Graph Sampling for Visual Analytics

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

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

EI 2017, 29 January - 2 February, Burlingame, California USA

February 27, 2019

MISSISSIPPI STATE
UNIVERSITY

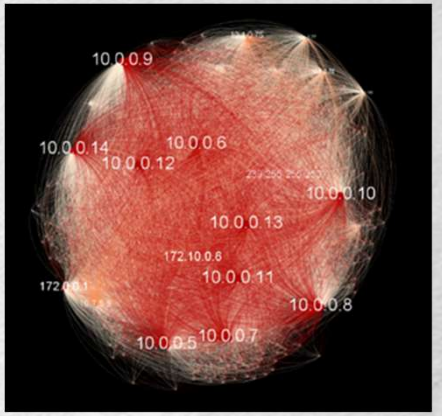
Outline

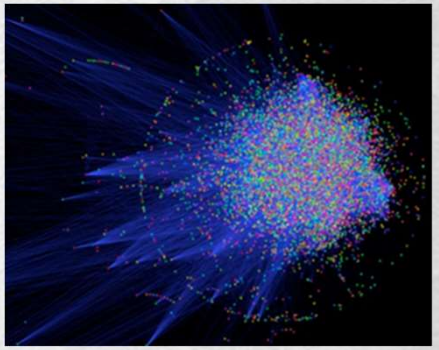
- Introduction
- Evaluation
- Sampling Methods
- Graph Datasets
- Statistical and Visual Comparison
- Analysis
- Conclusion & Discussion
- Contributions
- Ongoing work

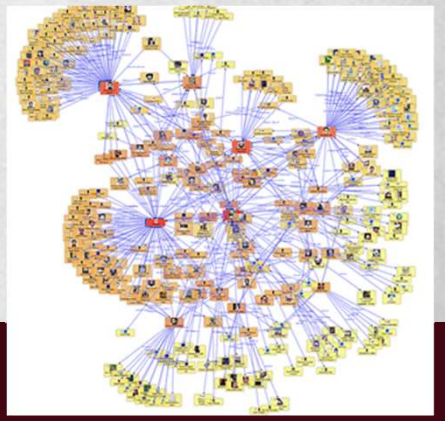
Introduction (1)

Motivations

- Visualization
  - To display all nodes and edges is impossible
- Estimation or calculation
  - To calculate graph properties on a large graph is costly
- Data
  - No complete data
  - To obtain data time-consuming







Introduction (2)

Problems

- Given a huge graph, how to get a representative sample?
- Given several sampling methods, which sampling method is best?
- How to compare those sampling results?
- How do we measure success?

Evaluation (1)

- Skew divergences reflects the average difference between two probability density distributions
- KL Divergence = $\sum(P * log\frac{P}{Q})$
- To smooth the two PDFs
- $SD(P, Q, \alpha) = KL[\alpha P + (1 - \alpha)Q || \alpha Q + (1 - \alpha)P]$, where α is 0.99.

Evaluation (2)

Graph properties distribution (8): Directed Graph

- Degree(DD)
- Average neighbor degree(ANDD)
- Degree centrality(DCD)
- Node betweenness centrality(NBCD)
- Edge betweenness centrality(EBCD)
- Local clustering coefficient(LCCD)
- Closeness centrality(CCD)
- Eigenvector centrality(EVCD)

Graph properties(9): Undirected Graph

- In-degree(InDD)
- Out-degree(OutDD)
- In degree centrality(InCD)
- Out degree centrality(OutCD)
- Average neighbor degree(ANDD)
- Node betweenness centrality(NBCD)
- Edge betweenness centrality(EBCD)
- Closeness centrality(CCD)
- Eigenvector centrality(EVCD)

