis. Degree and clustering coefficient in sparse random intersection graphs. 2013. doi: 10.1214/12-AAP874 2

[6] P. Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007. doi: 10.1016/j.socnet.2007.04.002 2, 4

[7] H. Cai. A note on jointly modeling edges and node attributes of a network. *Statistics & Probability Letters*, 121:54–60, 2017. doi: 10.1016/j.spl.2016.10.014 3

[8] D. Cheng, Y. Zou, S. Xiang, and C. Jiang. Graph neural networks for financial fraud detection: a review. *Frontiers of Computer Science*, 19(9):1–15, 2025. doi: 10.1007/s11704-024-40474-y 1

[9] F. Chiericetti, A. Dasgupta, R. Kumar, S. Lattanzi, and T. Sarlós. On sampling nodes in a network. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 471–481, 2016. doi: 10.1145/2872427.2883045 2

[10] S. Citraro, V. Pansanella, and G. Rossetti. Structure-attribute similarity interplay in diffusion dynamics on social networks. In *International Conference on Discovery Science*, pp. 425–439. Springer, 2024. doi: 10.1007/978-3-031-78980-9_27 1, 3

[11] C. Doerr and N. Blenn. Metric convergence in social network sampling. In *Proceedings of the 5th ACM workshop on HotPlanet*, pp. 45–50, 2013. doi: 10.1145/2491159.2491168 2

[12] S. Dolev, Y. Elovici, and R. Puzis. Routing betweenness centrality. *Journal of the ACM (JACM)*, 57(4):1–27, 2010. doi: 10.1145/1734213.1734219 2

[13] E. Estrada and J. A. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(5):056103, 2005. doi: 10.1103/PhysRevE.71.056103 2

[14] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pp. 417–426, 2019. doi: 10.1145/3308558.3313488 3

[15] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978. doi: 10.1016/0378-8733(78)90021-7 2

[16] M. Gentner, I. Heinrich, S. Jäger, and D. Rautenbach. Large values of the clustering coefficient. *Discrete Mathematics*, 341(1):119–125, 2018. doi: 10.1016/j.disc.2017.08.020 2

[17] M. Ghavipour and M. R. Meybodi. Irregular cellular learning automata-based algorithm for sampling social networks. *Engineering Applications of Artificial Intelligence*, 59:244–259, 2017. doi: 10.1016/j.engappai.2017.01.004 1, 2

[18] L. A. Goodman. Snowball sampling. *The annals of mathematical statistics*, pp. 148–170, 1961. doi: 10.1214/aoms/1177705148 2

[19] S. J. Hardiman and L. Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 539–550, 2013. doi: 10.1145/2488388.2488436 2

[20] P. Hu and W. C. Lau. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*, 2013. doi: 10.48550/arXiv.1308.5865 1, 2

[21] M. Jaouadi and L. B. Romdhane. A graph sampling-based model for influence maximization in large-scale social networks. *IEEE Transactions on Computational Social Systems*, 11(1):144–160, 2022. doi: 10.1109/TCSS.2022.3216587 2

[22] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.-H. Cui, and A. G. Percus. Reducing large internet topologies for faster simulations. In *International Conference on Research in Networking*, pp. 328–341. Springer, 2005. doi: 10.1007/11422778_27 2

[23] S. Kumar and H. Sundaram. Task-driven sampling of attributed networks. *arXiv preprint arXiv:1611.00910*, 2016. doi: 10.48550/arXiv.1611.00910 2, 3

[24] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of bfs (breadth first search). In *2010 22nd International Teletraffic Congress (LTC 22)*, pp. 1–8. IEEE, 2010. doi: 10.1109/ITC.2010.5608727 2

[25] M. Kurant, A. Markopoulou, and P. Thiran. Towards unbiased bfs sampling. *IEEE Journal on Selected Areas in Communications*, 29(9):1799–1809, 2011. doi: 10.1109/JSAC.2011.111005 2

[26] C.-H. Lee, X. Xu, and D. Y. Eun. Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. *ACM SIGMETRICS Performance evaluation review*, 40(1):319–330, 2012. doi: 10.1145/2318857.2254795 2

[27] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631–636, 2006. doi: 10.1145/1150402.1150479 2

[28] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, 2005. doi: 10.1145/1081870.1081893 2

[29] R.-H. Li, J. X. Yu, L. Qin, R. Mao, and T. Jin. On random walk based graph sampling. In *2015 IEEE 31st international conference on data engineering*, pp. 927–938. IEEE, 2015. doi: 10.1109/ICDE.2015.7113345 2

[30] M. Lin, W. Li, and S. Lu. Balanced influence maximization in attributed social network based on sampling. In *Proceedings of the 13th international conference on web search and data mining*, pp. 375–383, 2020. doi: 10.1145/3336191.3371833 2, 3

[31] Y. Luo, Y. Huang, G. Luo, K. Qin, and A. Chen. Edge convolutional networks: Decomposing graph convolutional networks for stochastic training with independent edges. *Neurocomputing*, 549:126430, 2023. doi: 10.1016/j.neucom.2023.126430 2

