Context-aware Sampling of Large Networks via Graph Representation Learning

Zhiguang Zhou, Chen Shi, Xilong Shen, Lihong Cai, Haoxuan Wang,
Yuhua Liu, Ying Zhao and Wei Chen

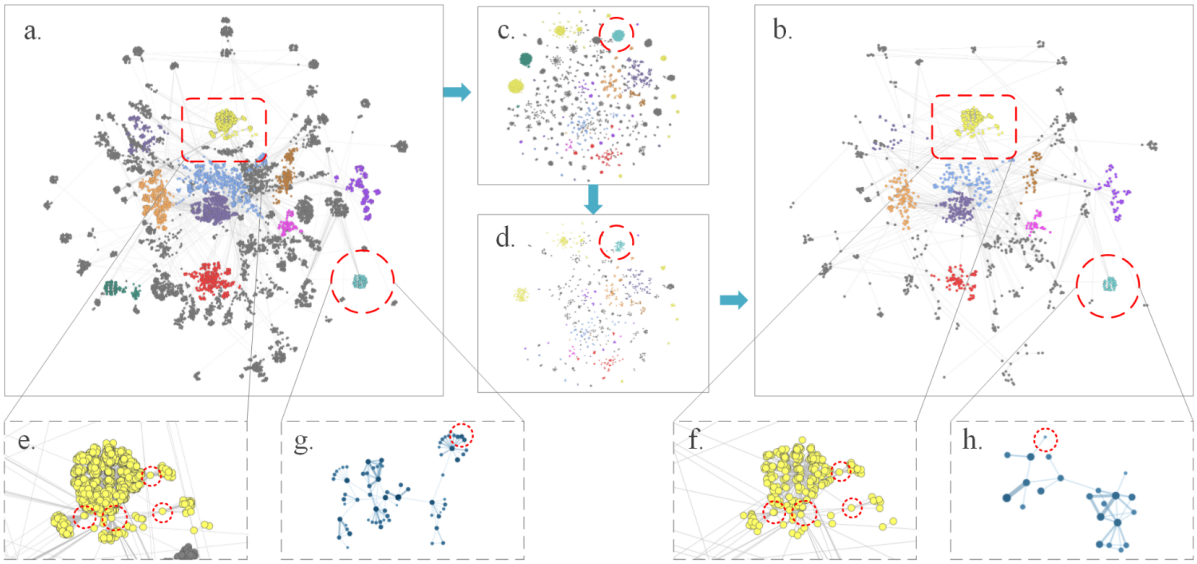


Fig. 1. A case for a *Webbase* data (16k nodes, 26k edges) based on our context-aware sampling method. (a) presents the original graph with a node-link diagram. (c) presents scatterplots obtained through GRL (node2vec) and dimensionality reduction (t-SNE). (e) highlights a local structure of interest in (a), and the circled nodes are of significance. (g) presents an aggregated layout of (a), in which each supernode represents a community feature. Our sampling method is conducted on (c), and the sampled scatterplots are presented in (d) with a contextual structure of interest highlighted by a red circle. (b) presents the corresponding sampled graph, with the significant features retained such as bridging nodes highlighted in (f) and graph connections presented in (h).

Abstract—Numerous sampling strategies have been proposed to simplify large-scale networks for highly readable visualizations. It is of great challenge to preserve contextual structures formed by nodes and edges with tight relationships in a sampled graph, because they are easily overlooked during the process of sampling due to their irregular distribution and immunity to scale. In this paper, a new graph sampling method is proposed oriented to the preservation of contextual structures. We first utilize a graph representation learning (GRL) model to transform nodes into vectors so that the contextual structures in a network can be effectively extracted and organized. Then, we propose a multi-objective blue noise sampling model to select a subset of nodes in the vectorized space to preserve contextual structures with the retention of relative data and cluster densities in addition to those features of significance, such as bridging nodes and graph connections. We also design a set of visual interfaces enabling users to interactively conduct context-aware sampling, visually compare results with various sampling strategies, and deeply explore large networks. Case studies and quantitative comparisons based on real-world datasets have demonstrated the effectiveness of our method in the abstraction and exploration of large networks.

Index Terms—Graph sampling, Graph representation learning, Blue noise sampling, Graph evaluation

1 Introduction

- *Zhiguang Zhou, Chen Shi, Xilong Shen, Lihong Cai, Haoxuan Wang and Yuhua Liu are with School of Information, Zhejiang University of Finance and Economics. E-mail: {zhgzhou1983, sc73048, 180110910420, cailihong, wanghaoxuan, liuyuhua}@zufe.edu.cn.*
- *Ying Zhao is with Central South University. E-mail: zhaoying@csu.edu.cn.*
- *Wei Chen is with State Key Lab of CAD & CG, Zhejiang University. E-mail: chenwei@cad.zju.edu.cn.*
- *Ying Zhao and Wei Chen are corresponding authors.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

As a ubiquitous data structure, network is always employed to encode relationships among entities in a variety of application areas, such as social relationships between people and financial transactions between companies [5, 57]. Graph visualization offers an interactive and exploratory means allowing users to gain structural insights [2] and sense implicit contextual features of networks. However, with the increase of data sizes, the visual exploration and analysis of networks are seriously influenced, because nodes and edges overlap with each other and generate much visual clutter in large graph visualizations, making it a complicated and time-consuming task to visually explore structural features of significance [50].

Graph sampling is commonly used to reduce the visual clutter and address scalability issues in the visual exploration of large networks, by means of which a subset of nodes and edges are selected on behalf of the original large graph. Over the past few decades, numerous ef-

forts have been paid on the design of sampling strategies, ranging from node-based and edge-based schemes [4, 26] to transversal-based and semantic-based schemes [4, 23, 56]. However, such strategies largely focus on sampling efficiency and randomness of sampling results, paying little attention to the preservation of significant contextual structures.

Contextual structures, formed by nodes and edges with tight relationships, are always of great significance for the exploration and interpretation of networks, such as bridging nodes, connected paths and aggregated communities [19,46]. For example, it is quite necessary to identify the contextual structures of crowd movement network for the diagnosis and spread prevention of infectious diseases [44]. However, it is a tough task to preserve contextual structures in the sampled network based on traditional sampling strategies, because contextual structures often have three characteristics: concealment in location, irregularity in scale, and complexity in structure. For example, nodes with tight relationships (in a community) may be difficult to find in large networks due to their concealed locations, because they are easily laid out far away from each other. Also, contextual structures are immune to scale, that is few nodes and edges would rather present a tough contextual structure (a small complete graph). Thus, it is really hard to give a comprehensive definition of contextual structures because their formations are too complicated to find a regular pattern.

As an effective way to represent and identify contextual structures of large networks [5], GRL has been widely applied in a variety of research areas, such as graph classification, graph query, graph mining, et al [12, 33]. It transforms nodes into vectors to quantitate the structural features of networks. Numerous GRL models have been proposed to train and represent nodes according to their local contexts in the network, such as deepwalk [39], node2vec [11], and struc2vec [42]. A family of biased random walks are developed in the course of corpus generation allowing an efficient exploration of diverse neighborhoods for given nodes [32]. Thus, network structures are well represented in a vectroized space obtained by GRL (e.g. a contextual structure of interest is highlighted as shown in Figure 1a and Figure 1c). We believe that it would be a feasible way to conduct graph sampling in the vectorized space, and the contextual structures would be preserved as far as possible (e.g. the contextual structure is well preserved in the sampled graph as shown in Figure 1d and Figure 1b).

However, there are still several problems to overcome for the preservation of contextual structures in the vectorized space obtained through GRL. **P1**: GRL is able to encode the contextual structures with vectorizied representation, but the vectorized space is too complicated to gain insights due to its high dimensions. **P2**: It is a difficult task to define a graph sampling model to preserve contextual structures captured by GRL, since they are represented with data distributions in the vectorized space rather than topological relationships in the original network space. **P3**: It is also difficult to conduct a unified graph sampling scheme to preserve various kinds of contextual structures in the vectorized space due to their respective characteristics. **P4**: It is another tough task to evaluate the sampled graphs from a variety of perspectives, and further demonstrate that the contextual structures of significance are well retained in the sampled graphs.

