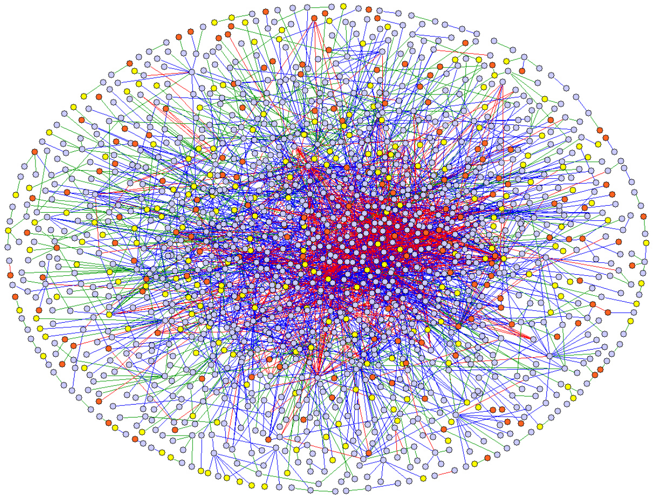
Monotone Sampling of Networks

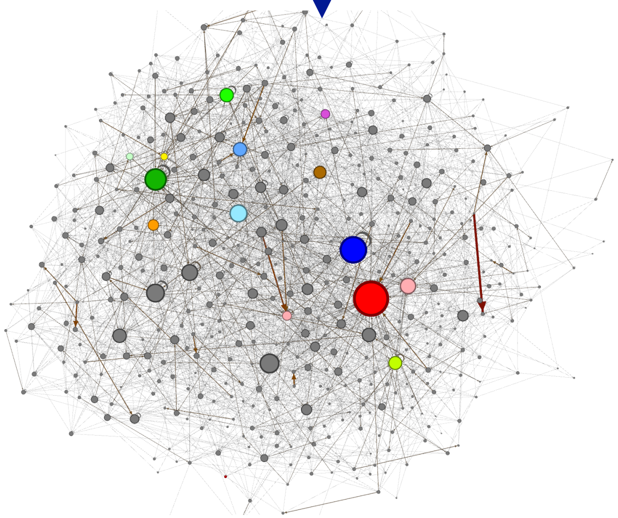
Tim Grube, Benjamin Schiller - P2P Networks - TU Darmstadt
Thorsten Strufe - Datenschutz und Datensicherheit - TU Dresden
DyNaK II / Nancy, France / 15.09.2014

Motivation



$$AverageDegree = ?$$
$$ClusteringCoefficient = ?$$
$$Diameter = ?$$

?