Sampling Methods

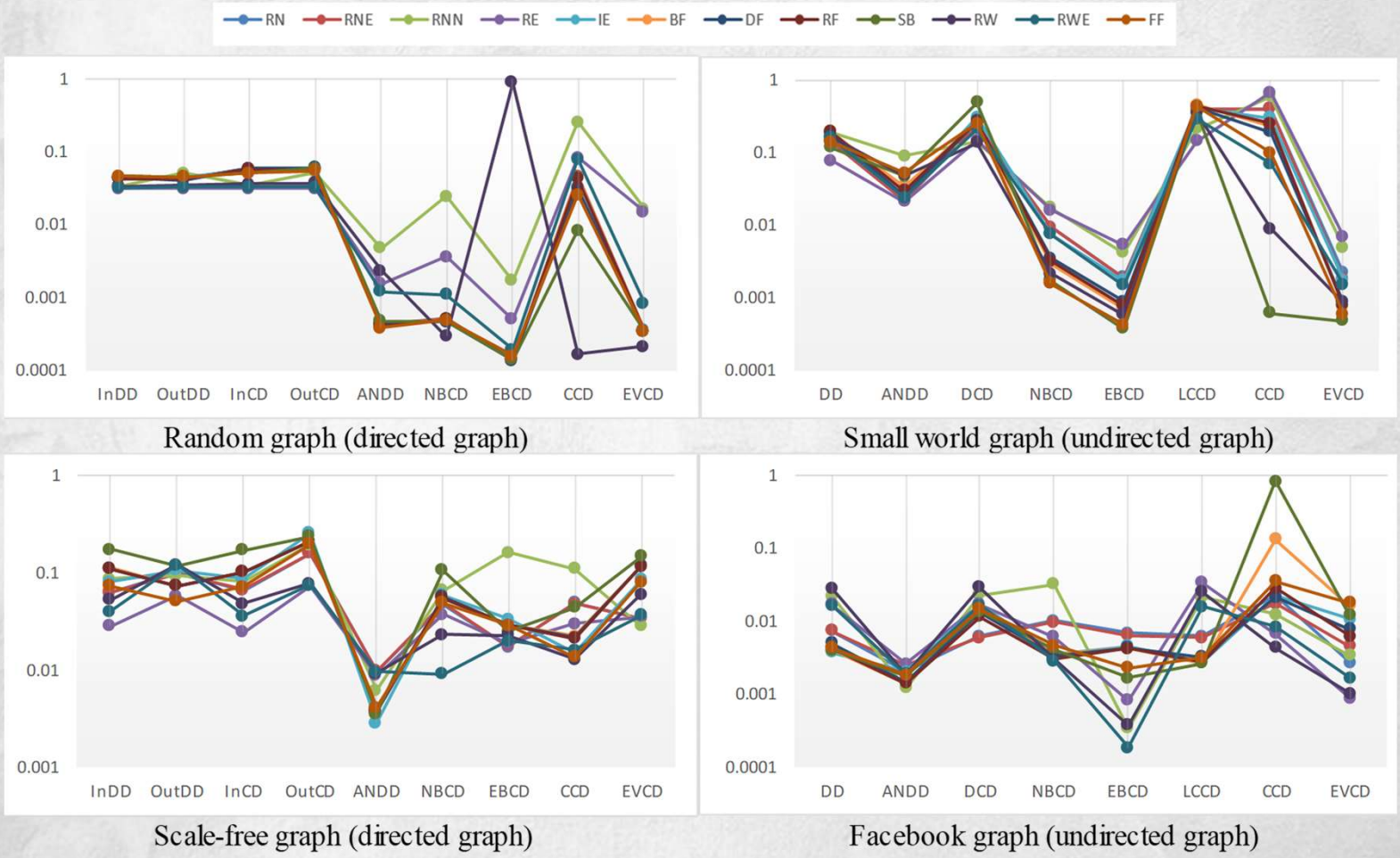
- Node Sampling
  - Random node (RN)
  - Random node-edge (RNE)
  - Random node-neighbor (RNN)
- Edge Sampling
  - Random edge (RE)
  - Induced edge (IE)
- Topology Based Sampling
  - Breadth-first (BF)
  - Depth-first (DF)
  - Random first (RF)
  - Snowball (SB)
  - Random walk (RW)
  - Random walk with escape (RWE)
  - Forest fire (FF)

