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]

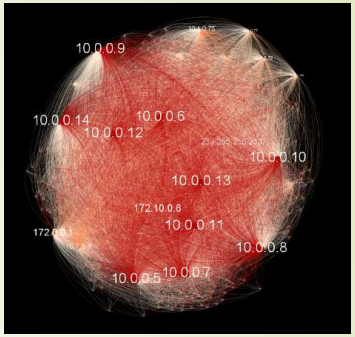
[1]Mississippi State University [2]Pacific Northwest National Laboratory

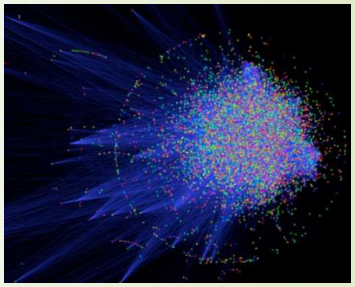
Outline

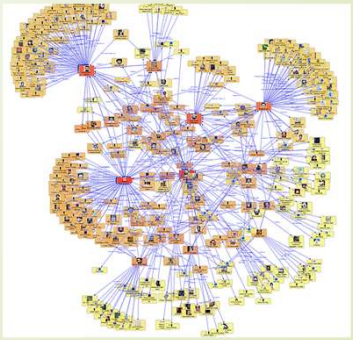
- **Introduction**
  - Motivation
  - Problems
- **Evaluation**
  - KS-Distance
  - Graph Type and Properties
- **Sampling Methods**
  - Node sampling
  - Edge sampling
  - Topology Based Sampling
- **Comparison of Sampling Results**
  - Statistical Comparison
  - Visual Comparison
  - Efficiency Comparison
- **Sampling on Large Graph**
  - Node Sampling
  - Out Degree Distribution
- **Conclusion**

Introduction

- **Motivation**
  - **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 is time-consuming.







Introduction

- **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?

Outline

- **Introduction**
  - Motivation
  - Problems
- **Evaluation**
  - KS-Distance
  - Graph Type and Properties
- **Sampling Methods**
  - Node sampling
  - Edge sampling
  - Topology Based Sampling
- **Comparison of Sampling Results**
  - Statistical Comparison
  - Visual Comparison
  - Efficiency Comparison
- **Sampling on Large Graph**
  - Node Sampling
  - Out Degree Distribution
- **Conclusion**

Evaluation

- **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
  - To calculate KS value on graph properties.

Evaluation

Graph properties(10): Directed Graph

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

Graph properties(7): 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)

Outline

- **Introduction**
  - Motivation
  - Problems
- **Evaluation**
  - KS-Distance
  - Graph Type and Properties
- **Sampling Methods**
  - Node sampling
  - Edge sampling
  - Topology Based Sampling
- **Comparison of Sampling Results**
  - Statistical Comparison
  - Visual Comparison
  - Efficiency Comparison
- **Sampling on Large Graph**
  - Node Sampling
  - Out Degree Distribution
- **Conclusion**

Sampling algorithms

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

Outline

- **Introduction**
  - Motivation
  - Problems
- **Evaluation**
  - KS-Distance
  - Graph Type and Properties
- **Sampling Methods**
  - Node sampling
  - Edge sampling
  - Topology Based Sampling
- **Comparison of Sampling Results**
  - Statistical Comparison
  - Visual Comparison
  - Efficiency Comparison
- **Sampling on Large Graph**
  - Node Sampling
  - Out Degree Distribution
- **Conclusion**

Comparison of Sampling Results

Undirected Graph

- American Airlines connection data
  - 235 nodes 1297 edges
- Simulated random undirected Graph
  - Graph 1: 500 nodes 1260 edges
  - Graph 2: 500 nodes 1238 edges
  - Graph 3: 750 nodes 2871 edges
  - Graph 4: 1000 nodes 4989 edges
  - Graph 5: 1000 nodes 5092 edges
  - Graph 6: 1250 nodes 7884 edges