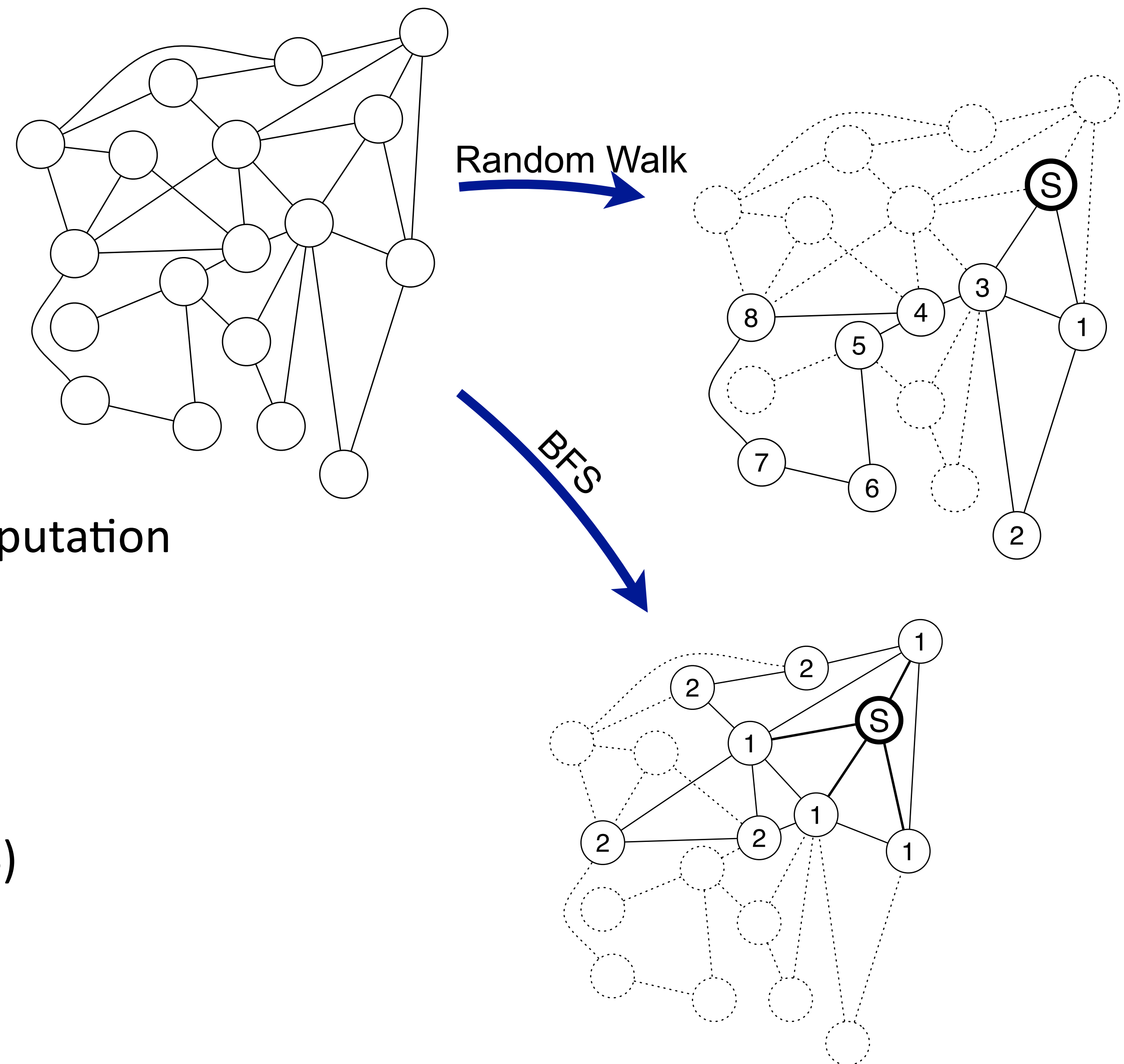


$$AverageDegree = d'_{avg}$$
$$ClusteringCoefficient = c'$$
$$Diameter = sp'_{max}$$

Outline

- Sampling
- Monotone Sampling
- Simulation
- Results
- Conclusion

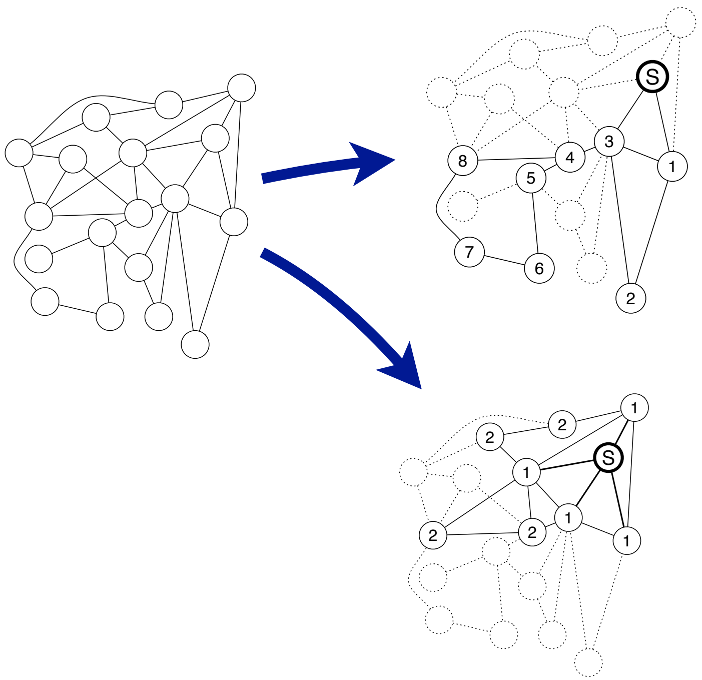
Sampling (1/2)



- Sampling
  - Reduce network size
  - Generates subgraph
  - Allows the property computation

- Based on
  - Random Walks
  - Breadth First Search (BFS)

Sampling (2/2)



- Many algorithms are biased

➡ Improvements:

  – Degree correction (Random Walk)

  – Randomized subset of neighbours (BFS)