Graph Datasets

Dataset	Graph Type	Model	# Vertices	# Edges
Random	Directed	Model	10,000	100,246
Small-World	Undirected	Model	10,000	21,895
Scale-Free	Directed	Model	10,000	18,838
Email	Directed	Real	265,214	420,045
Citation	Directed	Real	34,546	421,578
Internet	Directed	Real	10,876	39,994
Facebook	Undirected	Real	4,039	88,234
U.S. Flight	Undirected	Real	235	1,297

SNAP: https://snap.stanford.edu/data/

Statistical Comparisons (1)



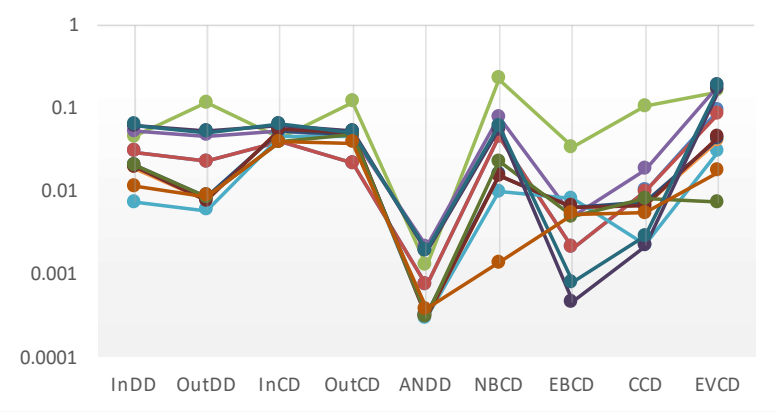
Random graph (directed graph)

Small world graph (undirected graph)

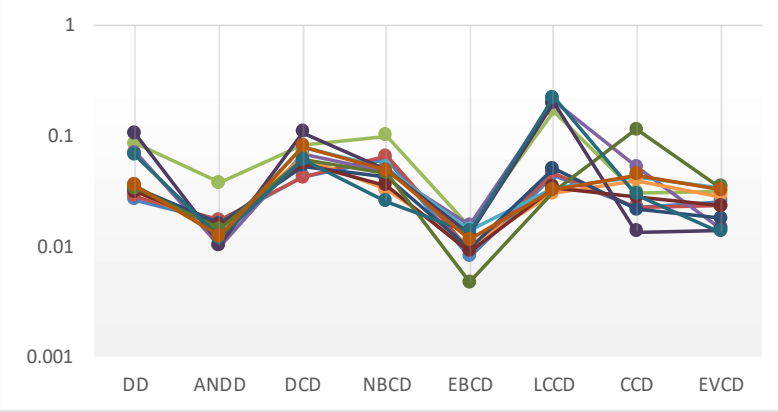
Scale-free graph (directed graph)

Facebook graph (undirected graph)

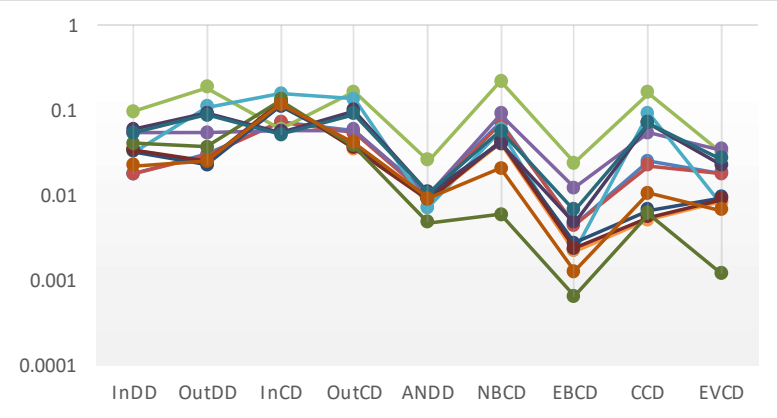
Statistical Comparisons Results (2)