In this paper, we propose a novel graph sampling method to simplify large graphs, especially with the contextual structures identified and preserved in the sampled graphs. Firstly, a GRL model is employed to encode contextual structures and a dimensionality reduction method is applied to transform the contextual structures into a low-dimensional vectorized space, where nodes sharing similar contextual features are visually distributed close to each other (**P1**). Then, we propose a novel blue noise sampling model to generate a subset of nodes in the vectorized space, guaranteeing that nodes with tight relationships are retained and the contextual features are well preserved in the sampled graph (**P2**). A set of desired objectives are further integrated into the sampling model to optimize the sampled graphs, in which topological features of significance are enhanced such as bridging nodes and graph connection (**P3**). Also, we utilize a group of metrics to evaluate the validity of our sampling method in contextual feature preservation from different perspectives, such as node importance, graph connection and community changes (**P4**). At last, a graph sampling framework is implemented to integrate sampling models, GRL and visual designs of metrics, and a rich set of interactions are also provided allowing users to intuitively evaluate different sampling strategies and easily explore structures of interest in large networks. The effectiveness and usefulness of our system are further demonstrated with case studies and quantitate comparisons based on real-world datasets. In summary, the main contributions of our work are:

- We utilize a GRL model (node2vec) to quantitate the contextual features of networks, offering important clues for graph sampling. To the best of our knowledge, it is the first to sample graphs with GRL.
- We design a multi-objective blue noise sampling model to simplify large networks, with the contextual structures and their topological features well retained in the sampled graphs.
- We propose a group of specific metrics enabling users to compare sampling strategies from different perspectives, and conduct case studies with real-world datasets to demonstrate the validity of our context-aware graph sampling method.

2 Related Work

We classify existing methods into four categories, including large graph visualization, graph sampling, metrics for graph evaluation and graph representation learning.

2.1 Large Graph Visualization

Graph visualization is widely used for network analysis [5]. Node-link diagram is a most intuitive layout scheme in which the nodes are represented as points and edges are represented as lines. Force-directed methods are employed to layout the node-link diagrams by optimizing graph drawing aesthetics [10, 19, 41]. With the increasing size of networks, the readability of node-link diagrams largely decreases due to visual clutter and scalability issues [9]. Two categories of methods are proposed: **(1) Graph clustering methods** aggregate groups of nodes and edges with similar properties to reduce the visual complexity of large graphs [7, 58]. ASK-GraphView [1] was proposed to organize a large graph into hierarchical structures allowing users to aggregate nodes to reduce visual clutter. To reduce edge crossings and emphasize directional patterns, a set of edge-based clustering methods are proposed in which edges with similar spatial distribution features are bundled together [6, 8, 15, 16]. **(2) Graph filtering methods** extract subgraphs of interest from original large graphs [20]. Hennessey et al. [14] took a set of graph metrics into account to obtain representative skeletons and simplify the visualization of large graphs, such as shortest path and distance to the central node. Yoghourdjian et al. [51] proposed Graph Thumbnails to enhance the readability of large graphs with high-level structures described with small icon-like glyphs. It can be seen that the underlying topological structures of networks are changed with clustering methods, which are still not preserved in the simplified graphs with filtering methods. It might generate a great deal of ambiguity that misleads the exploration of networks [49].

2.2 Graph Sampling

Graph sampling is another kind of filtering method, which also generates a subset of nodes or edges to simplify the original networks. Three categories are covered: **(1) Node-based Sampling.** Random Node (RN) sampling [26] is commonly used to randomly generate nodes from the original network. A set of graph properties are considered to improve the results of node-based sampling. For example, Random PageRank Node (RPN) defines the probability of nodes to be sampled as proportional to their PageRank weights [26, 36]. Random Degree Node (RDN) increases the probability of nodes with higher degree values to be sampled [4]. Hu et al. [18] designed a graph sampling method based on spectral sparsification, to reduce the number of edges and retain structural properties of original graphs. **(2) Edge-based Sampling.** Random Edge (RE) extracts a random subset of

edges from the original graph [52]. Random Node-Edge (RNE) [26] randomly selects a node followed with an adjacent edge. Random Edge-Node (REN) [13] selects nodes based on the randomly selected edges, and then generates a subgraph by adding edges whose nodes are present. Due to the sparsely connection of sampled graphs, RE and RNE usually miss structures of interest in original graphs [26]. **(3) Traversal-based Sampling.** Traversal-based sampling methods are proposed to preserve the connection relationships of networks, such as Depth First (DF, sampling nodes in a depth first order) [30] and Breadth First (BF, sampling nodes in a breadth first order) [23]. In the course of traversal-based sampling, nodes with higher degree and page rank values are easily retained in the sampled graphs. Similar to BF, Snow-Ball (SB) samples nodes in a fixed fraction of neighbors [50]. Forest Fire (FF) [27] samples a seed node with its adjacent edges and nodes recursively selected within a probability. Simple Random Walk (SRW) iteratively selects adjacent nodes at random, which easily falls into local traps [45]. Random Jump (RJ) will jump out of traps by randomly selecting other nodes within a certain probability in each iteration step [50].

2.3 Metrics for Graph Evaluation

To evaluate the validity of sampled graphs, a large number of metrics have been proposed that can be classified into two categories: **(1)** Numerous of quantitative metrics are proposed based on topological structures of networks, such as node degree, cluster coefficient, connectivity [17] and so on. Leskovec and Faloutsos [26] provided a set of criteria for a static snapshot of the graph, including in/out-degree distribution, weakly/strongly connected component, clustering coefficient, et al. Maiya et al. [30] introduced two degree-based measures (degree distribution similarity and hub inclusion) and two clustering-based measures (local/global clustering coefficient). In addition, some metrics combined with network analysis tasks are proposed. For example, Zhang et al. [54] presented precision and recall to effectively reflect the match quality of the clusters after graph sampling. The degree-sensitive neighborhood graph (dNNG) was introduced to design a shape-based metrics for neighborhoods in networks [34]. van Heeswijk et al. [48] designed metrics about the bisimulation structures of graphs. **(2)** Recently, a set of metrics have been proposed to evaluate graph sampling from a visualization perspective [50, 55]. Wu et al. [50] provided three important visual factors (high degree nodes, cluster quality and coverage area) to evaluate the influence on the perception of node-link diagrams caused by graph sampling strategies. Nguyen et al. [35] designed a family of shape-based quality metrics based on the new concept of 'proxy graph'. To select a well layout of graphs, Kwon et al. [24] presented the visual inspection of graphs in different layouts and provided corresponding aesthetic metrics.

2.4 Graph Representation Learning

Graph representation learning methods are widely utilized in the field of network analysis, which is able to mathematically depict the large network in a feature space, enabling users to easily capture network features [12, 53]. It can be classified into two categories [3]: **(1) Feature-based methods** formulate the vectorized representation of nodes by measuring a set of features. For example, van den Elzen et al. [47] flattened the adjacency matrix as a vector and attached derived attributes at the end of vectors. Pienta et al. [40] generated a vectorized signature with the combination of node attributes and topological structures. The graphlet kernel method obtained the vectorized features by counting the frequencies of graphlets, which are small, induced, and non-isomorphic subgraph patterns [24]. Narayanan et al. [33] further used graphlet kernels to measure the similarities between large graphs. **(2) Learning-based methods** have been proposed to transform network structures into a vectorized space by taking advantages of representation learning models, such as deepwalk [39], node2vec [11], struc2vec [42] and graphwave. In addition, key factors of learning process can be changed to optimize the learned features [28, 29]. For example, Kennedy et al. [21] presented a concept of graph landscape and used a combination of low-dimensional feature representation embedding, metric charts, mutual feature information representation for the analysis of graph corpus. Pei et al. [38] designed struc2gauss to learn node representations in the space of Gaussian distributions and performed network embedding based on global structural information.

3 Requirement Analysis and System Overview

In this section, a list of analytical requirements are summarized after detailed discussions with domain experts. The pipeline of our context-aware sampling framework is then presented to complete the large network exploration tasks.

3.1 Requirement Analysis

Our research focus is obtained through discussions with two domain experts (E1 and E2) in the fields of graph mining and visual analysis regarding their needs in exploring large networks. E1 has been working in a professional research institution for 7 years. During this period, E1 devoted himself to the researches about data mining and network analysis, especially for exploration of urban traffic networks. Thus, he has rich experience in graph visualization and mining. E2 is a senior data analyst working in an internet company for 5 years. He focuses on the data mining and exploratory analysis of social networks, and has strong demands for visualization and visual analytics of large graphs. We had meetings with two experts biweekly from October 2019 to March 2020. We began by interviewing the experts about their ongoing projects about large network exploration, such as graph classification, graph query, graph mining, and asked for any major issues they are troubled in the process. It turned out that graph sampling was their frequently-used method to give an overview of large networks, while the potential loss of significant structures in the sampled results largely limited its usability for further analysis. Although a variety of sampling strategies have been proposed to accelerate the calculation and simplify the visualization of large networks, it is still a difficult task to simultaneously retain the contextual structures of significance, such as bridging nodes, graph connection and communities. For example, random sampling strategies preserve overall structures in the sampled graph while local graph connections are always broken. Traversal-based sampling strategies largely preserve graph connections, which are easily troubled with local trapping. Specifically, the experts also claimed that it was a necessary task to provide visual cues for users to directly evaluate the features retained with different sampling strategies, and select a suitable sampling strategy for graph simplification and exploration. To sum up, we present four requirements as follows:

R1. Contextual structure representation. Contextual features are formed by nodes and edges with tight relationships, which are always difficult to capture through visual perception, especially when the networks are large and complex. In traditional graph sampling strategies, contextual structures cannot be well preserved at the same time. Therefore, a feature space where contextual structures of networks can be better represented is demanded by the experts, which will do great favors for subsequent graph sampling, evaluation and exploration.