Directed Graph

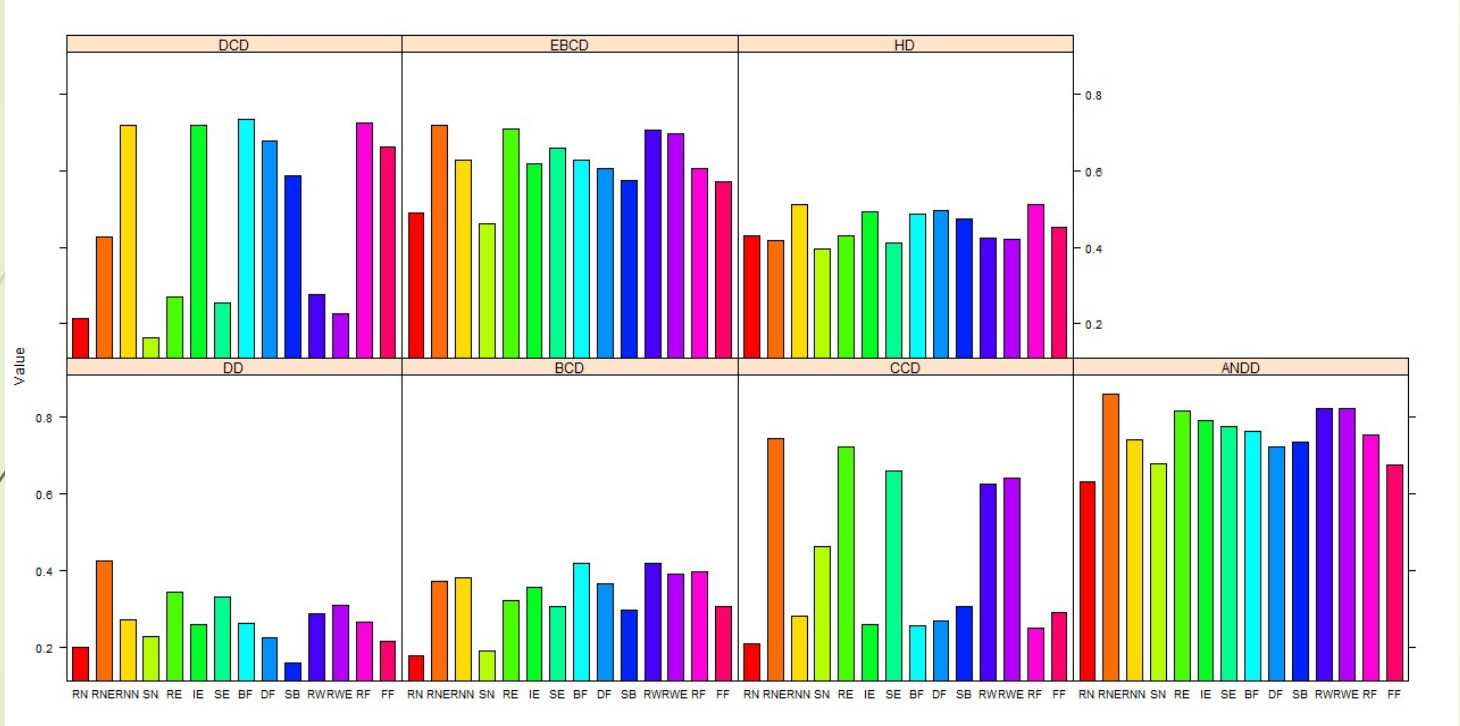
- VAST
  - 1214 nodes 15653 edges
- Simulated random directed Graph
  - Graph 1: 500 nodes 1284 edges
  - Graph 2: 500 nodes 1262 edges
  - Graph 3: 750 nodes 2859 edges
  - Graph 4: 1000 nodes 4900 edges
  - Graph 5: 1000 nodes 4954 edges
  - Graph 6: 1250 nodes 7732 edges

Simulated Graph: Random Graph created by Erdős–Rényi model

Statistical Comparison of Sampling Results

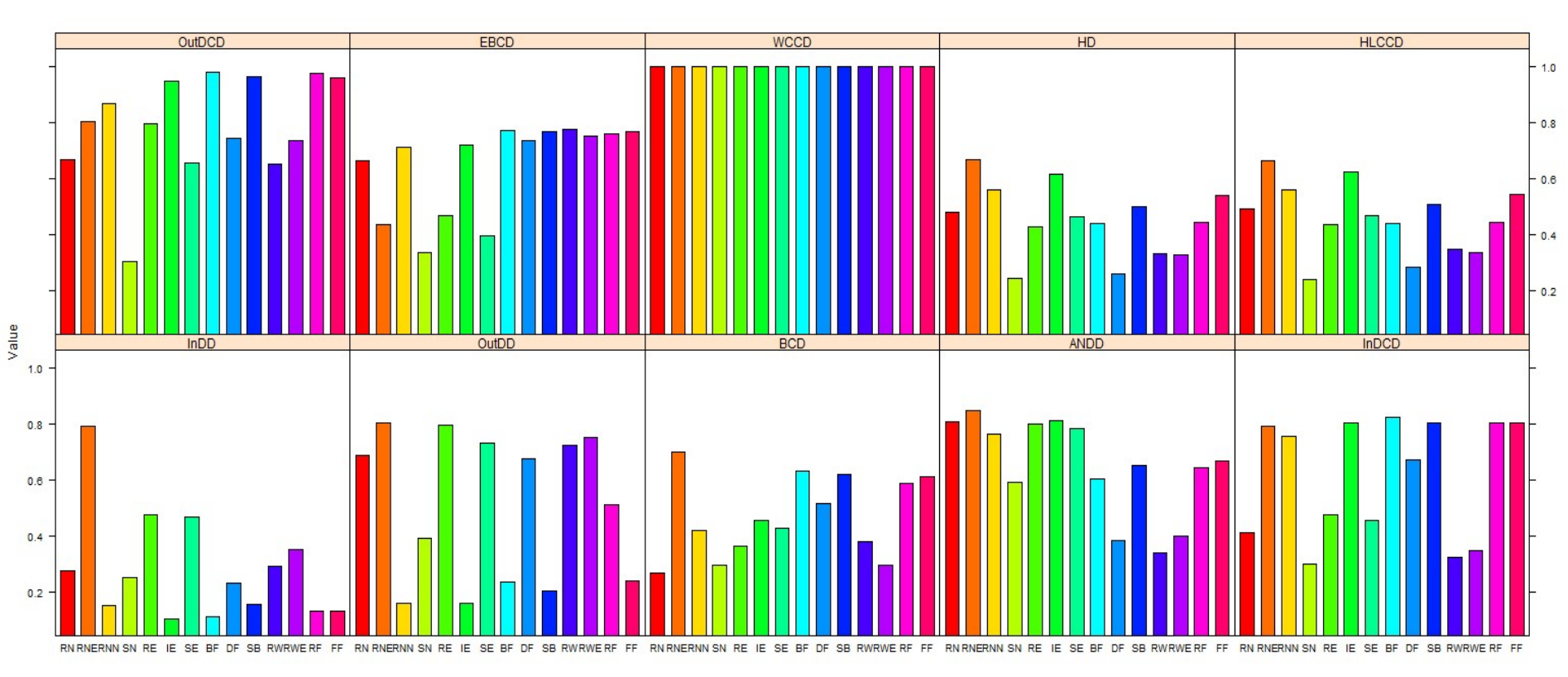
- **Comparison Categories (4):**
  - Between directed and undirected Graph
    - American airline connection data and VAST data
  - Between different graph type of undirected or directed graphs
    - American airline connection data and Simulated undirected graph
  - Between multiple graphs of same type but different sizes.
    - Simulated data with different number of nodes(500, 750, 1000, 1250)
  - Between multiple graphs of same size and same type
    - Simulated Graph(500 VS 500, 1000 VS 1000)