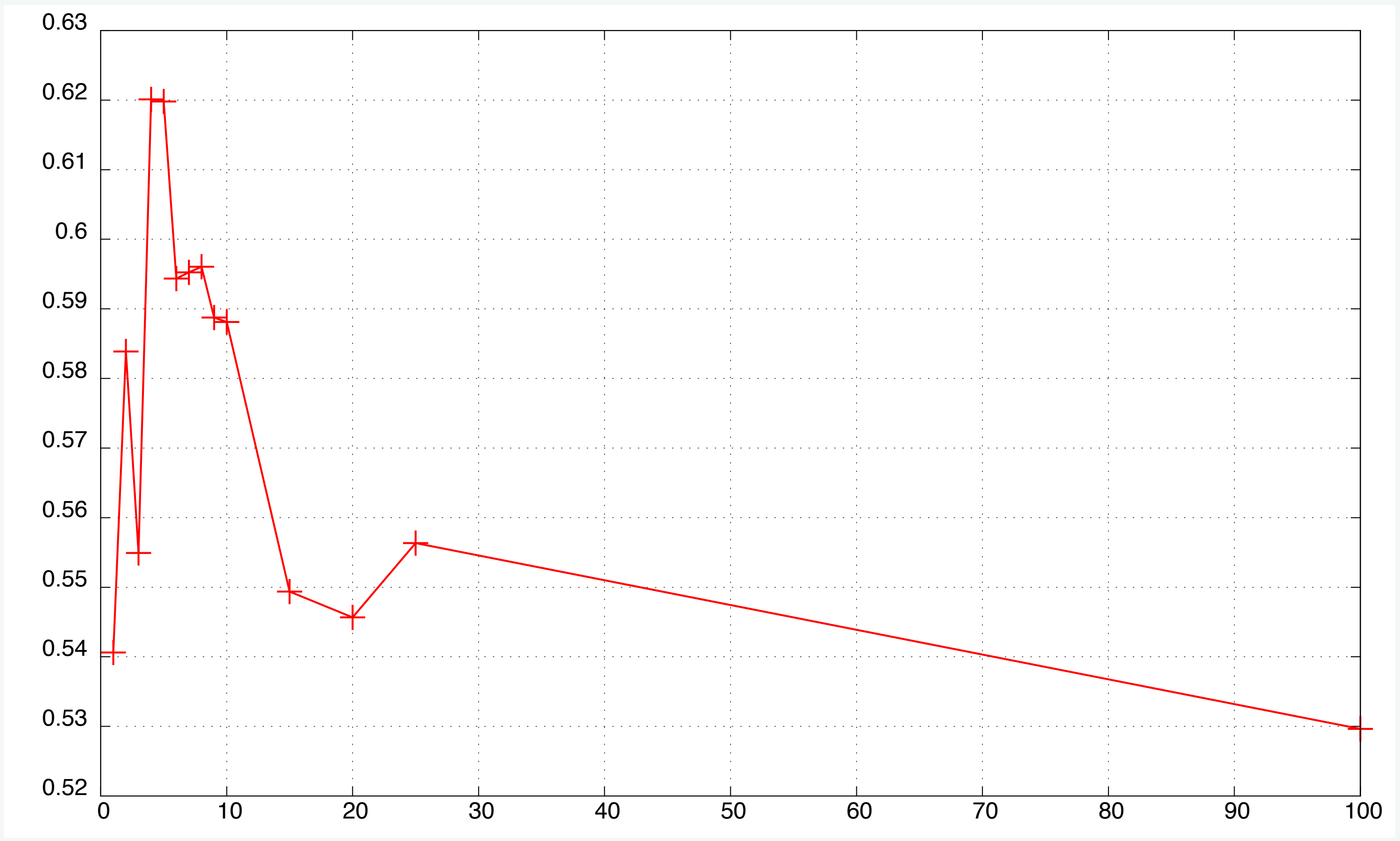
- Sampling reduces sizes

  - Reduced complexity of property computation

- But: **When to stop?** (w.r.t the desired properties)

