A Visual and Statistical Benchmark for Graph Sampling Methods

Fangyan Zhang[1] Song Zhang[1] Pak Chung Wong[2]

[1]Mississippi State University, [2]Pacific Northwest National Laboratory

ABSTRACT

Effectively visualizing large graphs containing millions of nodes and edges is challenging. Simplification algorithms developed to observe and analyze large graphs in a more feasible manner are indispensable for this task. Several varieties of sampling approaches for graph simplification have been proposed. It is still an open question, however, which single sampling technique produces the best representative sample. The goal of this paper is to evaluate commonly used sampling methods through a combined visual and statistical comparison. The visual comparison is facilitated by employing a fixed graph layout. The resulting benchmark can be incorporated into graph visualization and analysis tools.

Keywords: graph sampling, graph properties.

1 INTRODUCTION

Graphs are widely used for information visualization, particularly for those datasets that can be easily represented as a network [1]. Graph analysis and visualization [2] has evolved into a very active area of investigation over the last several decades. However, as the size of a graph grows, effectively displaying all of the nodes and edges becomes an extremely difficult challenge. Sampling algorithms, one type of simplification method, aim to reduce this drawing complexity by abridging massive graphs into a representative sample of the original[3]. Visualization and analysis that would otherwise be prohibitive to perform on the original large graph can be realized on the smaller sample. Ideally, the generated sample should contain nearly identical features as the original, allowing analysis of the small sample to yield the characteristics possessed by the original graph.

While numerous graph sampling techniques have been proposed[3], few comparisons between those methods, especially with regard to the effectiveness of the visualization, have been performed. Questions still remain on how best to evaluate those approaches and which sampling method is more suitable for a particular application. In order to answer those questions, proper graph properties and metrics for useful comparison must be identified. Evaluation metrics for the visualization of graph sampling methods, in particular, are lacking[4]. Therefore, it is valuable to construct a benchmark from meaningfully comparison across approaches. The results will help users choose effective sampling algorithms for their application needs and graph analysis.

In this paper, we design a benchmark for comparing a number of graph sampling methods. Our comparison considers two complementary aspects: how effectively the method preserves the visual properties and how well they preserve the statistical properties of the graph. In order to properly compare graph sampling methods for visualization, we fixed the graph layout in both the original and sampled graphs. As a result, the differences among the sampling methods become more apparent and discernable.

The main contributions of our work are as follows:

- We build a benchmark for evaluating graph sampling methods for both visual and statistical properties.
- We implement various graph sampling techniques and analyze them within the benchmark.
- We apply the benchmark across two different data sets.

2 VISUAL AND STATISTICAL BENCHMARK

2.1 Graph Sampling Methods

Within the benchmark, we implement Random Node sampling (RN), Random Edge sampling (RE), Random Node-Edge sampling (RNE), Random PageRank Node sampling (RPN), Random Node Neighbor sampling (RNN), Random Walk sampling (RW), Random Jump sampling (RJ), Forest Fire sampling (FF), Induced Edge sampling (IE), Streaming Nodes sampling (SN) and Streaming Edge sampling (SE). We apply these sampling methods to two datasets: American airline connections and VAST challenge data 2013. The American airline connections data are a static graph while VAST data are a dynamic multigraph. Due to space limitation, we provide the results of only a subset of these methods in this poster abstract for illustration purposes. We provide full results in the supplementary document.

2.2 Statistical Comparison

We utilize several statistical properties we want to preserve when sampling a graph. These properties form distributions on graphs. Hence we employ Kolmogorov-Smirnov D-statistic to evaluate the similarity between the original graph and a sampled graph based on those properties. We list the properties for static graph and dynamic graph separately.

2.2.1 Statistical properties for static graphs

We utilize the following static graph properties:

- In-Degree distribution (in-deg).
- Out-Degree distribution (out-deg).
- Weakly connected component distribution (wcc).
- Strongly connected component distribution (scc).
- Hops distribution [5].
- Hops on the largest weakly connected component (hops in largest wcc).

2.2.2 Statistical properties for dynamic graphs

We utilize the following dynamic graph properties:

- Densification Power Law (DPL). This property tests the power law relation between the number of nodes and the number of edges.
- The distribution of connected component (cc).
- The distribution of the largest singular value of graph adjacency matrix (sng-val).

2.3 Visual Comparison

We use Gephi[6] for visual comparison between sampling methods. We first drew the original graph and use the layout of the original graph for all sampled graph, i.e., the same node in all sampled graphs will occupy the same location as in the original graph.

For the American airline connection data, we used the geospatial layout. For the VAST challenge data, we used the force-directed layout.

3 RESULTS

We applied the sampling techniques to American airline connections dataset to retain about 25% percentage of nodes. A part of the result is shown below. We include the full result in the supplementary material.