Category 1: Between directed and undirected Graph (Undirected Graph)



- Data: American Airlines connection data. Sampling rate: average(10%-50%)

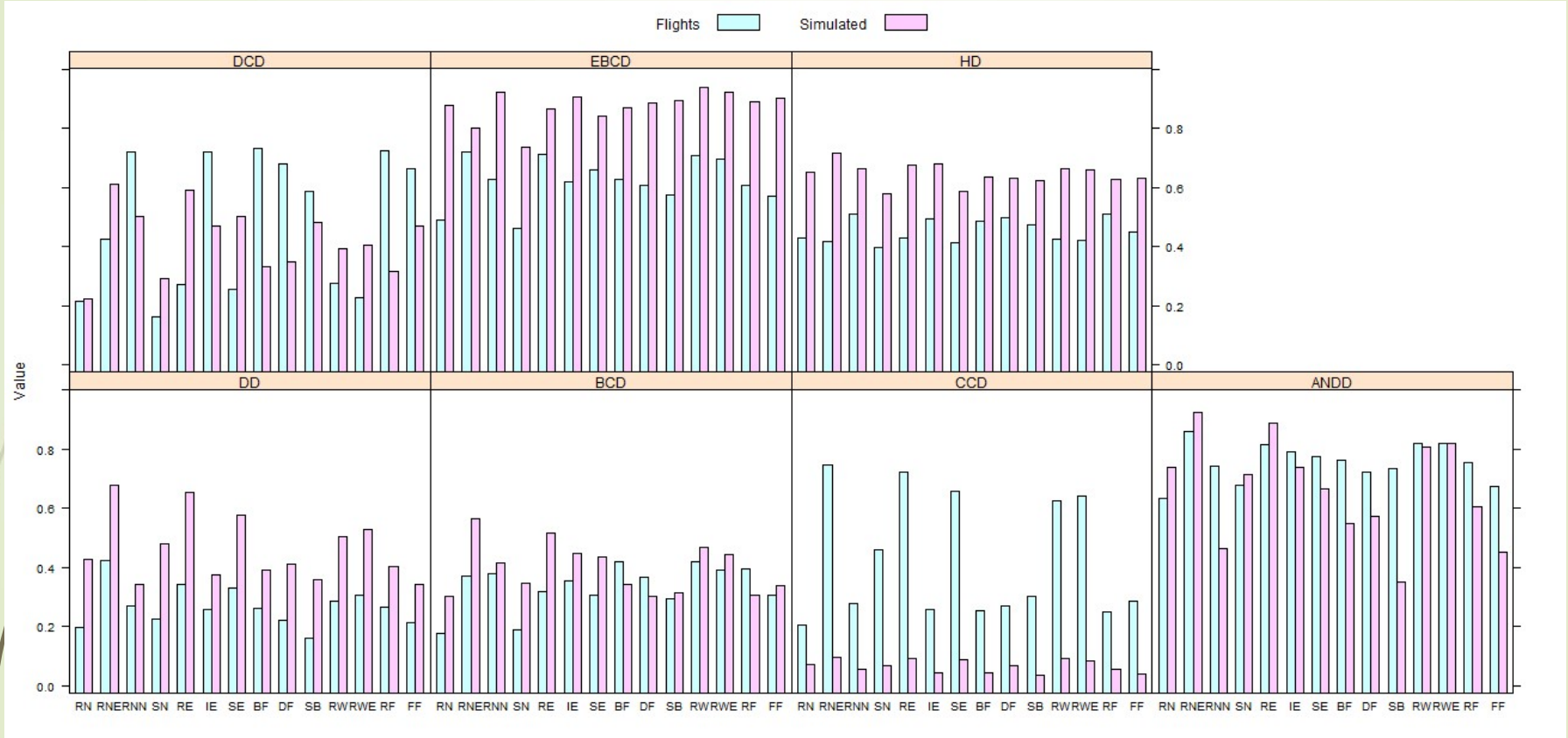
Category 1: Between directed and undirected Graph (directed Graph)



- Data: VAST 2013 Netflow data.