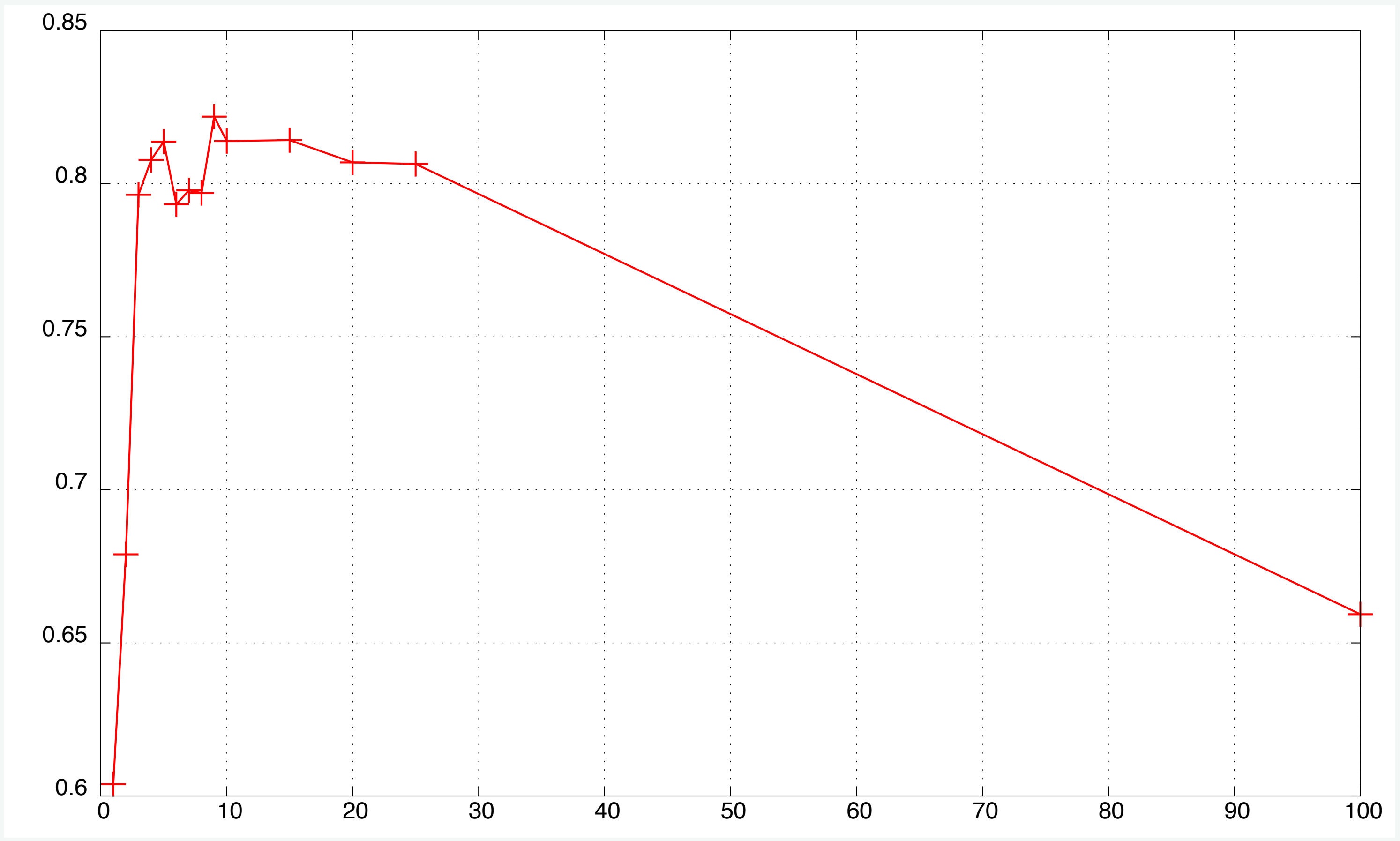
➡ Quality not traceable

Finding the best Algorithm



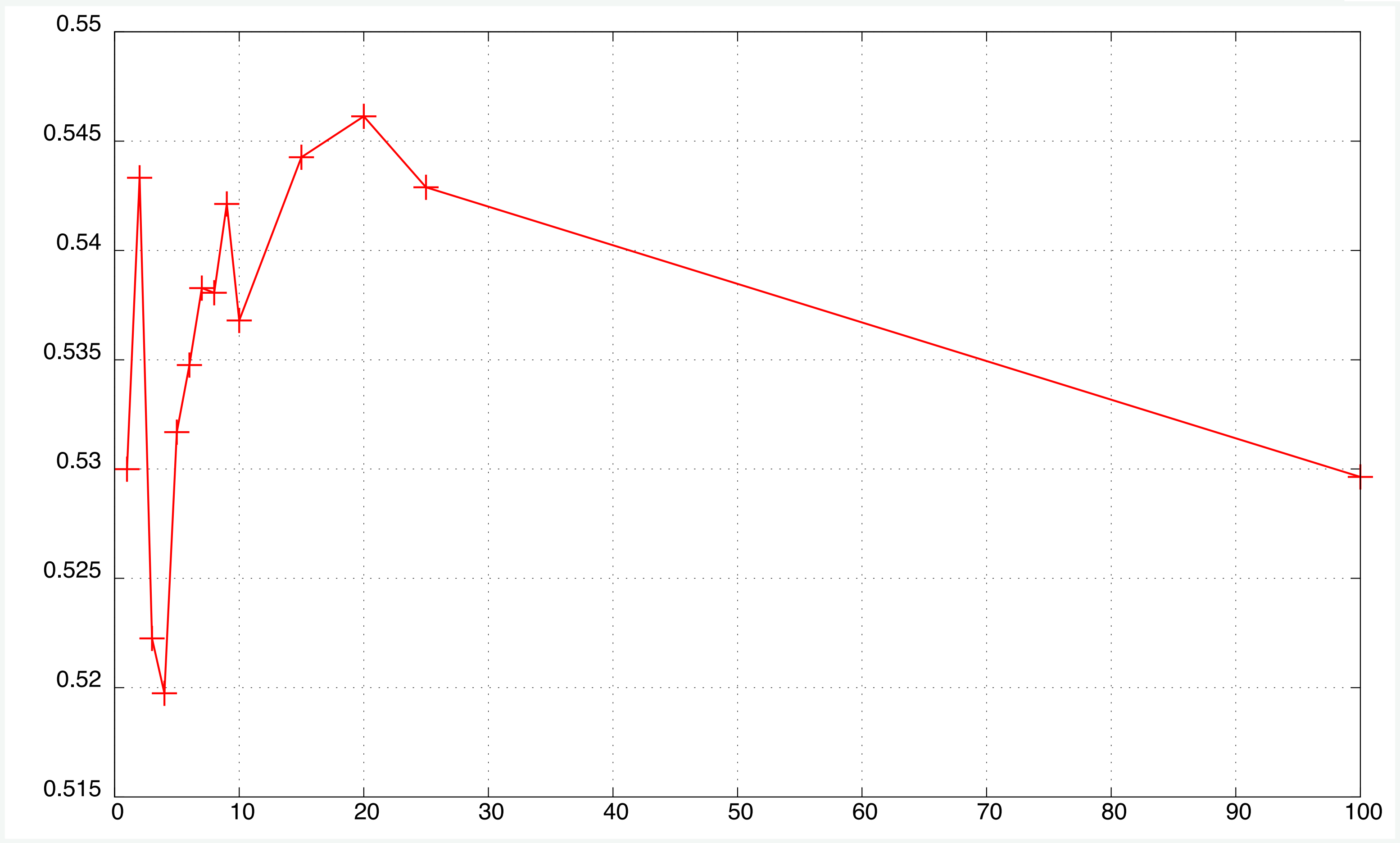
- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