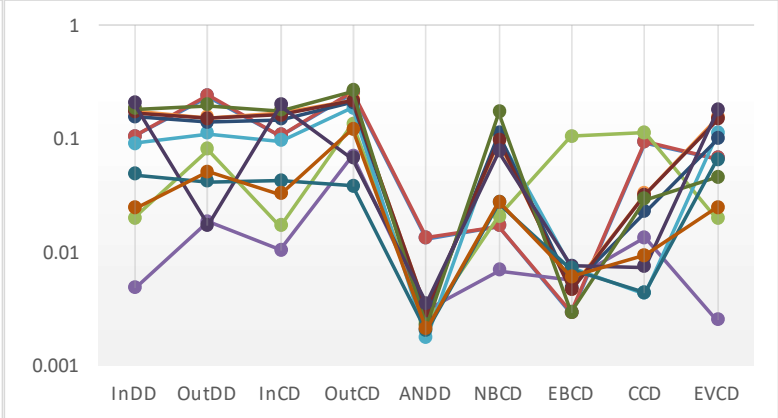
Citation graph (directed graph)



U.S. flight graph (undirected graph)

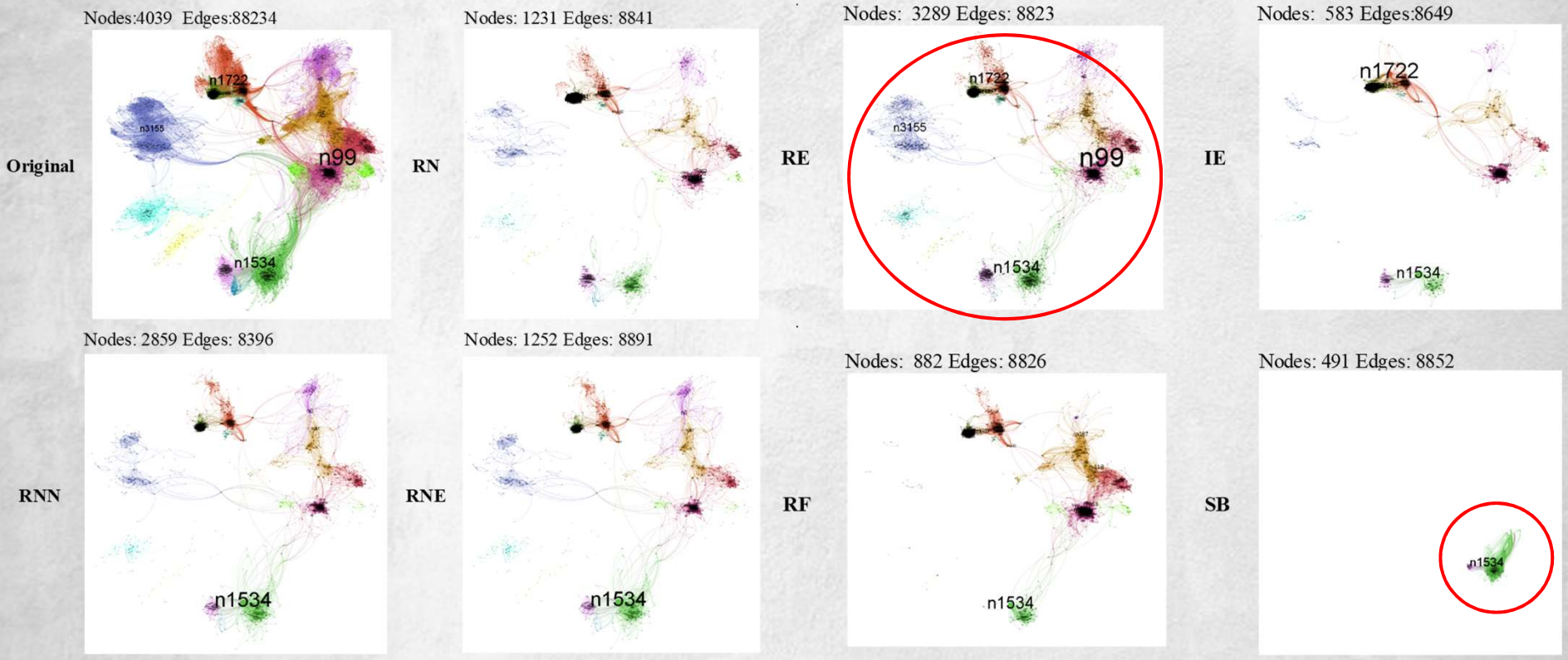


Internet graph (directed graph)



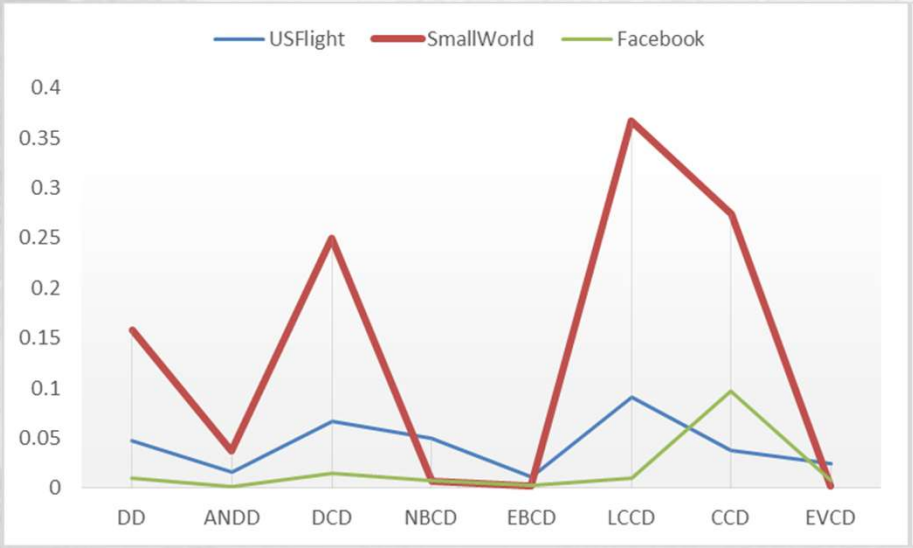
Email graph (directed graph)