R2. Context-aware sampling. In traditional sampling strategies, a subset of nodes or edges are selected based on different rules, which often focus on the efficiency of sampling, but ignore the preservation of contextual structures. Compared with an overview of the large network, the experts demand a more meaningful sampled graph, in which significant contextual features are preserved, enabling users to conduct further graph calculation and exploration. Therefore, it is necessary to design a unified context-aware sampling model for preservation of contextual structures with a variety of objectives integrated.

R3. Sampled graph evaluation. A lot of strategies have been proposed to sample graphs, and the performance is always quantified through metrics based on graph structural properties such as degree distribution and clustering coefficient. However, it is still a difficult issue to evaluate the effectiveness of different sampling strategies, since the contextual structures are comprised of features from different perspectives which cannot be well measured with general metrics. Two requirements were identified by the experts according to their experience. First, a comprehensive group of specific metrics are required to evaluate the effectiveness in the preservation of contextual structures.

Second, case studies based on real-world datasets are demanded to further demonstrate the validity of our context-aware graph sampling.

R4. Graph sampling framework. A lot of sampling strategies can be employed to simplify and explore large networks. However, it is really difficult to trade off different sampling strategies in the course of network exploration to achieve a desired simplification. According to the requirements of experts, a sampling framework including different sampling strategies is demanded by means of which the validity of sampled results can be perceived and evaluated in real time, enabling users to sample large networks with suitable sampling strategies and reasonable sampling rates.

3.2 System Overview

Motivated by the identified requirements, we designed a visualization framework enabling users to conduct context-aware sampling for the exploration of large networks. The system pipeline is presented in Figure 2. Firstly, large networks are loaded into the visualization system, and two categories of features are calculated in advance, including contextual structures represented by a GRL model and topological properties obtained with traditional statistical methods (R1). Then, a novel blue noise sampling strategy is conducted to generate a random sample in the vectorized space, with the contextual structures preserved as far as possible in the sampled graph. Several topological properties are further integrated into the sampling model to enhance the contextual structures, including node betweenness and graph connection (R2). Furthermore, a group of metrics are utilized to evaluate effectiveness of the proposed sampling strategy from different perspectives, such as node importance, node connection and community changes (R3). Also, a set of visual designs and interfaces are provided in the visualization framework, enabling users to compare the metrics intuitively and explore the networks interactively (R4).

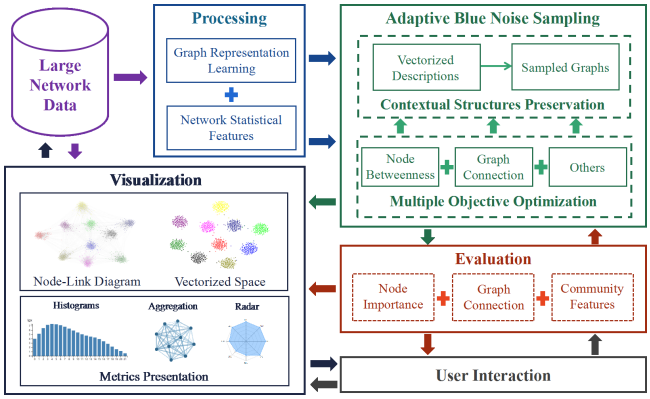


Fig. 2. The pipeline of our system which is comprised of sampling model, visual abstraction and quantitative evaluation.

4 Context-aware Graph Sampling

In this section, we detail the courses about transforming networks into vectorized space with context representation models and selecting those points with multiple contextual structures as far as possible.

4.1 Context Representation

It is a tough task to preserve the contextual structures in sampled graphs with traditional sampling strategies [11]. In recent years, GRL is widely applied to learn vectors to encode structural information of large networks. In this paper, we utilize node2vec to embed contextual structures of original large networks into quantitative vectors (R1).

Corpus Generation: Firstly, a random walk of fixed length l is simulated from a given source node u. w_i denotes the ith node in the walk, starting with $w_i = u$, which is generated by the following distribution:

$$P(w_i = x | w_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & if(v,x) \in E \\ 0 & otherwise \end{cases} \quad (1)$$

where π_{vx} is the unnormalized transition probability between nodes v and x, and Z is the normalizing constant [39].

Then, the transition probabilities π_{vx} on edges $(v;x)$ leading from v is defined as $\alpha_{pq}(t;x)$ in a unweighted graph, according to the walk that traverses edge $(t;v)$ and resides at node v. d_{tx} denotes the distance of shortest path between node t and node x. p controls the likelihood of immediately revisiting a node in the walk, while q allows the search to differentiate the approximate BFS behavior $(q > 1)$ and DFS-like exploration $(q < 1)$. Specifically, there are three situations: (1) if $d_{tx} = 0$, $\alpha_{pq}(t;x) = 1/p$; (2) if $d_{tx} = 1$, $\alpha_{pq}(t;x) = 1$; (3) if $d_{tx} = 2$, $\alpha_{pq}(t;x) = 1/q$. Based on the operations [11], enough sequences are collected to generate a corpus.

Vectorized Representation: In the model of node2vec, a skip-gram architecture is utilized to train the distributed representation of network nodes. The objective is to maximize the log-probability of the following objective function, which represents the likelihood of preserving network neighborhoods of nodes:

$$\max_f \sum_{u \in V} logPr(N_s(u)|f(u)) \quad (2)$$

where $Ns(u)$ is generated through the 2^{nd} order random walk [31, 32]. Two standard assumptions are used to tackle the optimization problem. In the first assumption, the likelihood is factorized based on conditional independence. In the second assumption, the probabilities are calculated through a softmax function, based on the symmetry in the feature space. With the above assumptions, the objective is simplified as:

$$\max_f \sum_{u \in V} \left[-logZ_u + \sum_{n_i \in N_s(u)} f(n_i) \cdot f(u) \right] \quad (3)$$

Since it is time-consuming to resolve the per-node partition function Z_u, negative sampling is employed to achieve approximation [32]. Therefore, vectorized representations of nodes are obtained with multiple contextual structures embedded through node2vec.

Dimension Reduction: With GRL, nodes are represented as vectors in a high-dimensional space (more than 50 dimensions), where nodes with similar contextual structures are close to each other. However, the relationships of high-dimensional vectors are difficult to perceive through visual perception.t-Distributed Stochastic Neighbor Embedding (t-SNE) [25] is an effective dimensionality reduction method, which is capable of enhancing local features while retaining global features in the dimensionality reduction space [46]. Thus, we utilize t-SNE to project high-dimensional vectors into a two-dimensional space, in which contextual structures of interest are visually enhanced. Figure 1a presents the node-link diagram of a network, in which each community is shaded in a different color. Corresponding nodes with the same color (such as yellow) are distributed close to each other as shown in Figure 1c, indicatiing that contextual structures are well embedded in the vectorized space obtained with GRL(node2vec).

4.2 Context-aware Sampling

Given the vectorized space obtained by context representation, we design a novel multi-objective blue noise sampling model integrating multiple objectives such as node importance and graph connection to preserve contextual structures of interest in the sampled graph.