Finding the best Algorithm



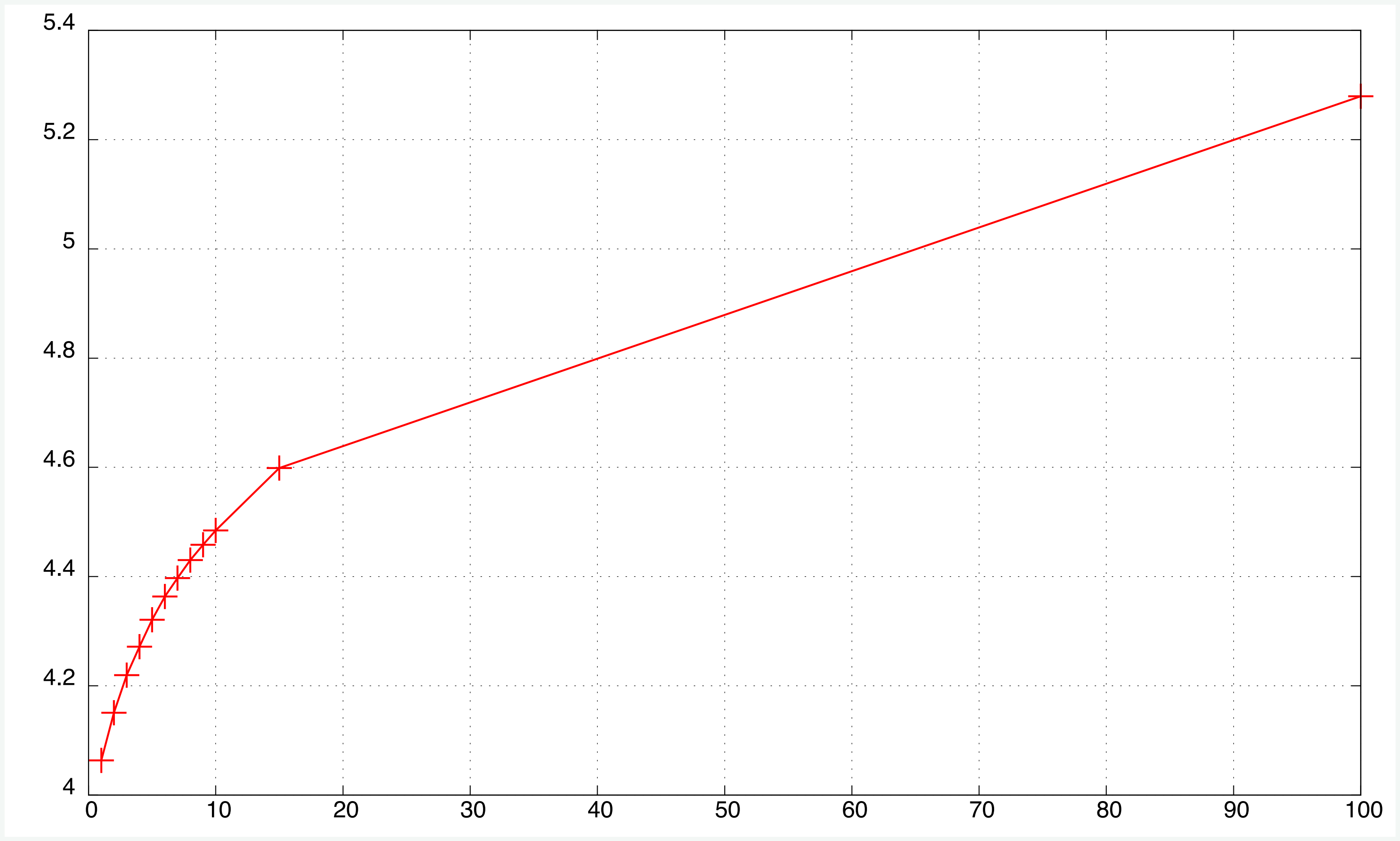
- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