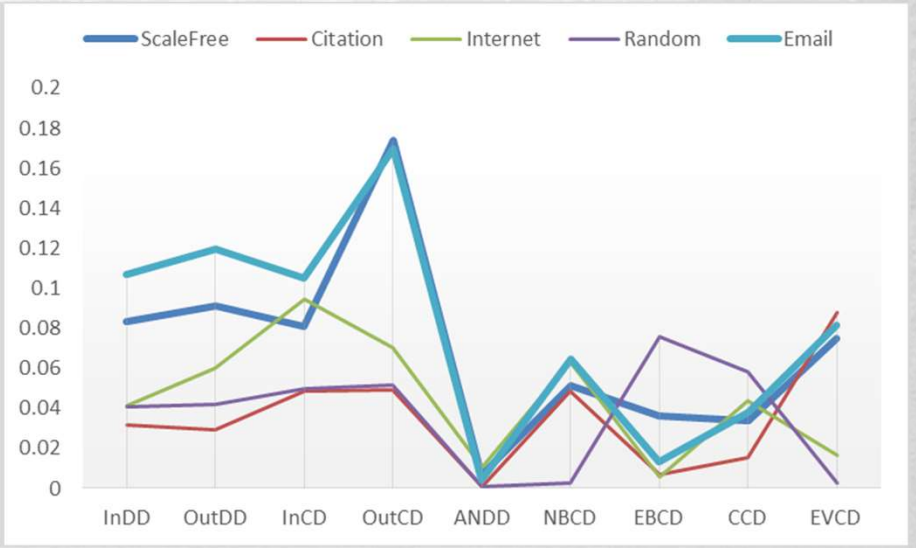
Visual Comparison Results



Facebook graph; Sampling rate: 10 % on edges

Analysis: Comparison between Graph Types



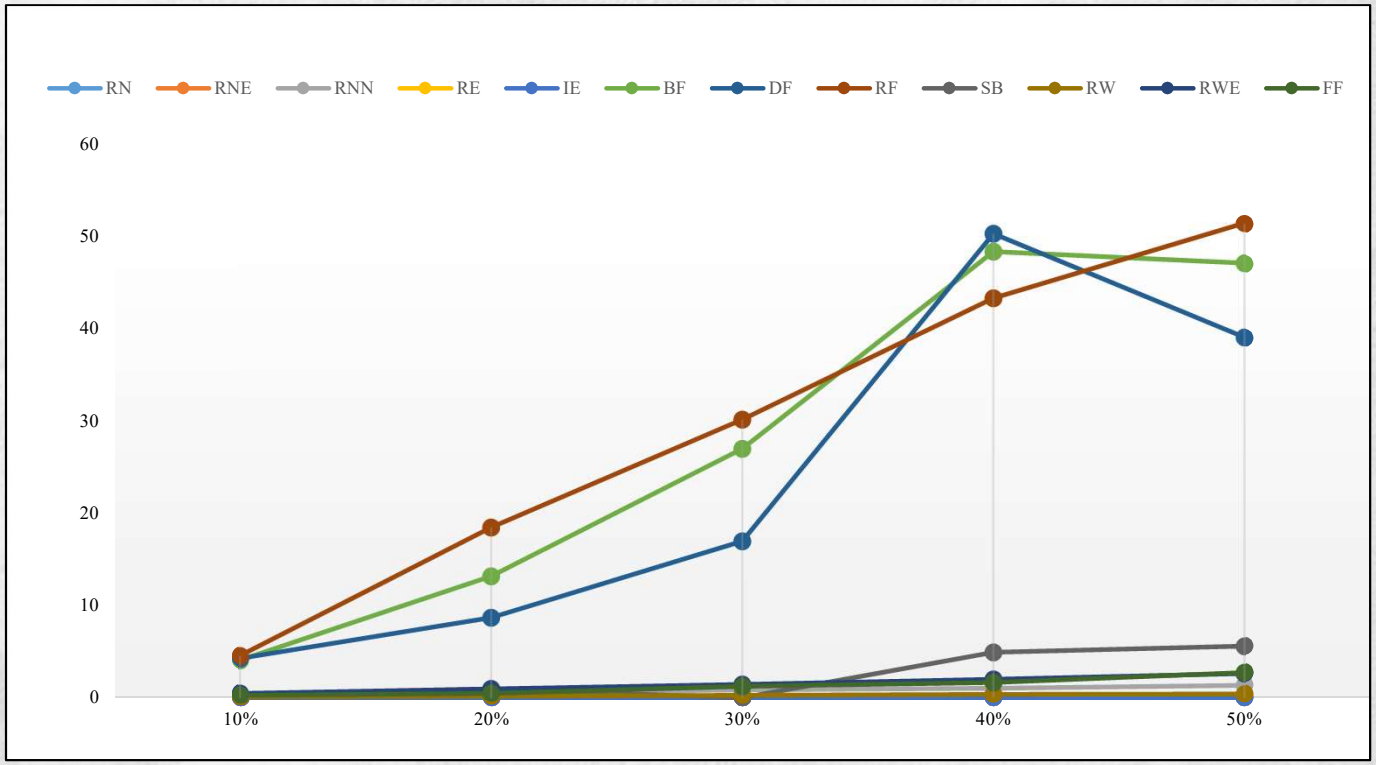


- Graph type has significant influence on sampling results.
- Graph type should be considered in sampling.