4.2.1 Adaptive Blue Noise Sampling

As shown in Figure 1c, contextual structures are well preserved in the vectorized space where network nodes with tight relationships are distributed close to each other. To preserve the contextual structures distributed all over the original graph, we conduct an adaptive blue noise sampling [37] in the vectorized space to generate a subset of nodes (as shown in Figure 1d), such that all contextual structures will be retained in the sampled graphs, no matter where they are distributed or how much their sizes are (R2). Figure 3a presents the illustration of adaptive blue noise sampling, and three steps are detailed as follows:

Step.1 A root node is randomly selected and a Poisson disk is generated centered with it. For a local area determined with the Poisson

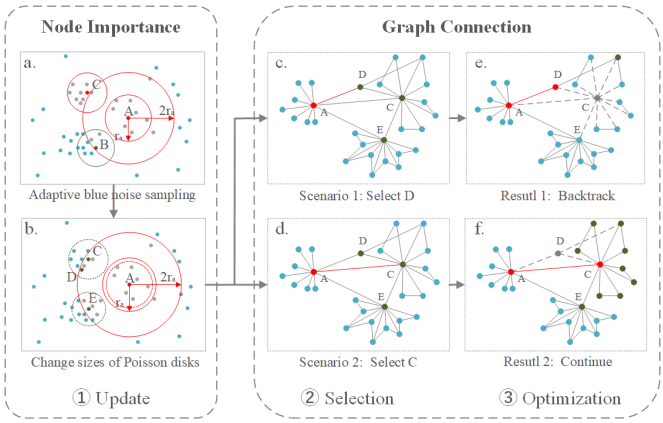


Fig. 3. Illustration of our context-aware sampling method. (a) illustrates the adaptive blue noise sampling method. (b) presents the size changes of Poisson disks according to node importance and spatial density. (c)-(f) illustrates the selection and optimization of sampled nodes for the retention of graph connection.

disk, a node is randomly selected as a representative. The radius of the Poisson disk is adaptively determined according to its local density. Equation 1 presents the calculation of radius:

$$R = r_a / f(p) \tag{4}$$

where $f(p)$ is the density estimation [56] of point p. r_a is a user-defined parameter to specify the sampling rate by adjusting radius.

Step.2 When a representative node (shaded in red) is specified, those nodes in the same Poisson disk are labeled as inactive nodes (shaded in grey), while those nodes located in an annular area between R and 2*R are labeled as active (shaded in blue). Then, an active node will be randomly selected as another root node to generate a new Poisson disk, as displayed in Figure 3a.

Step.3 Repeat from Step 1 to Step 2, until all nodes are labeled as representative or inactive. Thus, the adaptive blue noise sampling is completed.

With the adaptive blue noise sampling, the spatial distribution of nodes will be largely retained in the vectorized space. There must be a node specified as a representative in each local area with similar contextual structures, so that none of the contextual features will be missed in the sampled graphs.

4.2.2 Multi-objectives Optimization

Since the blue noise sampling is conducted in a feature space, there are still topological features of significance missed in the sampled graphs. For example, bridging nodes with larger betweenness are not remarkable in the local area that will be easily represented with the other nodes. Also, graph connection is not considered in the course of blue noise sampling, which will easily lead to disconnected sampled results. To enhance multiple contextual structures in the sampled graphs, we integrate a set of optional objectives into the sampling model to preserve various topological features such as node importance and graph connection (R3).

Node Importance: In the course of blue noise sampling, one point is randomly selected as a representative in a Poisson disk. Points with rich contextual structures will be easily discarded due to randomness, such as bridging nodes. To increase the probability of being representatives for those important nodes, we integrate betweenness into the sampling model and update the estimation of radius for a Poisson disk as follows:

$$r = \frac{r_a}{\alpha f(p_i) + \beta f(b_i)} \tag{5}$$

where $f(p_i)$ is still the density estimation of point p_i, and $f(b_i)$ is the betweenness estimation of Poisson disk p_i. In addition, α and β weight the density and betweenness respectively, which can be specified according to user requirements ($\alpha + \beta = 1$ and $\alpha \neq 0$). In this work, they are both initialized as 0.5. It can be seen that the radius of Poisson disk will be small when the density or betweenness of P is high, which increases the probability of important points being selected as representatives. As shown in Figure 3a and Figure 3b, b_1 (Poisson disk of node A) is small so that its Poisson disk radius increases as a new r_a.

Graph Connection: Although the nodes located nearby in the vectorized space likely connect with each other, the connection of sampled nodes will not be guaranteed due to the randomness of sampling. Inspirded by traversal-based graph sampling strategies [5], we optimize graph connection by means of a traversal across different Poisson disks. The course of graph connection based optimization is illustrated in Figure 3c and Figure 3f. Nodes C, D and E are the candidates and C and D are located in a same Poisson disk. If node D is selected (Figure 3c), the candidates of D will be only one point (shaded in green as shown in Figure 3e). The sampled graph will be not connected because contextual structures containing D and E disconnect with each other. Thus, we backtrack to reselect C, which would be a suitable sample owing to its connection with more contextual structures as shown in Figure 3d and Figure 3f. It can be seen that backtracking operations are capable of enhancing the connection relationships of sampled graphs.

4.3 Pseudocode

The pseudocode of our context-aware sampling method is presented in Algorithm 1.

Algorithm 1: Context-aware sampling

Input: V: the node set; B: node betweenness data set; r: radius of user input; N: node neighbor data set; α: adjustable parameter;
Output: S: sampled set;
Algorithm: node2vec: graph representation learning;
t-SNE: dimension reduction;
Dis: calculating the distance between a pair of node2vec points;
KDE: kernel density estimation;
$\beta = 1 - \alpha$;
$V_t = t - SNE(node2vec(V))$;
$kde = KDE(V_t)$;
Add vj into $P(vi)$,$Dis(vj, vi) \leq \frac{r}{\alpha \star kde(vi) + \beta \star B(vi))}$,all $vj \in V$,all $vi \in V$;
$B(vi) = Average(B(P(vz))), vz \in V, vi \in P(vz)$;
Add vi into S;
Delete vj from V, all $vj \in P(vz)$;
Add vi into D,all $vj \in N(vi)$ and $vj \in V$;
while $D \neq \oslash$ **do**
  $B(vi) = MAX(B(P(vz)))$,$vz \in D$;
  Add vi into S;
  Delete vj from V, all $vj \in P(vz)$ and $vj \in V$;
  Delete vj from D, all $vj \in P(vz)$ and $vj \in D$;
  Add vi into D,all $vj \in N(vi)$ and $vj \in V$;
end while;

5 Graph Sampling Framework

We further develop a visualization framework enabling users to visually evaluate the effectiveness of different sampling strategies in the preservation of contextual structures and exploration of latent features of interest in large networks.

5.1 Sampling Strategies and Evaluation Metrics

In our system, categories of sampling strategies [26, 50] are listed including: (1) **Node-based**: *Our method* (OUR), *Random Node Sampling* (RN); (2) **Edge-based**: *Random Edge Sampling* (RE), *Total Induction Edge Sampling* (TIES, focusing on the connectivity of sampled networks based on RE); (3) **Traversal-based**: *Simple Random Walk* (SRW), *Random Jump* (RJ), *Induced Subgraph Random Walk*

Table 1. Evaluation Metrics

Level	Metrics	Motivation
Node	Average Betweenness degree (ABD)	to weight node betweenness in the graph. A sampled graph with larger ABD will retain more bridging nodes [26].
	Average closeness centrality (ACC)	to weight closeness centrality in the graph. If the change of ACC is smaller, it means that nodes with central roles are largely retained in a sampled graph [26].
Connection	Average shortest path length (APL)	to weight shortest paths between each pair of nodes in the graphs. If a sampled graph presents APL with fewer changes, it means that node relationships and graph connectivity are better preserved [30].
	Connected component (CC)	to count the connected components in the graphs. If a sampled graph presents large CC, it means that the graph connection is seriously destroyed based on the sampling operation [26].
Community	Similarity of community (QSC)	to quantify the distribution changes of community histograms. If QSC is small, it means that the community features decrease or disappear based on the sampling operation.
	Structural stability of community (SSC)	to quantify the changes of communities with entropy. If communities share similar nodes, SSC will be small, meaning that community structures change little in the sampled graph.
Cluster	Local clustering coefficient (LCC)	to calculate the degree of local node relationships. If LCC is close to that of original graphs, it means the local structures are well retained [30].
	Global clustering coefficient (GCC)	to calculate the degree of global node relationships. Based on LCC and GCC, we want to evaluate the changes of clustering degrees between the original and sampled graphs [30].

(ISRW, aiming at the retention of local structures based on SRW). A set of quantitative metrics are further to evaluate the validity of sampled networks from different perspectives, as shown in Table 1.

5.2 Visualization and Interactions

We develop a visualization framework to conduct graph sampling and visually evaluate different sampling strategies.

(1) Node-link diagram: A node-link diagram is employed to layout networks with a force-directed model, and the latent structures are presented as shown in Figure 1a. When a sampling strategy is specified, the sampled graph is presented correspondingly with the nodes keeping their original positions, enabling users to visually capture changes of structures. In addition, a set of interactive tools are provided to focus on the structures of interest such as zooming in/out, point selection, region selection, community highlighting, betweenness mapping and connection visualization.