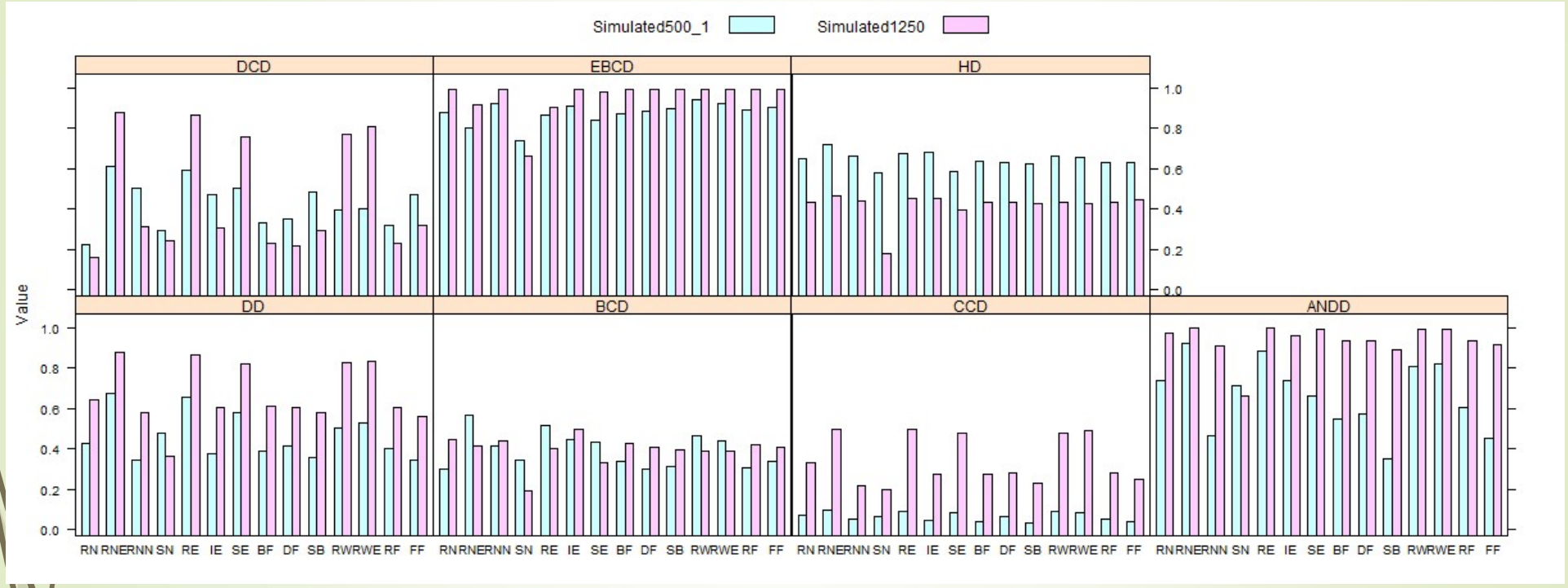
Sampling rate: average(10%-50%)