Finding the best Algorithm



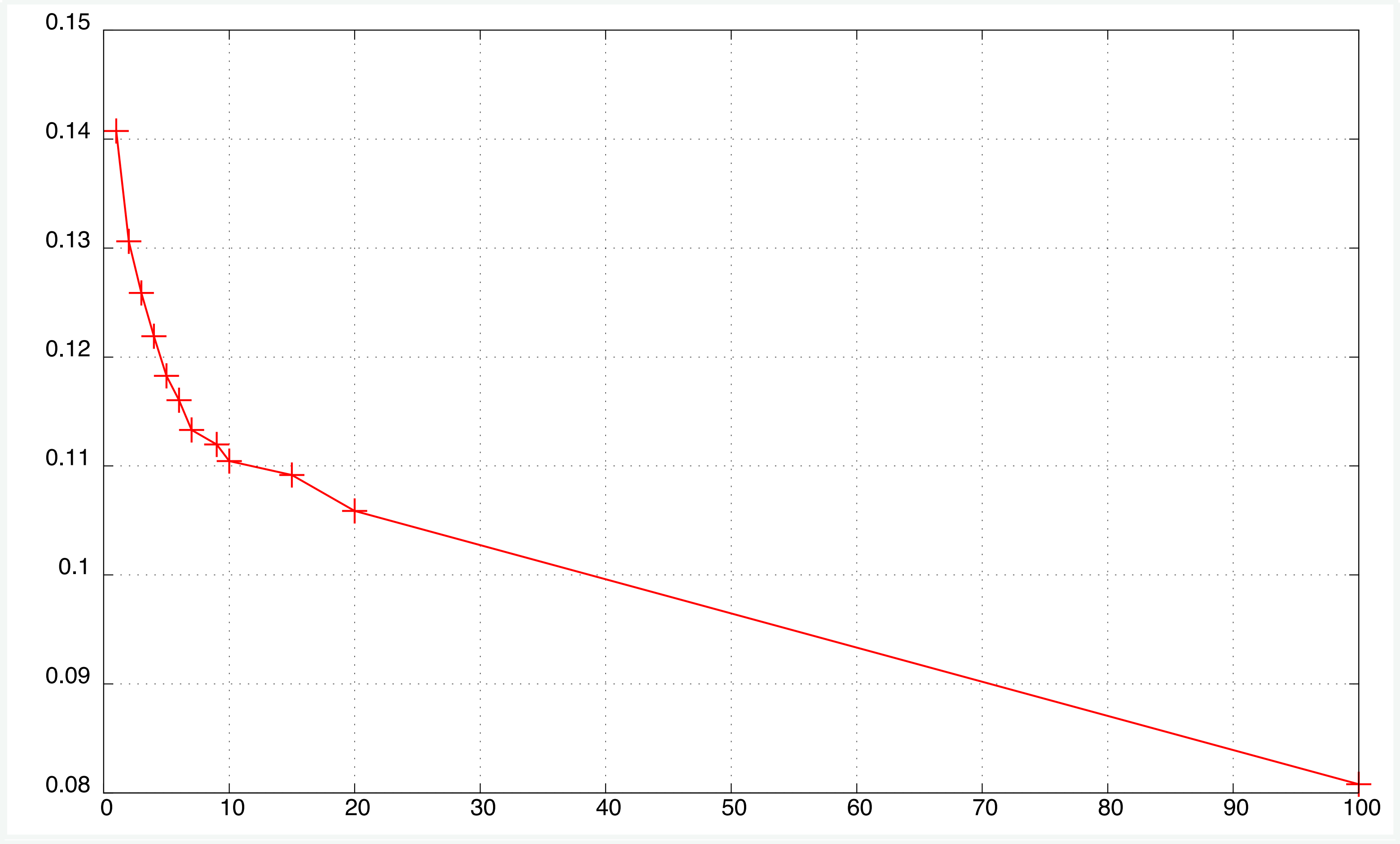
- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

