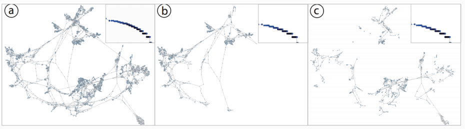
Evaluation of Graph Sampling: A Visualization Perspective

Paper by: Yanhong Wu, Nan Cao, Daniel Archambault, Qiaomu Shen, Huamin Qu, and Weiwei Cui

Presentation by: Austin Wallace.
March 28, 2017

What's better, B or C?



A little different, right?

Similar quantitative statistics

Very different perceptually

Problem: Analyzing large graphs

Large graphs are difficult to analyze even with state of the art techniques on high-end clusters

Can reach hundreds of millions, or even billions of nodes

One Solution: Graph sampling

Sampled graph often more desirable than small chunk of original graph

Makes analysis on large graphs tractable

Can be used for preliminary evaluation

One more problem: How to sample?

What is the best way to sample?

Should we pick nodes at random?

Traverse the graph?

Lots of solutions!

This paper focusses on five of the most widely used:

Random Node (RN)

Random Edge Node (REN)

Random Walk (RW)

Random Jump (RJ)

What? Why? How?

What:

Node-link unweighted networks (N: ~1000-20000)

Why:

Summarize topology

How:

RN, REN, RW, RJ, FF

Key Question: Perceptual Quality

What are the main factors that affect perceptual quality in a sampled graph?

How are those factors affected by the five sampling strategies?

Important Perceptual Qualities

Three identified:

Important Perceptual Qualities

Three identified

Coverage Area

Important Perceptual Qualities

Three identified

Coverage Area

Cluster Quality

Important Perceptual Qualities

Three identified

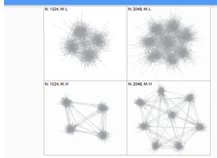
Coverage Area

Cluster Quality

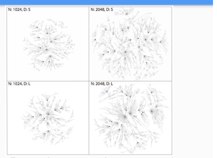
High Degree Nodes, and their preservation

In addition, 20% sampling rate was selected as a fair comparison rate

Graphs used: BA and Sah



Guaranteed cluster networks generated by Sah et al.'s model



Power law networks generated by a Barabasi-Albert model

How did they fare: Coverage Area

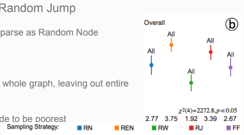
Best: Random Edge Node and Random Jump

Do not get trapped, but are not as sparse as Random Node

Random Walk is poorest

May not explore anywhere near the whole graph, leaving out entire sections

Researchers expected Random Node to be poorest



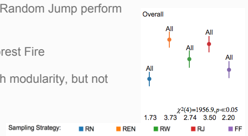
Forest Fire and Random Walk do better in less modular graphs

How did they fare: Cluster Quality

Best: Random Edge Node and Random Jump perform best

Poorest: Random Node and Forest Fire

Random Walk depends on graph modularity, but not graph size



How did they fare: High Degree Nodes

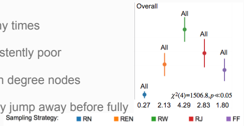
Best: Random Walk

- Can visit the same node many times

Poorest: Random node is consistently poor

- Not at all biased towards high degree nodes

Random jump does well, but may jump away before fully exploring a high degree node



Random Edge Nodes is biased towards high degree

So, which is best?

Random Walk to preserve high-degree nodes

Random Jump or Random Edge Node to preserve global structure and cluster quality

Almost never use Random Node

Strengths

Substantial thought given to experiment design and neutralizing potential confounds

Depth of work: Pilot study, three formal studies

Useful, well explained, and nuanced recommendations

Weaknesses and limitations

Does not explore the laying out of graphs post-sampling.

Only used computer science students/graduates in their studies

Single sampling rate was tested

Potential future work

Improve metrics based on human feedback

Perceptual quality of graph abstraction, as opposed to sampling

Investigate time to complete tasks on sampled graphs, as well as accuracy

Investigate false positives, such as a sampled low degree