Category 2: between two undirected graphs with different graph type



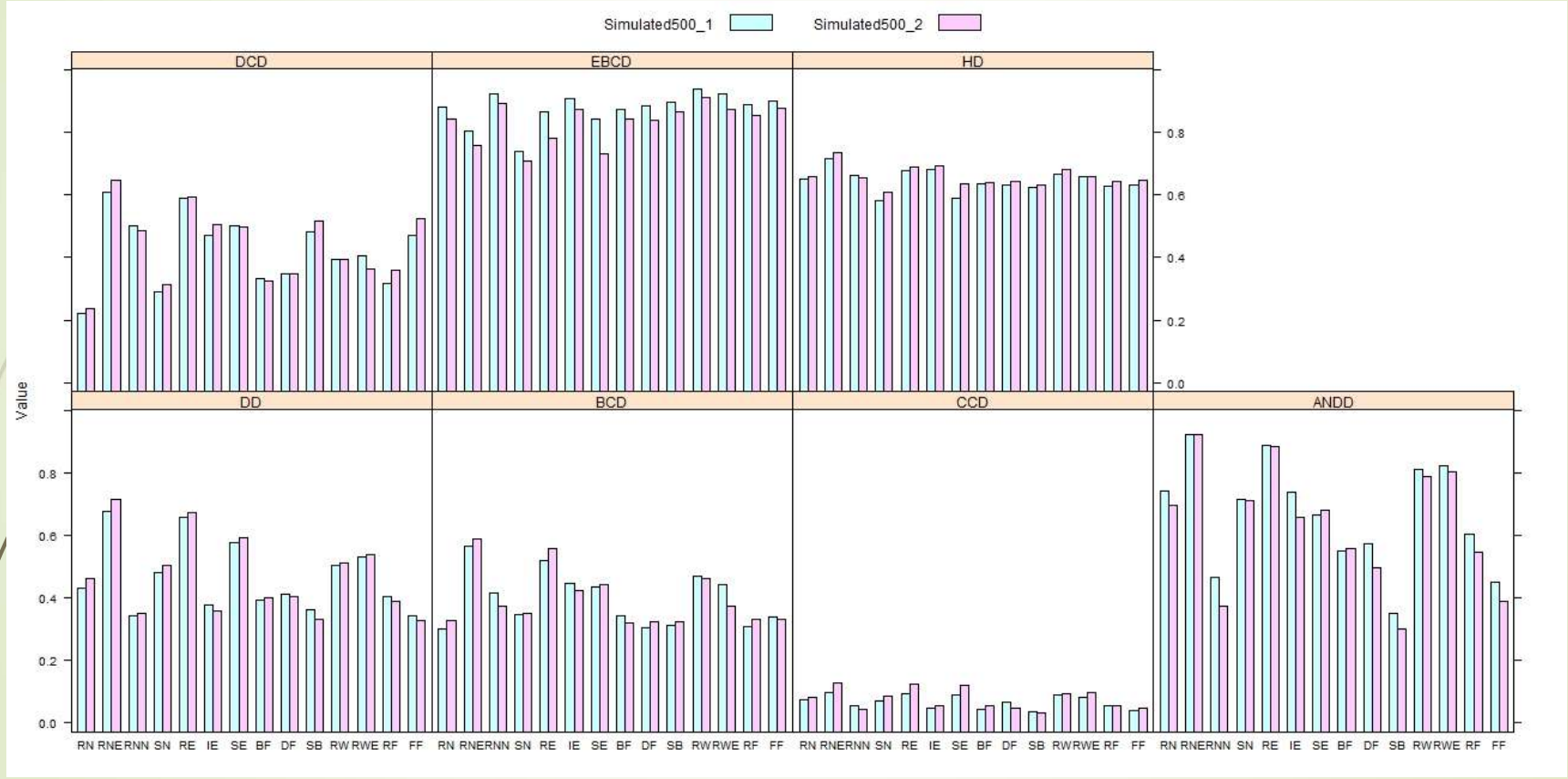
- Data: American Airline connection data VS Simulated 500_1. Sampling rate: average(10%-50%)

Category 3: between multiple graphs of same type but different sizes



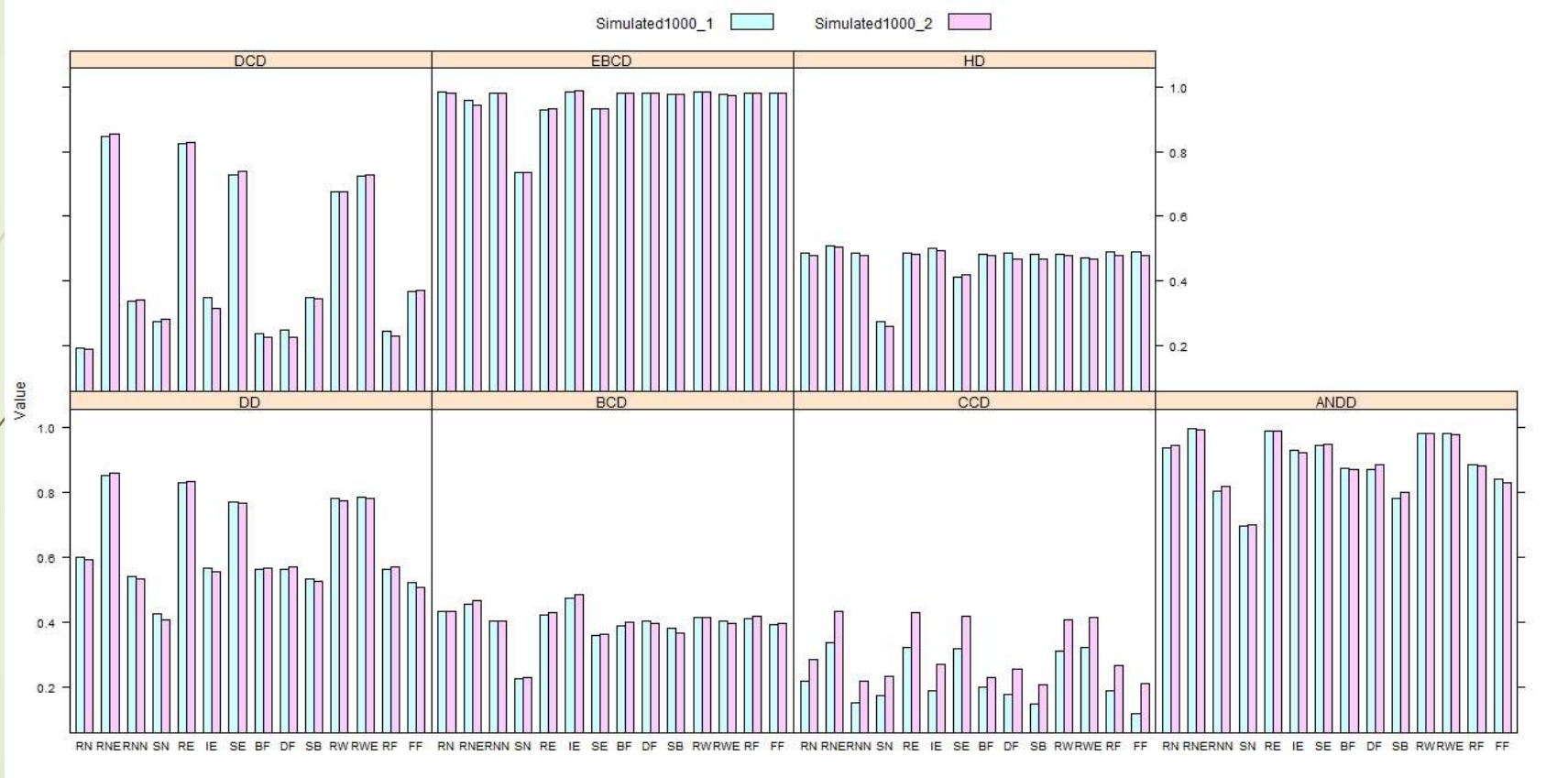
- Data: Simulated data 500_1 VS 1250. Sampling rate: average(10%-50%)