(2) Vectorized visualization: To visually present the contextual structures captured in the vectorized space, we utilize scatterplots to visualize 2D projections of vectors learned by node2vec as shown in Figure 1c. Each point corresponds to a node in the node-link diagram and the geometric distance of points represents the similarity of contextual structures between nodes. A set of interactive tools are also provided in the vectorized space, enabling users to easily perceive and compare the contextual features of network, such as POI (point of interest), ROI (region of interest) and FSM (feature correlation mapping).

(3) Metrics presentation: We also design a set of visual cues to present metrics as shown in Table 1, enabling users to easily compare feature preservation based on different sampling strategies.

Histograms are provided to present desired feature distributions. For example, a stacked bar chart is proposed for users to investigate the difference between original and sampled ABD histograms and estimate whether those nodes with higher betweenness are preserved in the simplified graphs.

Aggregation is designed to help users perceive community association and graph connection. We aggregate each component as a supernode and employ a force-directed model to display the graph components. Thus, a connected network is linked with a set of supernodes. Otherwise, unconnected supernodes will be separated and located around in the aggregation view.

Radar chart is provided to present the overview of metrics. Kinds of metrics are distributed in different directions (Node-based metrics are located in the upper left, connection-based metrics are located in the lower left, community-based metrics are located in the lower right, and cluster-based metrics are located in the upper right). When a sampling strategy or a sampling rate is specified, the polygons in the radar chart will change accordingly, enabling users to evaluate the effectiveness of sampled graphs and compare different sampling methods.

6 Evaluation

In this section, six networks [22, 43] are sampled to evaluate the effectiveness of our sampling method. We firstly compare the performance of different sampling strategies with desired metrics. Then, case studies are conducted based on three real-world datasets, and the validity

Table 2. Quantitative comparison

Strategies	Methods	WW							
		ABD	ACC	APL	CC	GCC	LCC	QSC	SSC
	Original	0.000573	0.099775	10.20919	1	0.024926	0.224315	#	#
Node-based	OURS	**0.004163**	**0.127242**	**8.051543**	**1**	**0.029533**	0.211212	**0.443543**	28.097672
	RNS	0.000776	0.001253	0.066563	1334.6	0.082446	0.014564	0.308587	56.625050
Edge-based	RES	0.003661	0.003505	1.220768	463.2	0.001232	0.001752	0.313593	51.936807
	TIES	0.003552	0.051480	1.387914	68.4	0.382521	0.315257	0.384059	37.014427
Traversal-Based	SRW	0.001491	0.195721	5.854871	1	0.048382	0.161847	0.248175	**14.762372**
	RJ	0.000478	0.002009	0.070126	1335.6	0.135725	0.015973	0.151873	56.625050
	ISRW	0.001054	0.206783	4.485272	1.2	0.123616	**0.222470**	0.311680	16.593352

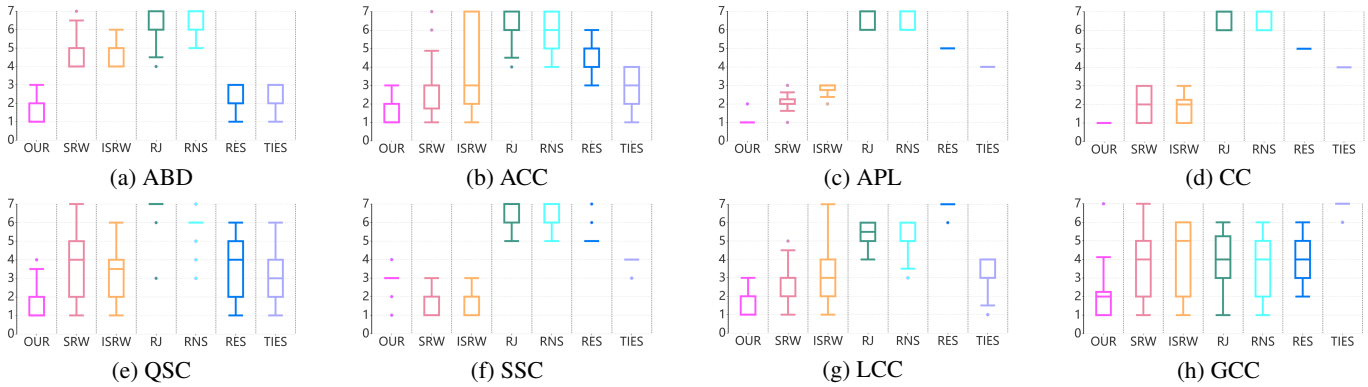


Fig. 4. Ranking comparison based on the metrics of different sampling results. Boxplots are employed to present the distribution of ranks, in which lower ranks mean that the sampling results perform better.

of our method are further discussed.

6.1 Quantitative Comparison

We have conducted a set of experiments to compare different sampling strategies. Specifically, each sampling operation is performed 10 times and the obtained metrics are weighted to make the comparison more convincing. The statistics of metrics for an example data is listed in Table 2. Further, the ranking results of sampling strategies for all datasets are presented in Figure 4.

Table 2 presents the eight metrics obtained by seven sampling strategies for a *Webbase* dataset. The sampling rate is specified as 10%. Obviously, our method performs better on ABD, ACC, APL, CC, GCC and QSC. Especially on the metric of CC, our method presents significant stability. SRW performs better on the metric of SSC, and ISRW performs better on metric of LCC. Our method performs a little inferior on the two metrics (2nd on LCC, 3rd on SSC). For LCC, our method is close to ISRW, both of which perform better than other methods for the preservation of local node relationships. For SSC, both SRW and ISRW are intentionally designed for preserving local structures. As a result, they often lose global features of networks as shown in ABD, GCC and QSC. Instead, our method gains more balanced representation of desired characteristics. The above results demonstrate that our context-aware sampling model effectively maintains the structural characteristics of original networks.

Figure 4 presents the overview of the ranking performance of metrics for all datasets. According to the ranks in Figure 4(a-c), our method outperforms on ABD, ACC and APL, indicating that important nodes with high betweenness degree are well preserved. To evaluate the connectivity of sampled graphs, we calculated the connected components of sampled graphs obtained by different sampling strategies. As shown in Figure 4d, node-based or edge-based sampling methods (RNS, RES, TIES) often generate multiple disconnected components, while traversal-based sampling methods (SRW, ISRW) generate connected graphs with greater probability. Our method traverses the Poisson disks to guarantee the connection of graphs, which also presents fine performance on CC. The above results confirm that our method can preserve the nodes and edges with contextual relationships. Figure 4(e-h) display the ranking performance of all sampling strategies on the four metrics on community and cluster. It can be seen that ISRW performs better in maintaining local structures, because it explicitly prioritizes the preservation of local nodes and edges, as shown in the ranking of SSC (Figure 4f) and LCC (Figure 4g). However, such preservation might lose the integrity of global structures as shown in the ranking results of GCC (Figure 4h). By contrast, our method performs better than ISRW on QSC and GCC, and better than most sampling strategies on SSC and LCC. We think that our method makes a good balance between the preservation of local features and global structures, and preserves feature distribution more reasonably and stably. Comparable results demonstrate the effectiveness of our method in preserving the contextual structures of networks.

6.2 Case Study

In this section, we conduct three case studies based on the real-world network datasets from a variety of application fields such as bitcoin trade, literature citation and webbase. According to section 6.1, we compare our sampling method with a set of representative methods, such as RNS (node-based), TIES (edge-based, better than RES) and ISRW (traversal-based, better than SRW, RJ).

Case 1. Contextual structure preservation

Communities are significant contextual structures, and the changes

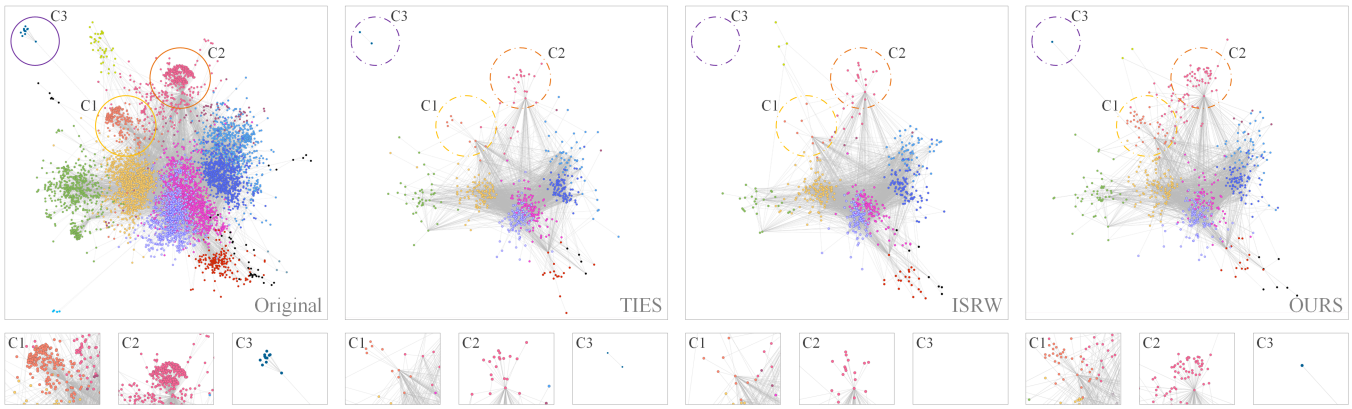


Fig. 5. Comparison of contextual structure preservation for a *Bitcoin* dataset (5k nodes and 21k edges). Community features of interest are circled in the original node-link diagram such as C1, C2 and C3. Three sampling strategies such as TIES, ISRW and Ours are performed and the graphs are presented with the retained community features circled. The detailed structures of community features are highlighted in the little figures at below.

