A Visual and Statistical Benchmark for Graph Sampling Methods

Fangyan Zhang[1] Song Zhang[1] Pak Chung Wong[2] J. Edward Swan II[1] T.J. Jankun-Kelly[1]
[1]Mississippi State University [2]Pacific Northwest National Laboratory

Abstract

Effectively visualizing large graphs is challenging[2]. Capturing the statistical properties of these large graphs is also difficult. Sampling algorithms, developed to more feasibly observe and analyze large graphs, are indispensable for this task. Many sampling approaches for graph simplification have been proposed[4]. However, it is still an open question which single sampling technique produces the best representative sample. We create a benchmark to evaluate commonly used sampling methods through a combined visual and statistical comparison.

Objective

To design a benchmark for comparing a number of graph sampling methods on directed and undirected graphs separately and use a number of statistical properties for comparison.

Sampling Methods

Node Sampling

- Random node (RN)
- Random node-edge (RNE)
- Random node-neighbour (RNN)
- Streaming nodes (SN)

Edge Sampling

- Random edge (RE)
- Induced edge (IE)
- Streaming edge (SE)

Topology-based Sampling

- Snowball (SB)
- Random walk (RW)[5]
- Random walk with escape (RWE)
- Forest fire (FF)
- Breadth-first (BF)
- Depth-first (DF)
- Random first (RF)

Sampling Rate

- All sampling rates refer to the number of edges instead of nodes.

Evaluation Technique

Kolmogorov-Smirnov (KS) D-statistic

- $D_n = \sup_x |F_n(x) - F(x)|$ $\sup_x$: supremum of the set of distance. F_n and F are distribution function.
- To evaluate the similarity between original and sampled graph based on graph properties.

Comparison Categories

- Between directed and undirected Graph
- Between different graph types of undirected or directed graphs.
- Between multiple graphs of same type but different sizes.
- Between multiple graphs of same size and type.

Graph Data

- American Airlines connection data
- VAST 2013 net-flow data
- Simulated random graph data up to 1 billon nodes and 500 billion edges.

Graph Properties

Directed Graph

- In-degree distribution (InDD)
- Out-degree distribution (OutDD)
- Betweenness centrality distribution (BB)
- Average neighbour degree distribution (ANDD)
- In-degree centrality distribution (InDCD)
- Out-degree centrality distribution (OutDCD)
- Hops distribution (HD)[1]
- Edge betweenness centrality distribution (EBCD)
- Weakly connected component distribution(WCCD)
- Hops distribution in largest weakly connected component (HLCCD)

Undirected Graph

- Degree distribution (DD)
- Betweenness centrality distribution (BB)
- Clustering coefficient (CCD)
- Average neighbor degree distribution (ANDD)
- Degree centrality distribution (DCD)
- Edge betweenness centrality distribution (EBCD)
- Hop distribution (HD)

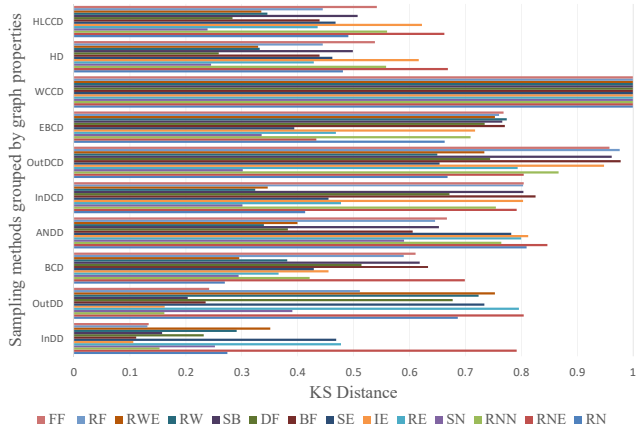
Statistical Comparison Results

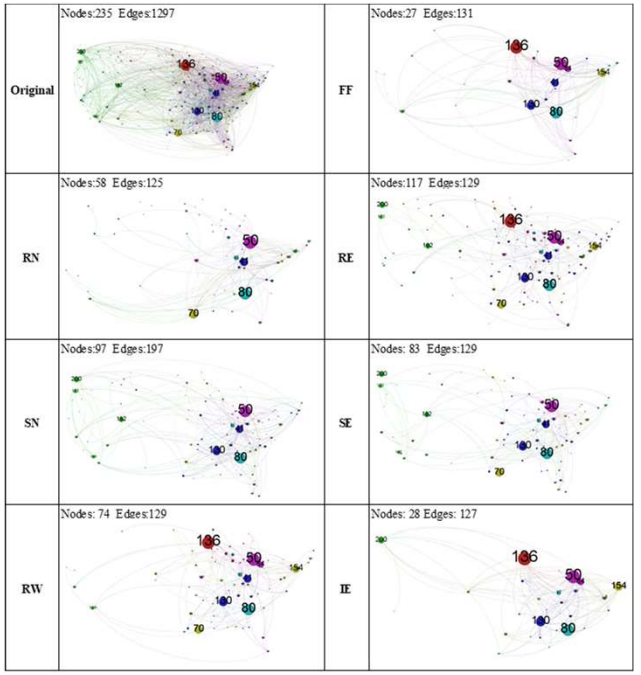
- KS values represent two distribution similarities.
- KS values are visualized using barchart.
- The comparison of sampling methods are based on and grouped by graph properties.
- The higher sampling rate is, the closer statistical properties to the original graph.
- The graph sampling methods are dependent on graph type, size etc.

Visual Comparison Results

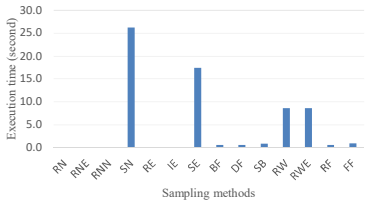
- Nodes position, label size, label color etc. are fixed.
- Gephi[3]is used to visualize the graph.
- From visual comparison, the differences between sample results can be intuitively understood.
- Edge-related sampling methods, for example, RE, IE, SE are biased towards high-degree nodes.

Statistical and Visual Comparison





Efficiency Comparison



Conclusion

- Sampling methods have different performance on graph properties，data type, graph sampling topology etc. No sampling method work well for all type of graphs.
- The benchmark helps users choose proper sampling methods in application.
- The benchmark provides an avenue to explore big graphs using appreciate sampling methods.

Reference

[1] J. Leskovec and C. Faloutsos, "Sampling from large graphs," *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 631–636, 2006.
[2] D. Rafiei and S. Curial, "Effectively visualizing large networks through sampling," in *Proceedings of the IEEE Visualization Conference*, 2005, p. 48.
[3] 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.
[4] P. Hu and W. Lau, "A survey and taxonomy of graph sampling," *arXiv.org*, pp. 1–34, 2013.
[5] B. Ribeiro and D. Towsley, "Estimating and Sampling Graphs with Multidimensional Random Walks," pp. 390–403, 2010..

Acknowledgement

THIS WORK IS SUPPORTED BY THE PACIFIC NORTHWEST NATIONAL LABORATORY UNDER THE U.S. DEPARTMENT OF ENERGY CONTRACT DE-AC05-76RL01830