Category 4: between multiple graphs of same size and same type



- Data: Simulated data 500_1 VS 500_2. Sampling rate: average(10%-50%)

Category 4: between multiple graphs of same size and same type

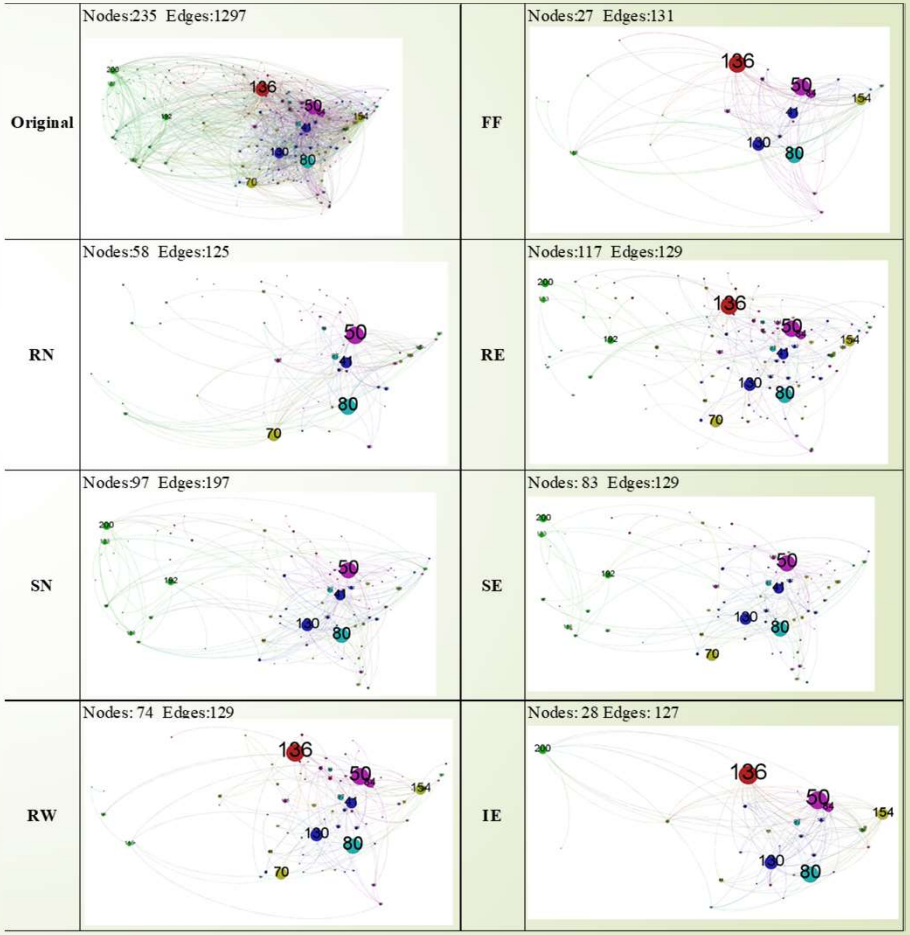


- Data: Simulated data 1000_1 VS 1000_2. Sampling rate: average(10%-50%)