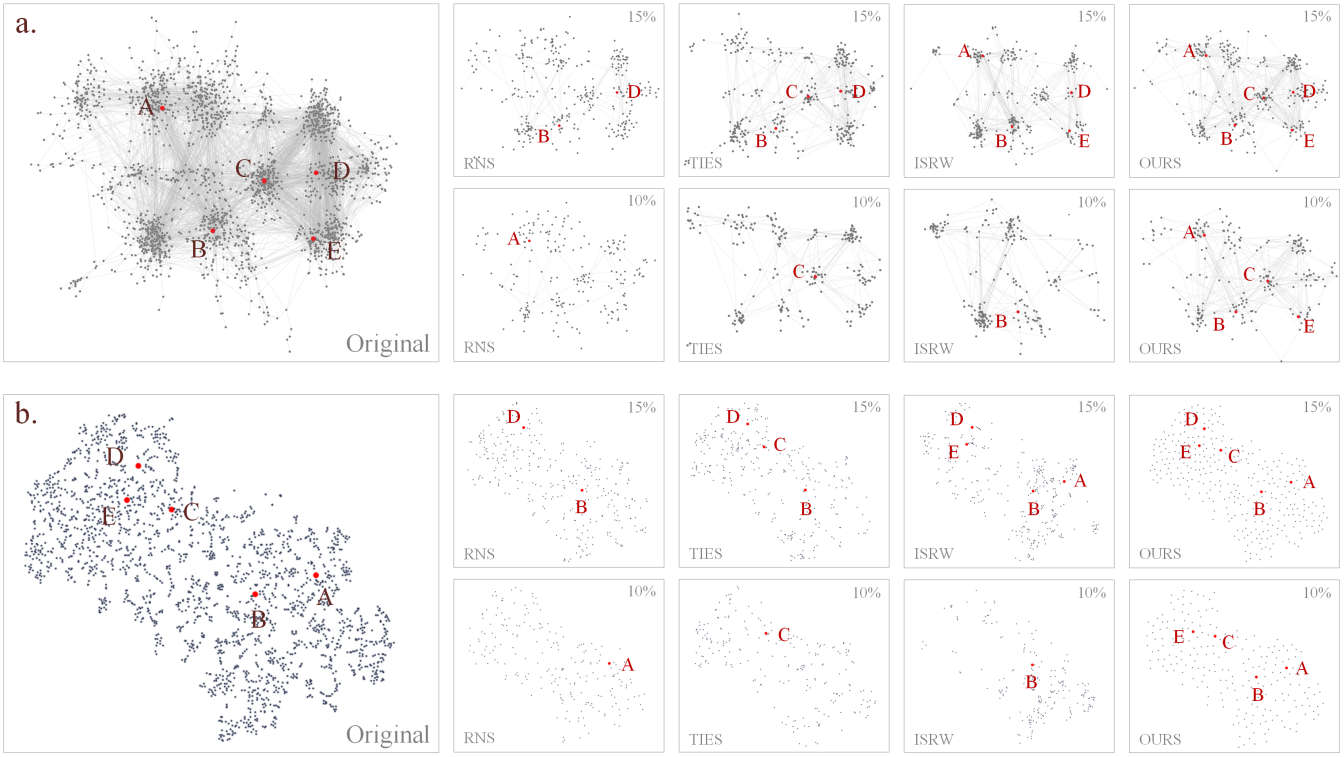


Fig. 6. Comparison of node importance preservation for a *IEEE_VIS* dataset (2k nodes and 8k edges). Nodes shaded in red (A, B, C, D and E) are significant owing to their higher betweenness values. The remained nodes are also highlighted in the sampled graphs obtained by RNS, TIES, ISRW and Ours (sampling rates are specified as 15% and 10%).

of communities will mislead users in the exploration of networks. Thus, we observe and compare the changes of communities to demonstrate that our method can preserve the contextual structures of interest like aggregated nodes.

As shown in Figure 5, there are obvious local communities (C1: yellow, C2: orange, C3: purple) in a bitcoin trading network *Bitcoin*. Each community represents a different trading mode respectively. In our experiments, we use a variety of sampling strategies to conduct graph simplification with the sampling rate specified as 10%. Sampled results are presented in Figure 5. It can be seen that ISRW retains the communities C1-C2 and loses community C3. Communities C1-C3 are retained in the sampled results of TIES, but their scales are obviously imbalance. For instance, scales of C1 and C2 are smaller than expected, and structures of C3 are evidently broken. By contrast, our method does not only retain communities C1-C3, but also preserves accounts of nodes as uniform as possible. Meanwhile, we further retain their connections and original structures in a relatively balanced way. The preservation of community features does great favors for further pattern recognition and analysis of bitcoin transaction. The experimental results prove the effectiveness of our algorithm in maintaining contextual structures of original networks.

Case 2. Important node preservation

In our method, we preserve nodes with larger betweenness by adjusting the radii of Poisson disks. Figure 6 presents the comparison of sampled results for a *IEEE_VIS* network. The node-link diagram is shown in Figure 6a and the corresponding projection is presented in Figure 6b.

There are five nodes with larger betweenness identified in the literature network, such as A, B , C, D and E in Figure 6a. These points are not only the key nodes in the literature network, but also important figures in the specific field. It can be observed that all points can be retained with our method and ISRW at the sampling rate specified as 15%. Two nodes are retained with RNS and three nodes are retained with TIES. Then, we further decrease the sampling rate as 10%. More significant nodes disappear in the sampled results, such as RNS, TIES and ISRW. By contrast, our method presents well stability, still retaining almost all significant nodes. To further demonstrate the effectiveness of parameters in our sampling model, we change α and β from $(\alpha = 0.5, \beta = 0.5)$ to $(\alpha = 0.8, \beta = 0.2)$. Obviously, the sampling operation with a new set of parameters pays little attention to node importance. Figure 7 presents the sampled graphs (sampling rate is specified as 20%) in which nodes with larger betweenness are highlighted. We can see that point C is missed in Figure 7a $(\alpha = 0.8, \beta = 0.2)$ while five points are retained in Figure 7b $(\alpha = 0.5, \beta = 0.5)$, which demonstrates the usefulness of parameters for our context-aware sampling. To sum up, the retention of significant nodes is of great importance for subsequent network analysis, graph calculation and literature data processing.

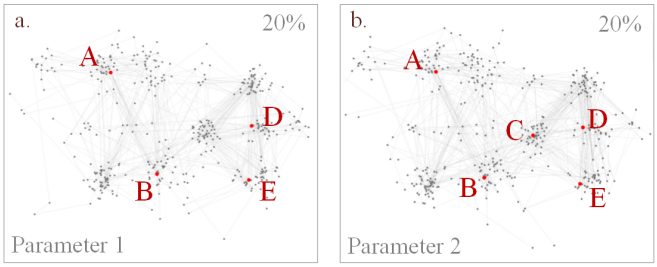


Fig. 7. Important nodes are retained based on different parameters. (a) A, B, D and E are retained with $(\alpha = 0.8, \beta = 0.2)$ (b) A ,B, C, D and E are all retained with $(\alpha = 0.5, \beta = 0.5)$.