Finding the best Algorithm



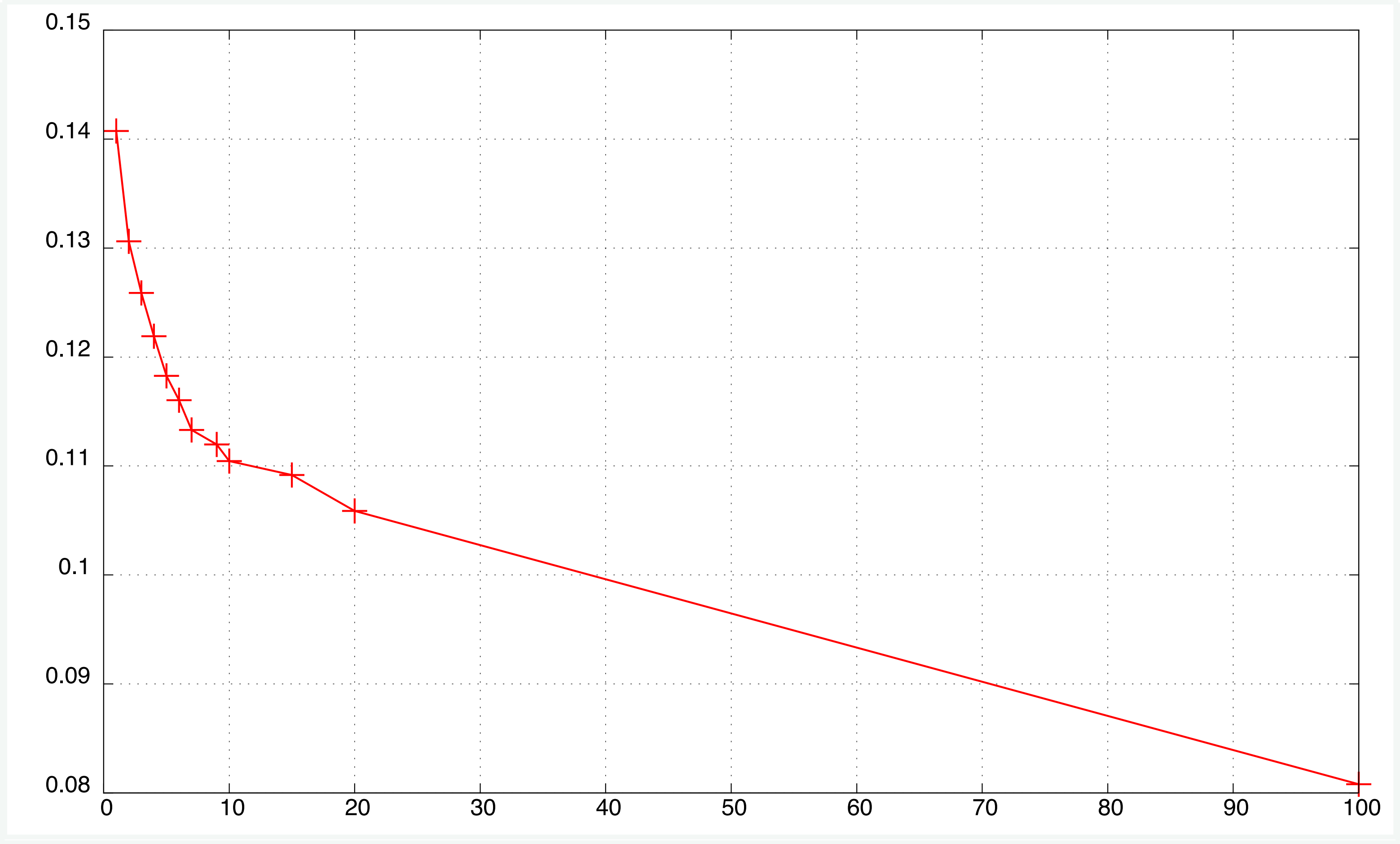
- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