Visual Comparison of Sampling Results

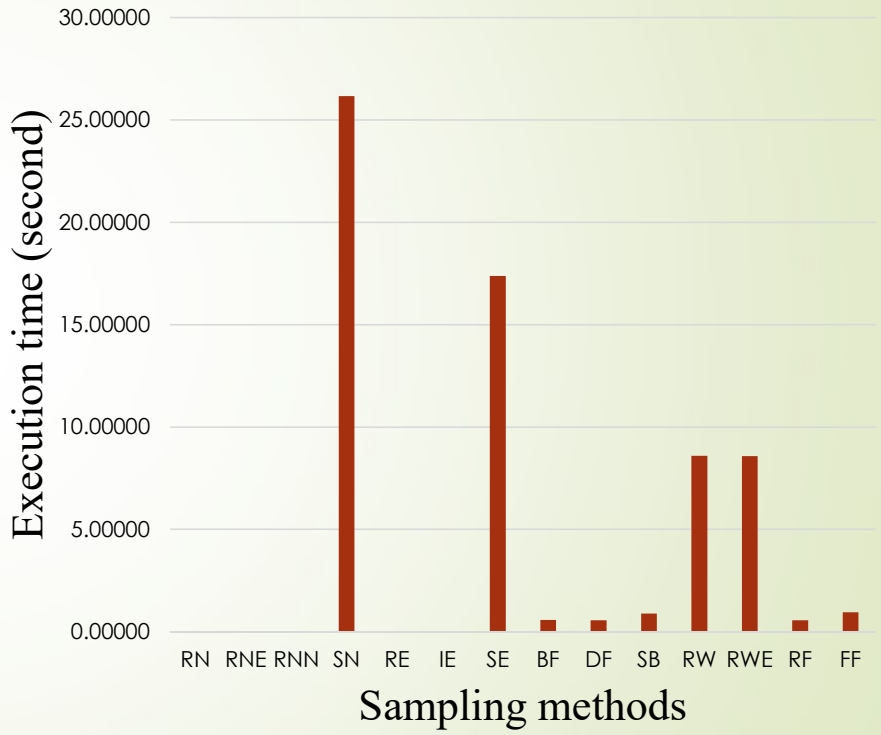
- **Visual Comparison**
  - **Data:** American Airlines connection data
  - **Graph Type:** undirected graph
  - **Sampling rate:** 10% on edges
  - **Visual Comparison Technique**
    - Fix nodes location, label size, color etc.
  - **Analysis**

    Edge-related sampling methods are biased towards high-degree nodes. For example, random edge sampling, induced edge sampling, streaming edge sampling are easy to sample high degree nodes (136,50, 80,130,70 etc.)



Efficiency Comparison

- **Execution Time**
  - **Data**
    - simulated undirected graph data
  - **Sampling rate**
    - Average of sampling result with sampling rate range from 10% to 50% on edges.



Outline

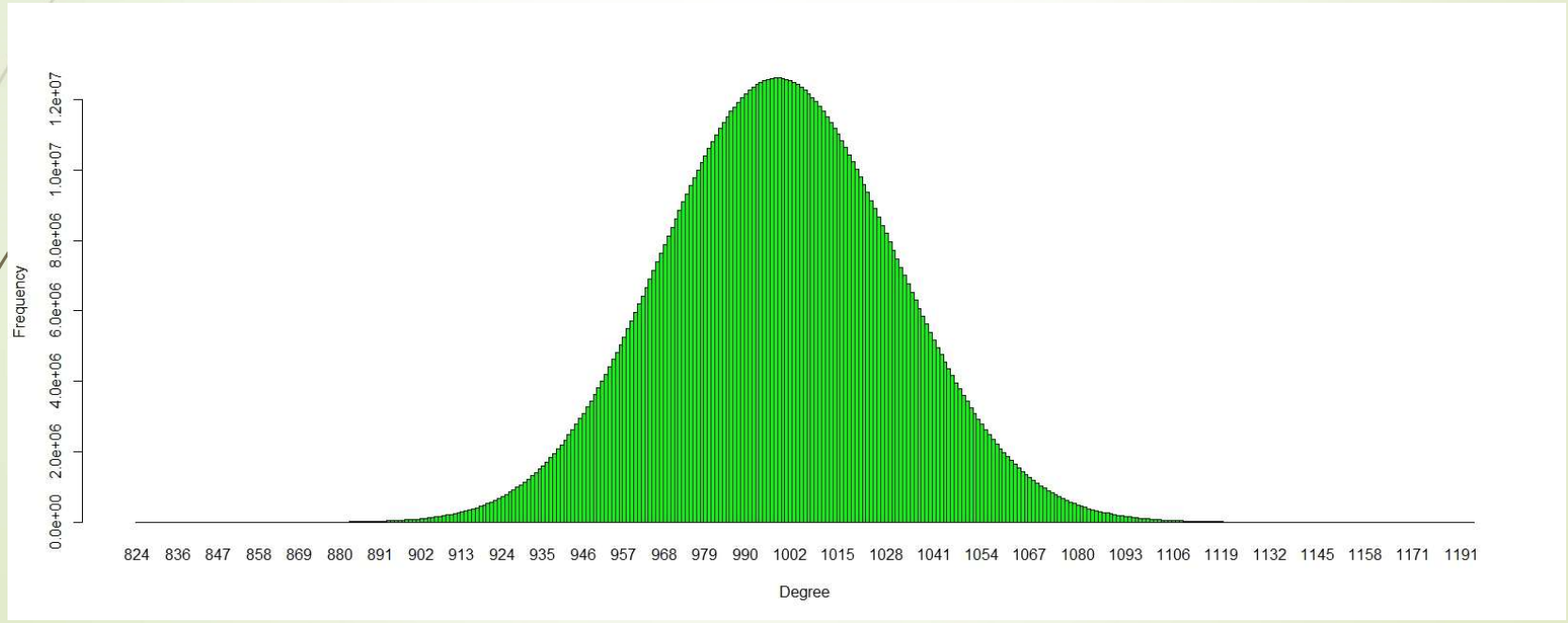
- **Introduction**
  - Motivation
  - Problems
- **Evaluation**
  - KS-Distance
  - Graph Type and Properties
- **Sampling Methods**
  - Node sampling
  - Edge sampling
  - Topology Based Sampling
- **Comparison of Sampling Results**
  - Statistical Comparison
  - Visual Comparison
  - Efficiency Comparison
- **Sampling on Large Graph**
  - Node Sampling
  - Out Degree Distribution
- **Conclusion**

