



Space-Efficient Sampling from Social Activity Streams

Nesreen K. Ahmed

Purdue University, CS Dept.
nkahmed@cs.purdue.edu

Jennifer Neville

Purdue University, CS Dept.
neville@cs.purdue.edu

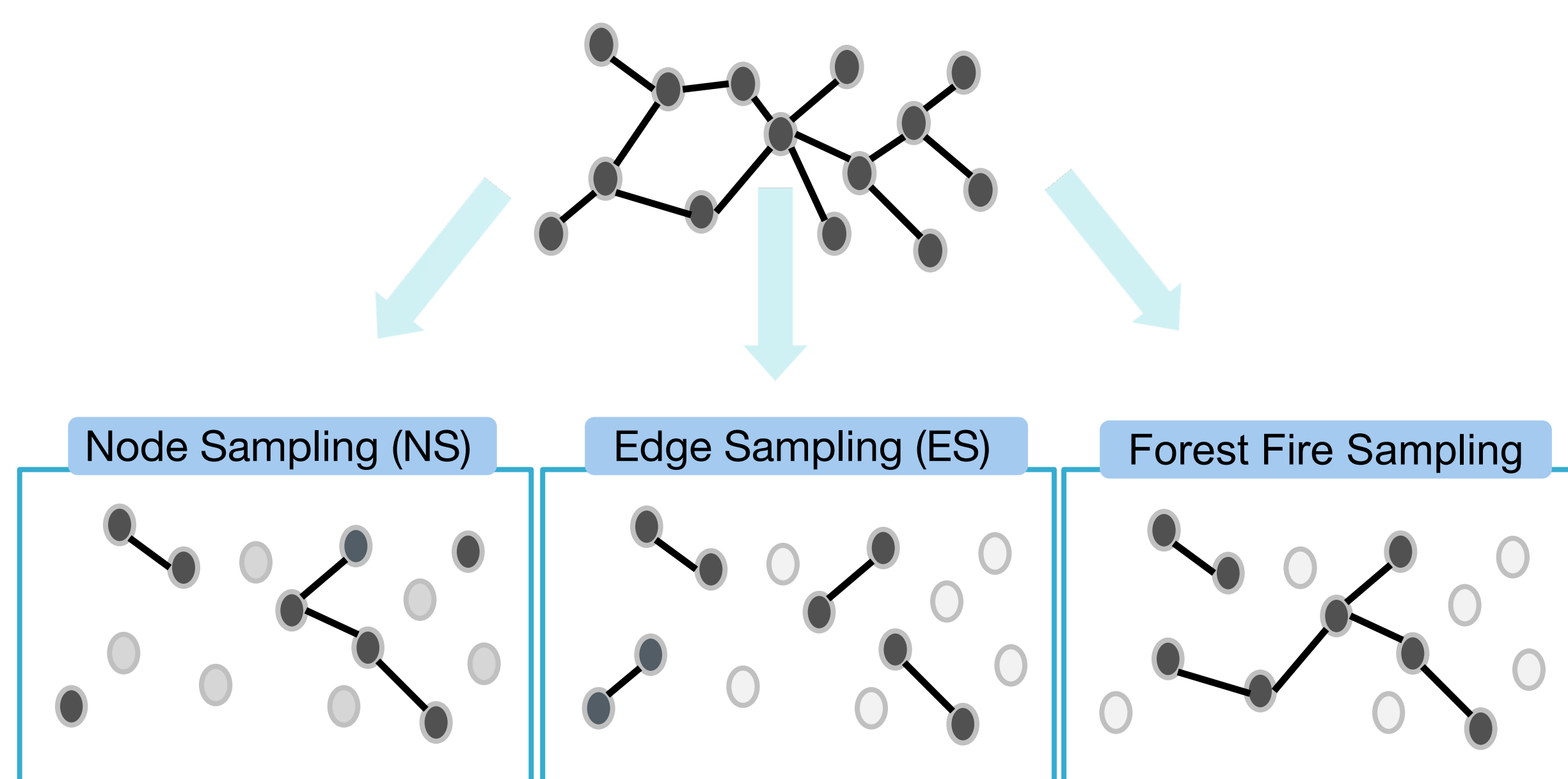
Ramana Kompella

Purdue University, CS Dept.
kompella@cs.purdue.edu

Motivation

- Studying complex networks is a challenging task due to their:
 - Heterogeneous and dependent structure
 - Large size
 - Evolution over the time
- Network Sampling** is used to select a subset of nodes/edges from the full graph
- A sample is representative of the full graph if it preserves characteristics of the graph
 - degree, path length, clustering coefficient
- Network Sampling can be broadly classified as:

(1) Node (2) Edge (3) Topology



Problem Definition

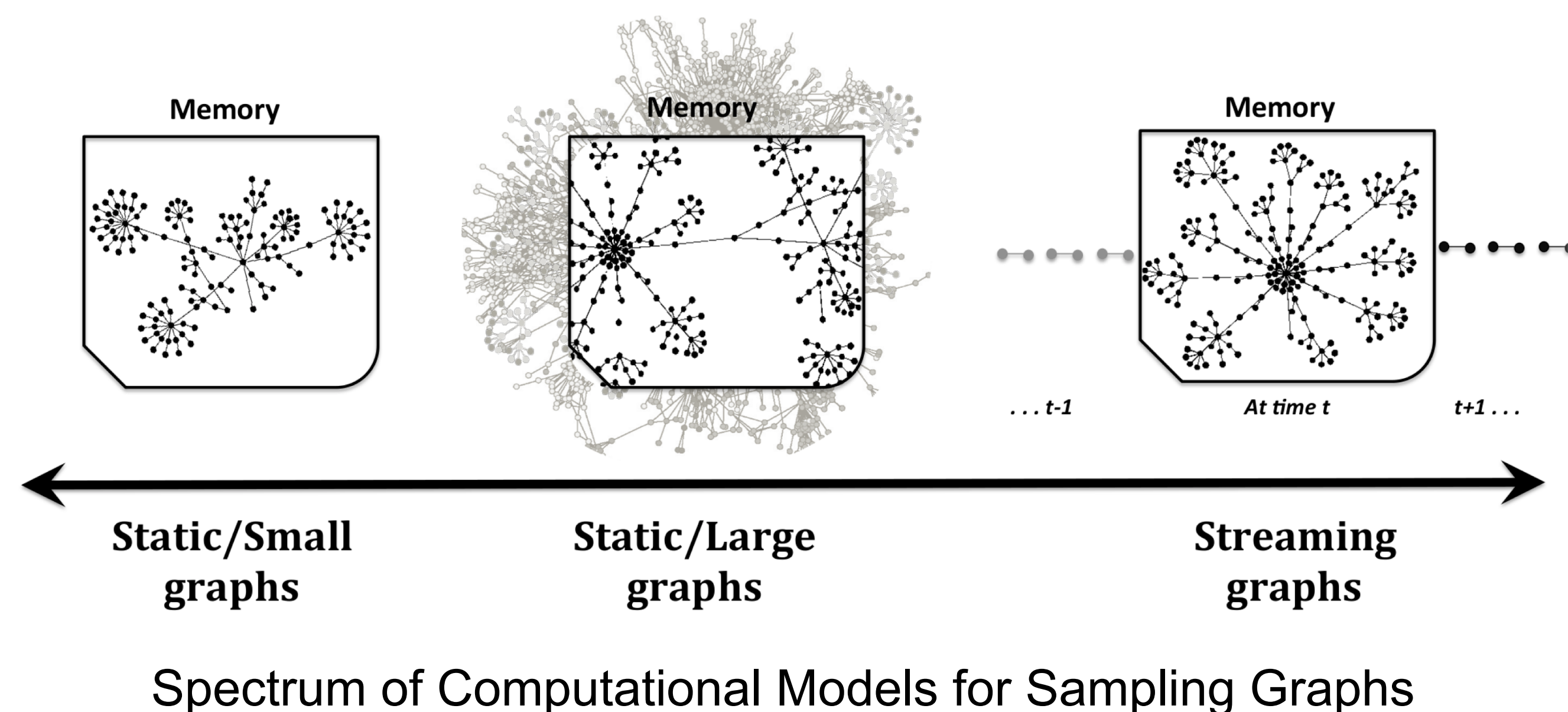
- Current sampling methods require access to either:
 - (a) entire graph
 - (b) node's neighbors
- This requires the graph to be memory-resident
- Big* graphs are too large/dynamic to fit in main memory
- Most *big* graphs evolve as a stream of edges over time
 - e.g. email communications, tweets in twitter hashtags