Analysis: Comparison between Graph Properties

- Only a few sampling methods act consistently well on certain graph properties across graph type
- For a certain graph data, some methods preserve certain graph properties very well

Analysis: Comparison in Efficiency



Facebook

- Random sampling methods are not sensitive to sampling rate
- Traversal-based sampling methods, the execution time grows rapidly with the sampling rate increasing

Analysis: Visual Comparison

- Spatial coverage
  - Random sampling methods have better spatial coverage than traversal-based sampling, in particular for a small sampling rate
- Clusters
  - Edge-related sampling methods (e.g., random edge) are better than node sampling and traversal-based sampling when the sampling rate is small

Conclusion & Discussion

- Sampling factors
  - graph type
  - graph property
  - sampling efficiency
  - visual requirements

Undirected Graph

Small World Graph

DD	ANDD	DCD	NBCD	EBCD	LCCD	CCD	EVCD
RE	RE	RW	FF	SB	RE	SB	SB
SB	RN	RNN	SB	FF	RNN	RW	FF
FF	RNE	RE	RW	RW	RWE	RWE	BF

US Flight graph

DD	ANDD	DCD	NBCD	EBCD	LCCD	CCD	EVCD
RN	RE	RNE	RWE	SB	BF	RW	RWE
RNE	RW	RN	BF	RN	SB	DF	RW
RF	RWE	DF	RF	RF	FF	RNE	RE

Facebook Graph

DD	ANDD	DCD	NBCD	EBCD	LCCD	CCD	EVCD
IE	RNN	RNE	RWE	RWE	BF	RW	RE
SB	RF	RN	RW	RNN	SB	RE	RW
BF	BF	RF	RF	RW	IE	RWE	RWE

Directed Graph

Scale-Free Graph

InDD	OutDD	InCD	OutCD	ANDD	NBCD	EBCD	CCD	EVCD
RE	FF	RE	RE	IE	RWE	RN	RW	RNN
RWE	RE	RWE	RWE	SB	RW	RNE	FF	RN
RW	RF	RW	RW	DF	RE	RE	IE	RNE

Email Graph

InDD	OutDD	InCD	OutCD	ANDD	NBCD	EBCD	CCD	EVCD
RE	RW	RE	RWE	IE	RE	RN	IE	RE
RNN	RE	RNN	RW	RWE	RN	RNE	RWE	RNN
FF	RWE	FF	RE	FF	RNE	SB	RW	FF

Citation Graph

InDD	OutDD	InCD	OutCD	ANDD	NBCD	EBCD	CCD	EVCD
IE	IE	FF	RNE	IE	FF	RW	RW	SB
FF	RF	SB	RN	RF	IE	RWE	IE	FF
BF	BF	RN	FF	BF	RF	RNE	RWE	IE

Random Graph

InDD	OutDD	InCD	OutCD	ANDD	NBCD	EBCD	CCD	EVCD
RE	RE	RE	RE	FF	RW	DF	RW	RW
RWE	RWE	RW	RWE	IE	DF	SB	SB	DF
RNN	RW	RNN	RW	RF	SB	BF	FF	BF

Internet Graph

InDD	OutDD	InCD	OutCD	ANDD	NBCD	EBCD	CCD	EVCD
RNE	DF	RWE	BF	SB	SB	SB	BF	SB
RN	BF	RW	RF	IE	FF	FF	RF	FF
FF	RF	RE	SB	RF	BF	BF	SB	IE

Contributions

- Studied a number of sampling methods and graph data
- Evaluated graph sampling methods with both visual and statistical properties, built a benchmark
- Suggested to choose proper sampling methods in application

Ongoing Work

- If a huge graph that fits into disk but not main memory,
  - How to make sampling in reasonable time?
  - How to use traversal-based sampling methods in short time?
  - How can we speed up the computation?
- Solution
  - Distributed computation framework
  - Developed a sampling package on spark

Questions?

Thanks!

Presenter: Fangyan Zhang
Email: fz56@msstate.edu

Acknowledgment

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

MISSISSIPPI STATE UNIVERSITY