Sampling on Big Graph

- Graph Generation
  - Erdős–Rényi model
  - parallel algorithm (by MPi4py)
  - Shadow II. Used 100 nodes, 2000 processors, each node: 512GB memory.
- Size:
  - 1 billion nodes
  - ~500 billion edges.
- Graph Storage:
  - about 10TB

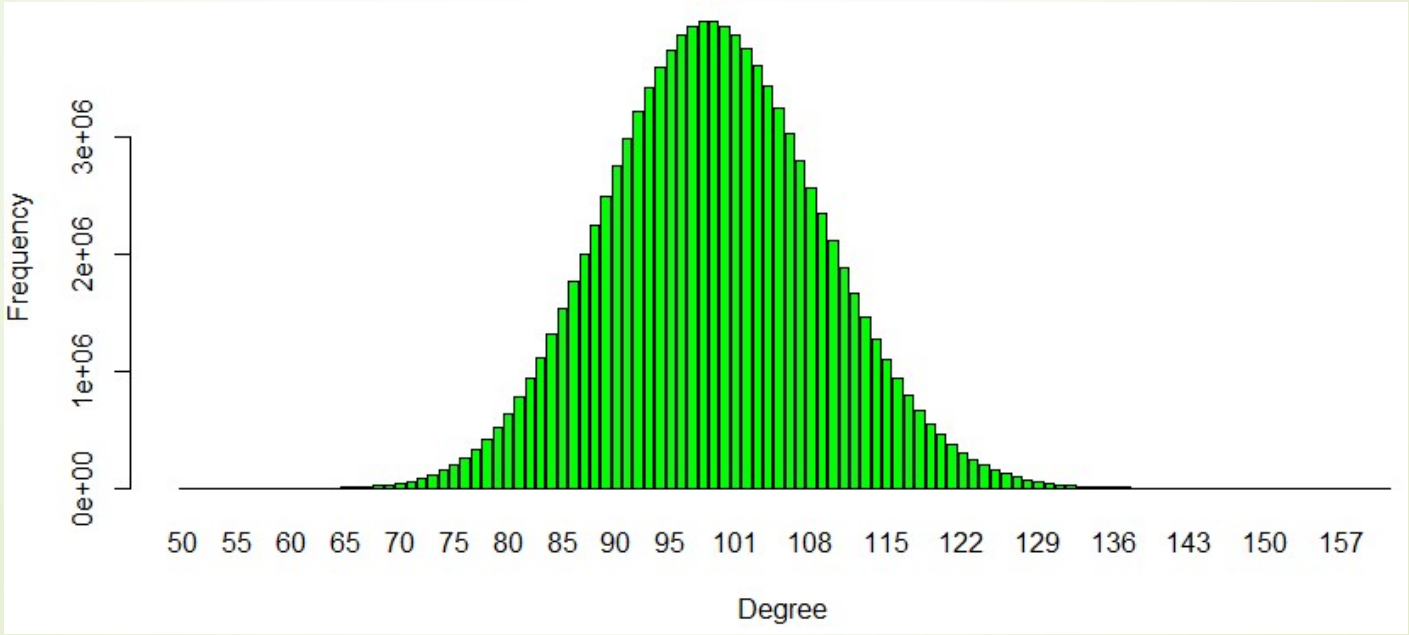
Sampling on Big Graph

- Original Graph
- Out Degree distribution



Sampling on Big Graph

- Node sampling. Sampling Rate: 10 %
- Out Degree Distribution:



Outline

- **Introduction**
  - Motivation
  - Problems
- **Evaluation**
  - KS-Distance
  - Graph Type and Properties
- **Sampling Methods**
  - Node sampling
  - Edge sampling
  - Topology Based Sampling
- **Comparison of Sampling Results**
  - Statistical Comparison
  - Visual Comparison
  - Efficiency Comparison
- **Sampling on Large Graph**
  - Node Sampling
  - Out Degree Distribution
- **Conclusion**

Conclusion

- No sampling method works well for all graphs.
- In visual comparison, the consistent graph layout facilitates comparison.
- The benchmark helps users choose proper sampling methods in applications.
- The benchmark provides an avenue to explore big graph.

Questions?

Thanks!

Acknowledgement

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