Goal

- Study how to *sample* from large graph streams

Contributions

- We propose a novel graph **stream sampling PIES**:
 - reservoir-based sampling
 - maintains a dynamic/changing representative sample from large graph streams
 - uses a single pass over the edges $O(|E|)$
- We propose streaming implementations for current sampling methods
 - Node, Edge, and Breadth First sampling

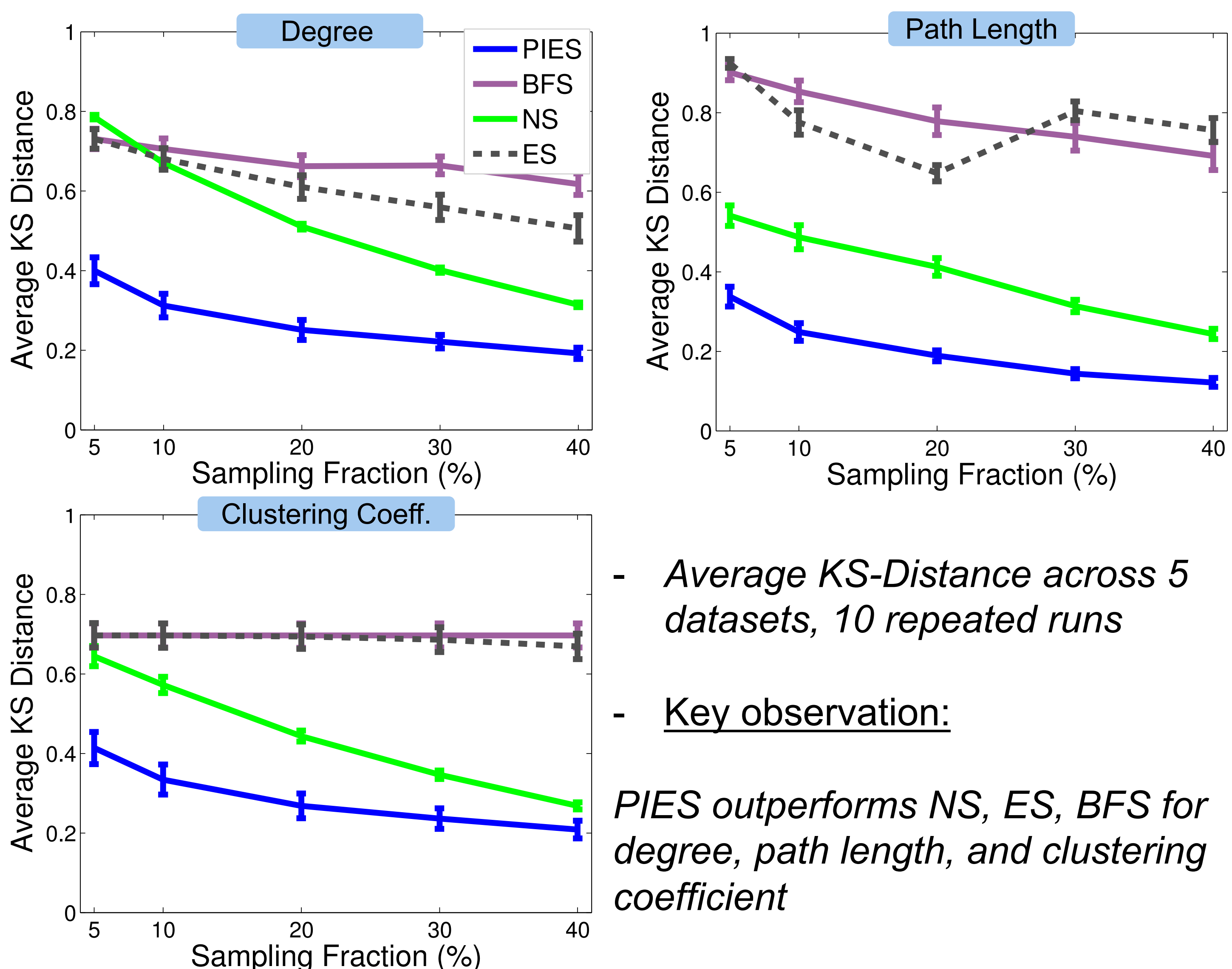


Algorithms

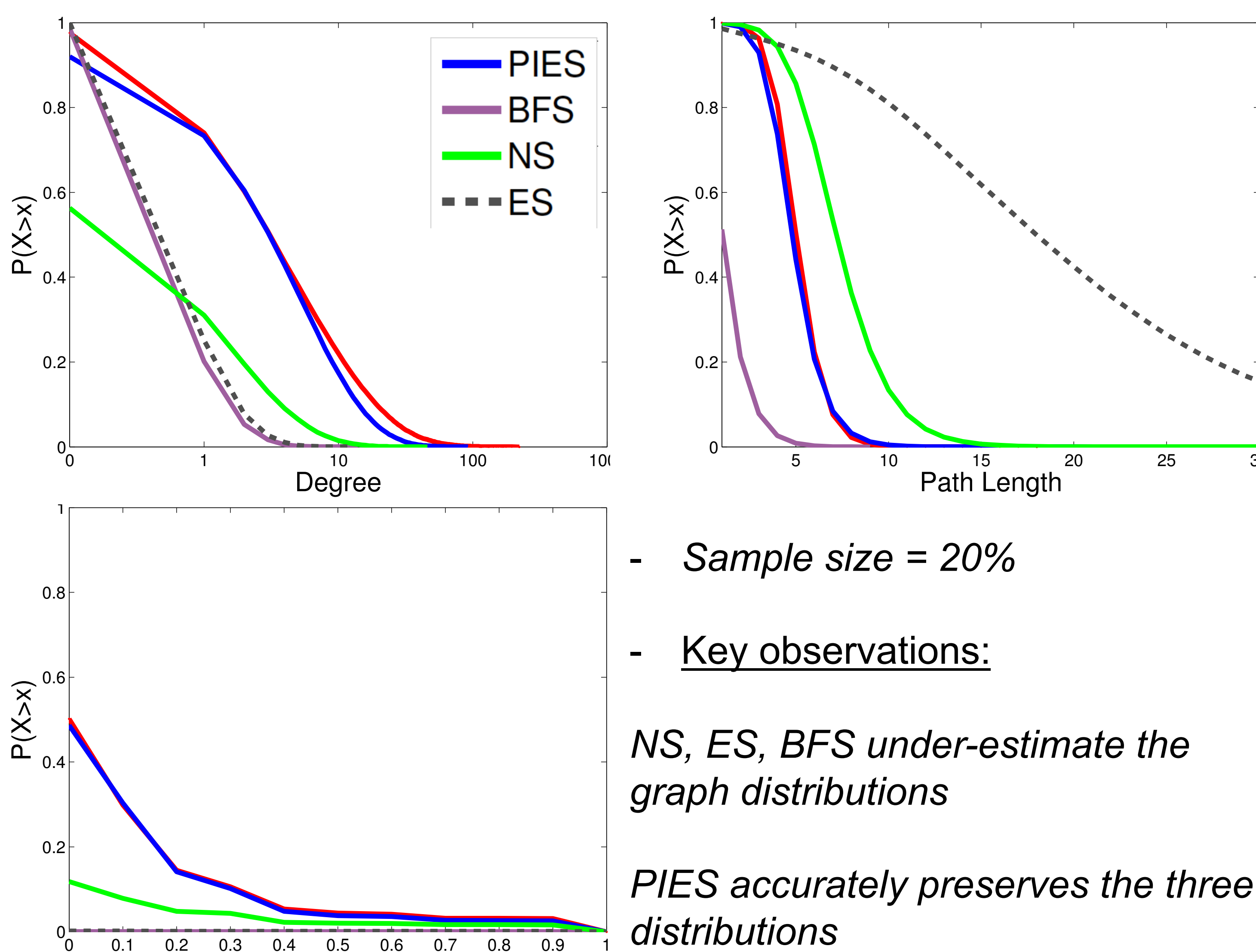
- PIES** *Partially Induced Edge Sampling*
 - using a reservoir to maintain a random sample
 - using partial induction in the forward direction of the stream to capture the connectivity of the nodes
- Implementation of Node Sampling
 - keep a reservoir with n nodes with min hash values
- Implementation of Edge Sampling
 - keep a reservoir with m edges with min hash values
- Implementation of Breadth-First Sampling
 - using a sliding window of size w

Results

Kolomogrov-Smirnov Statistic (KS distance)



Distributions for Facebook Network



Stream Evaluation for Facebook Network