Case 3. Graph connection preservation

Nodes and edges together constitute a rich set of network paths, which are important for graph analysis. If a critical path is missing in sampled graph, it is likely to bring much misreading for network understanding. In our method, blue noise sampling model is optimized with multiple targets integrated to retain key connectivity paths in the sampled graph as far as possible. As shown in Figure 8, *Webbase* is a

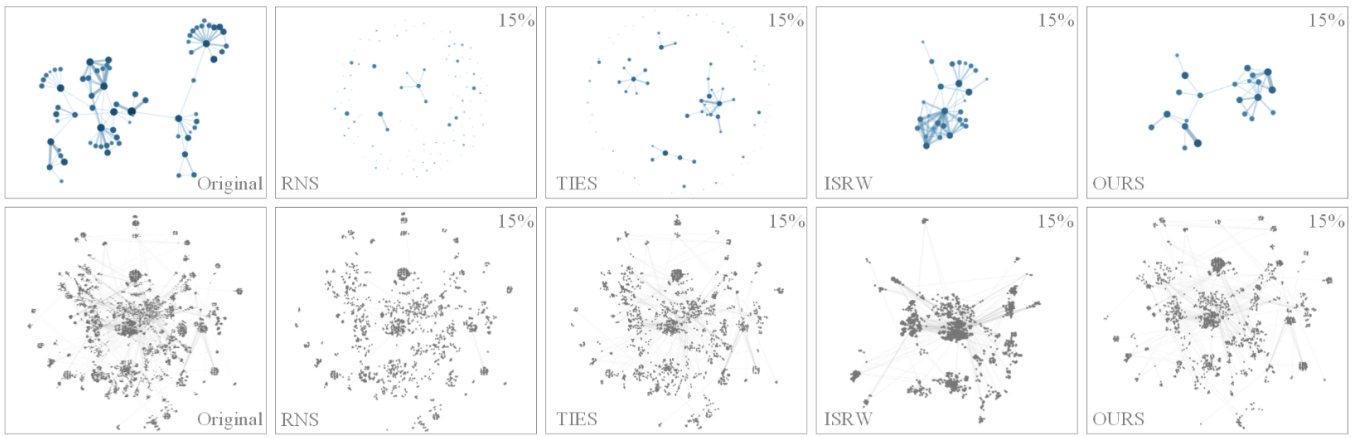
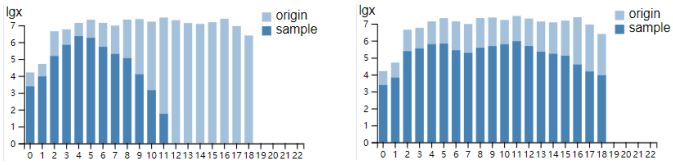


Fig. 8. Comparison of graph connection retained for different sampling strategies for a *Webbase* dataset. Node-link diagrams are presented at the bottom, and corresponding supernode-based diagrams are presented at on the top.



(a) ISRW (b) Ours

Fig. 9. Comparison of APL distribution histograms in which the changes of shortest paths between all pairs of nodes are statistically presented based on different sampling strategies.

connected network consisting of many connected communities. With the sampling rate specified as 15%, RNS generate a large number of disconnected components, which loses the association of communities. Similarly, there are a large number of disconnected aggregations obtained by TIES. Fortunately, some communities are connected with each other, retaining original connectivity features. By contrast, ISRW and our method seem to well maintain the connectivity of original network. But, there is still significant difference between the aggregation results.

To further analyze the difference, we use distribution histograms of APL to observe shortest distances between nodes, as shown in Figure 9. Compared with original distribution, APL distributions of ISRW are concentrated in 0-11. It means that those nodes with smaller distances are retained, while a considerable part of nodes located far away from each other are missed. It indicates that a lot of critical paths might be lost in the sampled graph. By contrast, APL distributions of our method are located from 0 to 18, which present similar to that of original graph. It indicates that major paths of the original graph are still retained in the sampled graph. According to the comparison results, our method performs better than the other sampling strategies in maintaining graph connectivity and the key paths of networks.

6.3 Discussion

Compared with traditional sampling strategies, the context-aware sampling model proposed in this paper performs better on most of the metrics. The mainly advantages are that we conduct a blue noise sampling operation on a feature space obtained by GRL, where the contextual structures are well represented. In addition, multiple objectives are integrated into the sampling model to make the sampled graphs preserve more contextual structures of interest such as node importance and graph connectivity. Therefore, our sampling method cannot only reduce the visual clutter generated in the original large graph visualization, but also preserve and enhance the contextual structures of significance in the sampled graphs.

Meanwhile, there are still some issues not well resolved in this paper, which will be addressed in the future work. (1) Some errors are inevitably generated due to the randomness of GRL and approximation of dimensionality reduction with t-SNE. In the future work, we will specify an algorithmic model to quantify the inevitable errors for the evaluation of GRL. The results of evaluation are able to help users construct a more accurate mapping between expressions of features in the representation learning space and network space. (2) In addition to the errors generated in the courses of representation learning and dimensionality reduction, blue noise sampling will also bring sampled graphs with much uncertainty. In this paper, our visual designs are still not enough for a deeply optimization of sampled graphs. In the future work, we will further study a collaborative model to represent sampling uncertainty, integrate it into the course of blue noise sampling and design an uncertainty-aware sampling framework for the exploration of large networks. (3) In this paper, our proposed sampling method is based on a model of node2vec. Of course, more graph learning models have been studied to represent various structures of networks. In the future work, we will integrate other graph representation learning models into our sampling method, providing users with a set of sampling options, to obtain desired graph simplification according to their requirements.

7 Conclusion

In this paper, node2vec is utilized to represent the contextual structures of original networks and a multi-objective blue noise sampling model is designed to reduce the visual clutter of densely graph visualization with the contextual structures of interest preserved in the sampled graphs. Multiple metrics are employed to measure the validity of context-aware sampling method from different perspectives such as node importance, graph connection and community stability. In addition, a set of visual interfaces are provided enabling users to visually compare the sampling strategies and explore latent features of interest. Quantitative comparisons and case studies based on real-world datasets have demonstrated the effectiveness of our system in simplifying large networks with contextual structures preserved.

Acknowledgments

We would like to thank the reviewers for their thoughtful comments. The work is supported in part by the National Natural Science Foundation of China (No.61872314, 61802339, 41901363, 61872388, 61772456 and 61761136), the Humanities and Social Sciences Foundation of Ministry of Education in China (No.18YJC910017), the Natural Science Foundation of Zhejiang Province (No.LY18F020024 and LGF20G010003), the Major Humanities and Social Sciences Research Projects in Colleges of Zhejiang Province (No.2018QN021), the Open Project Program of the State Key Lab of CAD&CG of Zhejiang University (No.A2001) and the First Class Discipline of Zhejiang-A (Zhejiang University of Finance and Economics-Statistics).

REFERENCES

[1] J. Abello, F. V. Ham, and N. Krishnan. Ask-graphview: A large scale graph visualization system. *IEEE Transactions on Visualization & Computer Graphics*, 12(5):669–676, 2006. doi: 10.1109/TVCG.2006.120

[2] K. Ammar and M. T. Özsu. Experimental analysis of distributed graph systems. vol. 11, pp. 1151–1164, 2018. doi: 10.14778/3231751.3231764

[3] R. Angles and C. Gutierrez. Survey of graph database models. *ACM Computing Surveys*, 40(1):1–39, 2008. doi: 10.1145/1322432.1322433

[4] A.-L. Barabäi. Linked: The new science of networks. *American Journal of Physics*, 71(4):409–410, 2003. doi: 10.1063/1.1570778

[5] W. Chen, F. Guo, D. Han, J. Pan, et al. Structure-based suggestive exploration: A new approach for effective exploration of large networks. *IEEE Transactions on Visualization & Computer Graphics*, 25(1):555–565, 2019. doi: 10.1109/TVCG.2018.2865139

[6] W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li. Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization & Computer Graphics*, 14(6):1277–1284, 2008. doi: 10.1109/TVCG.2008.135

[7] N. Elmqvist and J. Fekete. Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. *IEEE Transactions on Visualization & Computer Graphics*, 16(3):439–454, 2010. doi: 10.1109/TVCG.2009.84

[8] E. R. Gansner, Y. Hu, S. North, and C. Scheidegger. Multilevel agglomerative edge bundling for visualizing large graphs. In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, pp. 187–194. IEEE, 2011. doi: 10.1109/PACIFICVIS.2011.5742389

[9] M. Ghoniem, J. D. Fekete, and P. Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *Proceedings of the 10th IEEE Symposium on Information Visualization*, pp. 17–24. IEEE, 2004. doi: 10.1109/INFVIS.2004.1

[10] H. Gibson, J. Faith, and P. Vickers. A survey of two-dimensional graph layout techniques for information visualisation. *Information Visualization*, 12(3-4):324–357, 2013. doi: 10.1177/1473871612455749

[11] A. Grover and J. Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 855–864. ACM, 2016. doi: 10.1145/2939672.2939754

[12] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *IEEE Database Engineering Bulletin*, 40(3):52–74, 2017.