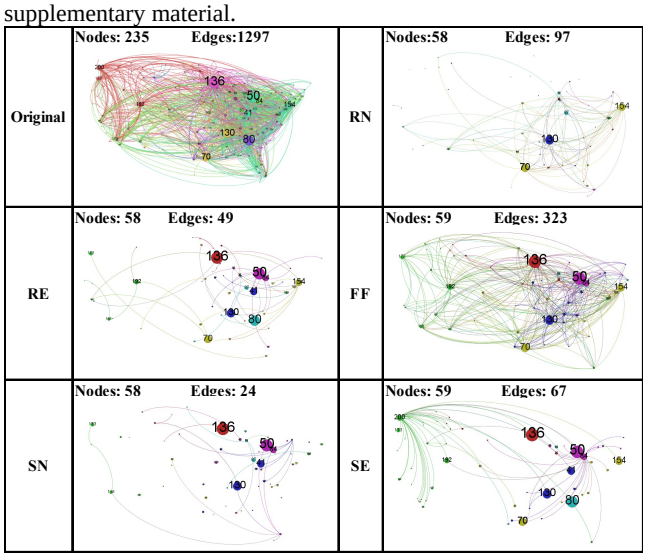

Figure 1: Visual comparison between sampled graphs for American airline connection data.

Table 1: Statistical comparison between sampled graphs for American airline connection data.

		Matrix					
		in-degree	out-degree	wcc	scc	hops	Hops-wcc
Algorithm	RN	0.384	0.313	1	0.014	0.833	0.848
	RE	0.487	0.513	1	0.013	0.924	0.924
	FF	0.252	0.052	1	0.153	0.688	0.684
	SN	0.608	0.738	1	0.013	0.924	0.939
	SE	0.507	0.809	1	0.013	0.879	0.879

Table 1 shows the statistical comparison between sampling algorithms on the American airline connection data.

We also applied the sampling techniques to VAST 2013 challenge 3 netflow dataset (week 3) to retain about 25% percentage of the nodes. Part of the result is shown below.

Table 2: Statistical comparison for VAST data.

		Matrix		
		DPL	CC	Sngl-value
Algorithm	RN	0.354	0.661	0.109
	RE	0.479	0.746	0.109
	FF	0.428	0.565	0.058
	SN	0.487	0.835	0.022
	SE	0.587	0.935	0.109

From the statistical comparison, we can compare different sampling methods quantitatively. Users can select a sampling algorithm based on the graph properties important to them, the type of graphs they are working with, and the capability of the sampling algorithm to preserve the graph properties.

Visual comparison results give a direct perception of the sampled graphs and allow qualitative analysis of the sampled graphs. For example, from the visual comparison results, we can conclude that the edge-sampling algorithms, e.g., random edge or forest fire, are biased toward high degree nodes. They are more suitable if high degree nodes are important in an application. For example, RE sampling in both datasets obtains more high degree nodes than RN sampling, even though the two sampling results have similar number of edges.

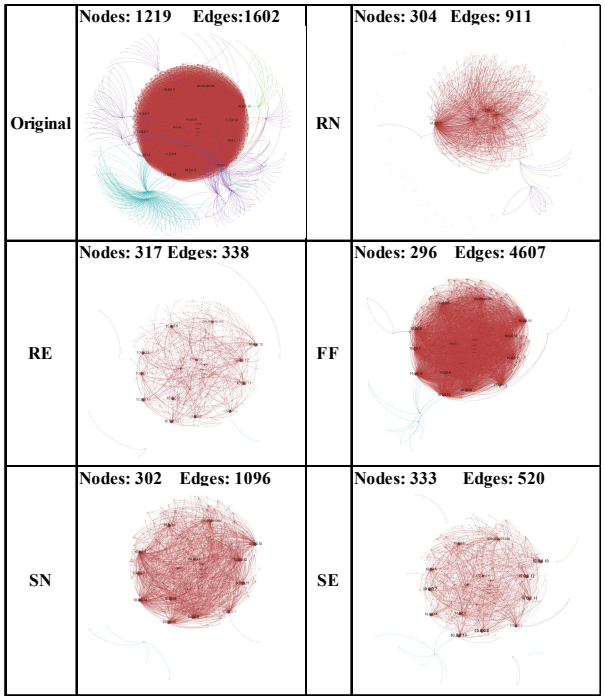

Figure 2: Visual comparison between sampled graphs for VAST data.

4 CONCLUSION AND DISCUSSION

Our visual and statistical benchmark evaluates a number of graph sampling methods based on their effectiveness in preserving both the quantitative statistical properties and qualitative visual properties of the large original graph. The results provide insight into the effectiveness of each sampling method, aiding users with their choices of these methods in applications.

REFERENCES

[1] I. Herman, etc.al., "Graph visualization and navigation in information visualization: a survey,"*TVCG.*, vol. 6, no. 1, pp. 24–43, 2000.

[2] E. A. Lopez-Rojas, "Social Network Analysis in the dataset US Air 97 with Pajek," *LiU*, no. 1, pp. 1–11, 2011.

[3] P. Hu and W. Lau, "A survey and taxonomy of graph sampling," *arXiv.org*, pp. 1–34, 2013.

[4] J. Leskovec etc.al. , "Sampling from large graphs," *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 631–636, 2006.

[5] C. R. Palmer, P. B. Gibbons, and C. Faloutsos, "ANF: a fast and scalable tool for data mining in massive graphs," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, 2002, p. 81.

[6] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," *Third Int. AAAI Conf. Weblogs Soc. Media*, pp. 361–362, 2009.