[32] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 105–113, 2011. doi: 10.1145/2020408.2020431 2

[33] B. Martinez-Seis, X. Li, and X. Wang. Measure community quality by attribute importance and density in social networks. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pp. 628–633. IEEE, 2019. doi: 10.1109/COASE.2019.8842970 3

[34] F. Meng, X. Rui, Z. Wang, Y. Xing, and L. Cao. Coupled node similarity learning for community detection in attributed networks. *Entropy*, 20(6):471, 2018. doi: 10.3390/e20060471 3

[35] Q. H. Nguyen, S.-H. Hong, P. Eades, and A. Meidiana. Proxy graph: Visual quality metrics of big graph sampling. *IEEE transactions on visualization and computer graphics*, 23(6):1600–1611, 2017. doi: 10.1109/TVCG.2017.2674999 2

[36] D. Rafiei. Effectively visualizing large networks through sampling. In *VIS 05. IEEE Visualization, 2005.*, pp. 375–382. IEEE, 2005. doi: 10.1109/VISUAL.2005.1532819 2

[37] M. Seufert, S. Lange, and T. Hoßfeld. More than topology: Joint topology and attribute sampling and generation of social network graphs. *Computer Communications*, 73:176–187, 2016. doi: 10.1016/j.comcom.2015.07.023 2, 3

[38] Y. Shang, Z. Wang, Y. Chen, X. Yang, Z. Ren, X. Zeng, and L. Xu. Hnf-dda: subgraph contrastive-driven transformer-style heterogeneous network embedding for drug–disease association prediction. *BMC biology*, 23(1):1–16, 2025. doi: 10.1186/s12915-025-02206-x 2

[39] Z. Su, Y. Liu, J. Kurths, and H. Meyerhenke. Generic network sparsification via degree-and subgraph-based edge sampling. *Information Sciences*, 679:121096, 2024. doi: 10.1016/j.ins.2024.121096 2

[40] Q. Tan, X. Zhang, X. Huang, H. Chen, J. Li, and X. Hu. Collaborative graph neural networks for attributed network embedding. *IEEE Trans-*

actions on Knowledge and Data Engineering, 36(3):972–986, 2023. doi: 10.1109/TKDE.2023.3298002 1

[41] C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier. Sampling from social networks with attributes. In *Proceedings of the 26th international conference on world wide web*, pp. 1181–1190, 2017. doi: 10.1145/3038912.3052665 2

[42] T. Wang, Y. Chen, Z. Zhang, T. Xu, L. Jin, P. Hui, B. Deng, and X. Li. Understanding graph sampling algorithms for social network analysis. In *2011 31st international conference on distributed computing systems workshops*, pp. 123–128. IEEE, 2011. doi: 10.1109/ICDCSW.2011.34 2

[43] D. R. White and F. Harary. The cohesiveness of blocks in social networks: Node connectivity and conditional density. *Sociological Methodology*, 31(1):305–359, 2001. doi: 10.1111/0081-1750.00098 2, 5

[44] Y. Wu, N. Cao, D. Archambault, Q. Shen, H. Qu, and W. Cui. Evaluation of graph sampling: A visualization perspective. *IEEE transactions on visualization and computer graphics*, 23(1):401–410, 2016. doi: 10.1109/TVCG.2016.2598867 2

[45] S.-H. Yoon, K.-N. Kim, J. Hong, S.-W. Kim, and S. Park. A community-based sampling method using dpl for online social networks. *Information Sciences*, 306:53–69, 2015. doi: 10.1016/j.ins.2015.02.014 2

[46] M. I. Yousuf, I. Anwer, and R. Anwar. Empirical characterization of graph sampling algorithms. *Social Network Analysis and Mining*, 13(1):66, 2023. doi: 10.1007/s13278-023-01060-5 2, 4

[47] M. I. Yousuf and S. Kim. Guided sampling for large graphs. *Data mining and knowledge discovery*, 34(4):905–948, 2020. doi: 10.1007/s10618-020-00683-y 1

[48] F. Zhang, S. Zhang, P. C. Wong, H. Medal, L. Bian, J. E. Swan II, and T. Jankun-Kelly. A visual evaluation study of graph sampling techniques. *Electronic Imaging*, 29:110–117, 2017. doi: 10.2352/ISSN.2470-1173.2017.1.VDA-394 1, 2

[49] J. Zhao, P. Wang, J. C. Lui, D. Towsley, and X. Guan. Sampling online social networks by random walk with indirect jumps. *Data Mining and Knowledge Discovery*, 33:24–57, 2019. doi: 10.1007/s10618-018-0587-5 2

[50] Z. Zhou, C. Shi, X. Shen, L. Cai, H. Wang, Y. Liu, Y. Zhao, and W. Chen. Context-aware sampling of large networks via graph representation learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1709–1719, 2020. doi: 10.1109/TVCG.2020.3030440 2

[51] M. Zitnik, M. Agrawal, and J. Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018. doi: 10.1093/bioinformatics/bty294 1, 3