[13] 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. ACM, 2013. doi: 10.1145/2488388.2488436

[14] D. Hennessey, D. Brooks, A. Fridman, and D. E. Breen. A simplification algorithm for visualizing the structure of complex graphs. In *Proceedings of the 12th International Conference on Information Visualisation*, pp. 616–625. IEEE, 2008. doi: 10.1109/IV.2008.37

[15] D. Holten. Hierarchical edge bundles: visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization & Computer Graphics*, 12(5):741–748, 2006. doi: 10.1109/TVCG.2006.147

[16] D. Holten and J. J. van Wijk. Force-directed edge bundling for graph visualization. *Computer Graphics Forum*, 28(3):983–990, 2010. doi: 10.1111/j.1467-8659.2009.01450.x

[17] S. Hong, Q. H. Nguyen, A. Meidiana, J. Li, et al. Bc tree-based proxy graphs for visualization of big graphs. In *Proceedings of the 2018 IEEE Pacific Visualization Symposium*, pp. 11–20. IEEE, 2018. doi: 10.1109/PacificVis.2018.00011

[18] J. Hu, S.-H. Hong, and P. Eades. Spectral vertex sampling for big complex graphs. In *Proceedings of the the 8th International Conference on Complex Networks and Their Applications*, vol. 882, pp. 216–227. Springer, 2019. doi: 10.1007/978-3-030-36683-4_18

[19] Y. Hu and L. Shi. Visualizing large graphs. *Wiley Interdiplinary Reviews Computational Stats*, 7(2):115–136, 2015. doi: 10.1002/wics.1343

[20] Y. Jia, J. Hoberock, M. Garland, and J. Hart. On the visualization of social and other scale-free networks. *IEEE Transactions on Visualization & Computer Graphics*, 14(6):1285–1292, 2008. doi: 10.1109/TVCG.2008.151

[21] A. Kennedy, K. Klein, A. Nguyen, and F. Y. Wang. The graph landscape: using visual analytics for graph set analysis. *Journal of Visualization*, 20(3):417–432, 2017. doi: 10.1007/s12650-016-0374-6

[22] S. Kumar, B. Hooi, D. Makhija, M. Kumar, et al. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 333–341. ACM, 2018. doi: 10.1145/3159652.3159729

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

[24] O. Kwon, T. Crnovrsanin, and K. Ma. What would a graph look like in this layout? a machine learning approach to large graph visualization. *IEEE Transactions on Visualization & Computer Graphics*, 24(1):478–488, 2018. doi: 10.1109/TVCG.2017.2743858

[25] V. D. M. Laurens and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008. doi: 10.2312/eurovisshort.20161164

[26] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 631–636. ACM, 2006. doi: 10.1145/1150402.1150479

[27] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187. ACM, 2005. doi: 10.1145/1081870.1081893

[28] M. Liu, J. Shi, K. Cao, J. Zhu, et al. Analyzing the training processes of deep generative models. *IEEE Transactions on Visualization & Computer Graphics*, 24(1):77–87, 2018. doi: 10.1109/TVCG.2017.2744938

[29] M. Liu, J. Shi, Z. Li, C. Li, et al. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization & Computer Graphics*, 23(1):91–100, 2017. doi: 10.1109/TVCG.2016.2598831

[30] 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 & Data Mining*, pp. 105–113. ACM, 2011. doi: 10.1145/2020408.2020431

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*, 2013.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, et al. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013.

[33] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, et al. graph2vec: Learning distributed representations of graphs. *CoRR*, abs/1707.05005, 2017.

[34] Q. H. Nguyen, S. Hong, and P. Eades. dnng: Quality metrics and layout for neighbourhood faithfulness. In *Proceedings of the 2017 IEEE Pacific Visualization Symposium*, pp. 290–294. IEEE, 2017. doi: 10.1109/PACIFICVIS.2017.8031607

[35] Q. H. Nguyen, S. Hong, P. Eades, and A. Meidiana. Proxy graph: Visual quality metrics of big graph sampling. *IEEE Transactions on Visualization & Computer Graphics*, 23(6):1600–1611, 2017. doi: 10.1109/TVCG.2017.2674999

[36] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[37] A. Parada-Mayorga, D. L. Lau, J. H. Giraldo, and G. R. Arce. Blue-noise sampling on graphs. *IEEE Transactions on Signal and Information Processing over Networks*, 5(3):554–569, 2019. doi: 10.1109/TSIPN.2019.2922852

[38] Y. Pei, X. Du, J. Zhang, G. Fletcher, et al. struc2gauss: Structure preserving network embedding via gaussian embedding. *CoRR*, abs/1805.10043, 2018.

[39] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 701–710. ACM, 2014. doi: 10.1145/2623330.2623732

[40] R. Pienta, M. Kahng, L. Zhiyuan, J. Vreeken, et al. Facets: Adaptive local exploration of large graphs. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 597–605. SIAM, 2017. doi: 10.1137/1.9781611974973.67

[41] H. C. Purchase, E. Hoggan, and C. Görg. How important is the "mental map"? An empirical investigation of a dynamic graph layout algorithm. In *Proceedings of the 14th International Conference on Graph drawing*, pp. 184–195. Springer, 2007. doi: 10.1007/978-3-540-70904-6_19

[42] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo. Struc2vec: Learning

node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 385–394. ACM, 2017. doi: 10.1145/3097983.3098061

[43] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 4292–4293. AAAI, 2015.

[44] M. Salathë, M. Kazandjieva, J. W. Lee, P. Levis, et al. A high-resolution human contact network for infectious disease transmission. vol. 107, pp. 22020–22025, 2010. doi: 10.1073/pnas.1009094108

[45] A. D. Sarma, D. Nanongkai, G. Pandurangan, and P. Tetali. Distributed random walks. *Journal of the ACM*, 60(1):1–31, 2013. doi: 10.1145/2432622.2432624

[46] L. Tang, B. Bie, and D. Zhi. Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease. *American Journal of Infection Control*, 46(12):1375–1380, 2018. doi: 10.1038/ncomms15186

[47] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk. Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization & Computer Graphics*, 22(1):1–10, 2016. doi: 10.1109/TVCG.2015.2468078

[48] W. van Heeswijk, G. H. L. Fletcher, and M. Pechenizkiy. On structure preserving sampling and approximate partitioning of graphs. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pp. 875–882. ACM, 2016. doi: 10.1145/2851613.2851650

[49] Y. Wang, Q. Shen, D. Archambault, Z. Zhou, et al. Ambiguityvis: Visualization of ambiguity in graph layouts. *IEEE Transactions on Visualization & Computer Graphics*, 22(1):359–368, 2016. doi: 10.1109/TVCG.2015.2467691

[50] Y. Wu, N. Cao, D. Archambault, Q. Shen, et al. Evaluation of graph sampling: A visualization perspective. *IEEE Transactions on Visualization & Computer Graphics*, 23(1):401–410, 2017. doi: 10.1109/TVCG.2016.2598867

[51] V. Yoghourdjian, T. Dwyer, K. Klein, K. Marriott, et al. Graph thumbnails: Identifying and comparing multiple graphs at a glance. *IEEE Transactions on Visualization & Computer Graphics*, 24(12):3081–3095, 2018. doi: 10.1109/TVCG.2018.2790961

[52] S.-H. Yoon, K.-N. Kim, J. Hong, S.-W. Kim, et al. 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

[53] J. Yuan, C. Chen, W. Yang, M. Liu, et al. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7(1):1–31, 2021.

[54] J. Zhang, Y. Pei, G. H. L. Fletcher, and M. Pechenizkiy. Structural measures of clustering quality on graph samples. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 345–348. IEEE, 2016. doi: 10.1109/ASONAM.2016.7752256

[55] Y. Zhao, F. Luo, M. Chen, Y. Wang, et al. Evaluating multi-dimensional visualizations for understanding fuzzy clusters. *IEEE Transactions on Visualization & Computer Graphics*, 25(1):12–21, 2018. doi: 10.1109/TVCG.2018.2865020

[56] Z. Zhou, L. Meng, C. Tang, Y. Zhao, et al. Visual abstraction of large scale geospatial origin-destination movement data. *IEEE Transactions on Visualization & Computer Graphics*, 25(1):43–53, 2019. doi: 10.1109/TVCG.2018.2864503

[57] Z. Zhou, C. Shi, M. Hu, and Y. Liu. Visual ranking of academic influence via paper citation. *Journal of Visual Languages & Computing*, 48:134–143, 2018. doi: 10.1016/j.jvlc.2018.08.007

[58] M. Zinsmaier, U. Brandes, O. Deussen, and H. Strobelt. Interactive level-of-detail rendering of large graphs. *IEEE Transactions on Visualization & Computer Graphics*, 18(12):2486–2495, 2012. doi: 10.1109/TVCG.2012.238