Finding the best Algorithm



- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

Finding the best Algorithm



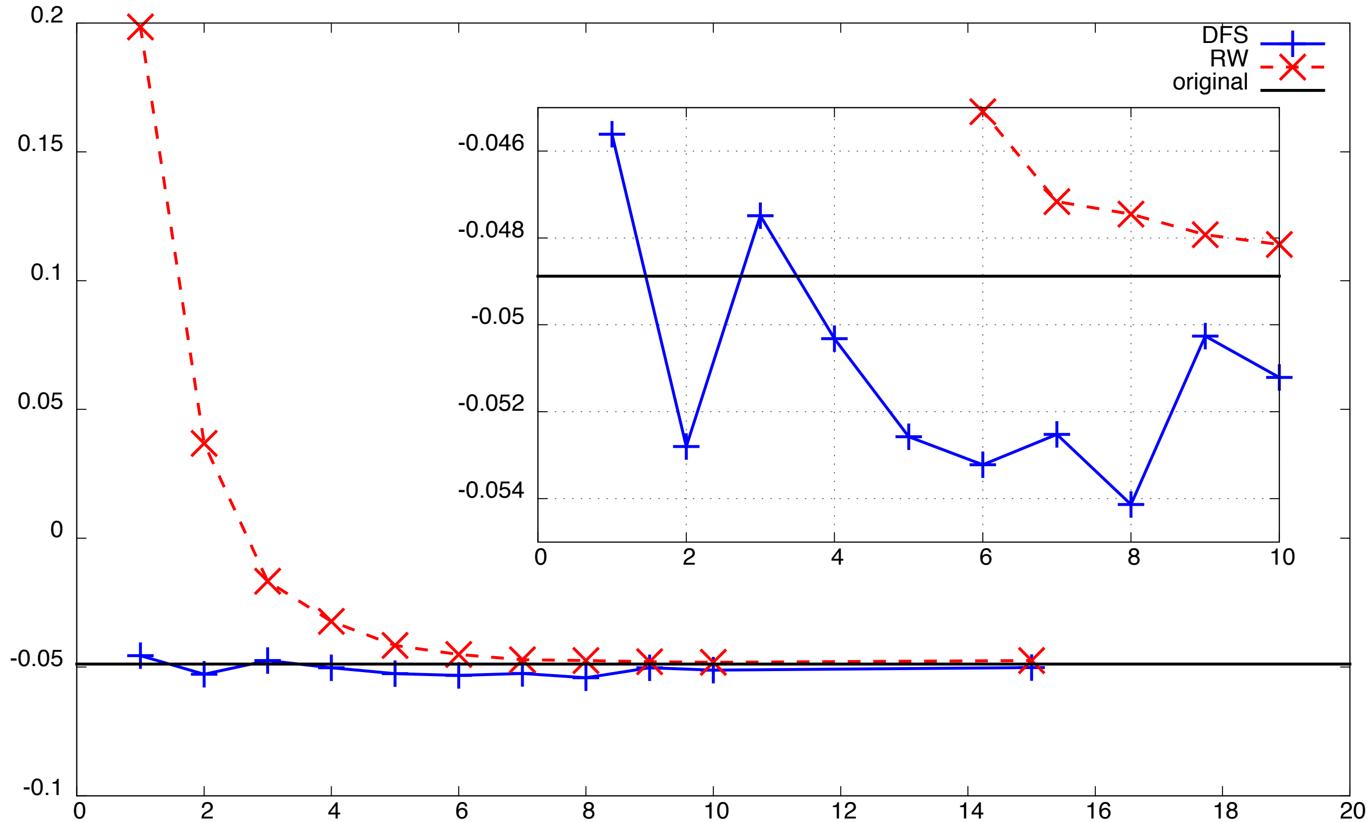
- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

➡ Some properties are converging **monotonously**

Idea: Exploit Monotonicity

- **Some** algorithms produce **monotonously converging properties**

» Allows **simple** and **good** extrapolation



- y-Axis:
  Property value
- x-Axis:
  Sampling Size (%)

- Compute samples and compare convergence of graph properties

Simulation: Setup (1/2)

- Sampling Algorithms: collected from related work

Random Walk
Random Jump
Random Walk with Degree Correction
Frontier Sampling
Random Stroll
Metropolized Hastings Random Walk
DFS
Random Stroll with Degree Correction

BFS
Snowball Sampling
Forest Fire
Respondent-driven Sampling

- Metrics: typical graph-theoretic properties (single scalar)

Characteristic Path Length
Diameter
Effective Diameter

Assortativity Coefficient

Average Clustering Coefficient

Average (In/Out) Degree
Maximum Degree

Simulation: Setup (2/2)

- Networks[1]
  - social
    - social-directed
    - social-undirected
  - Peer-to-Peer
  - Webgraph
- Runs

  $<$ 500.000 Nodes
  - 100 runs
  - 1-10%, 15%, 20%, 25%

  $>$ 500.000 Nodes
  - 20 runs
  - 1-10%, 15%
- Baseline

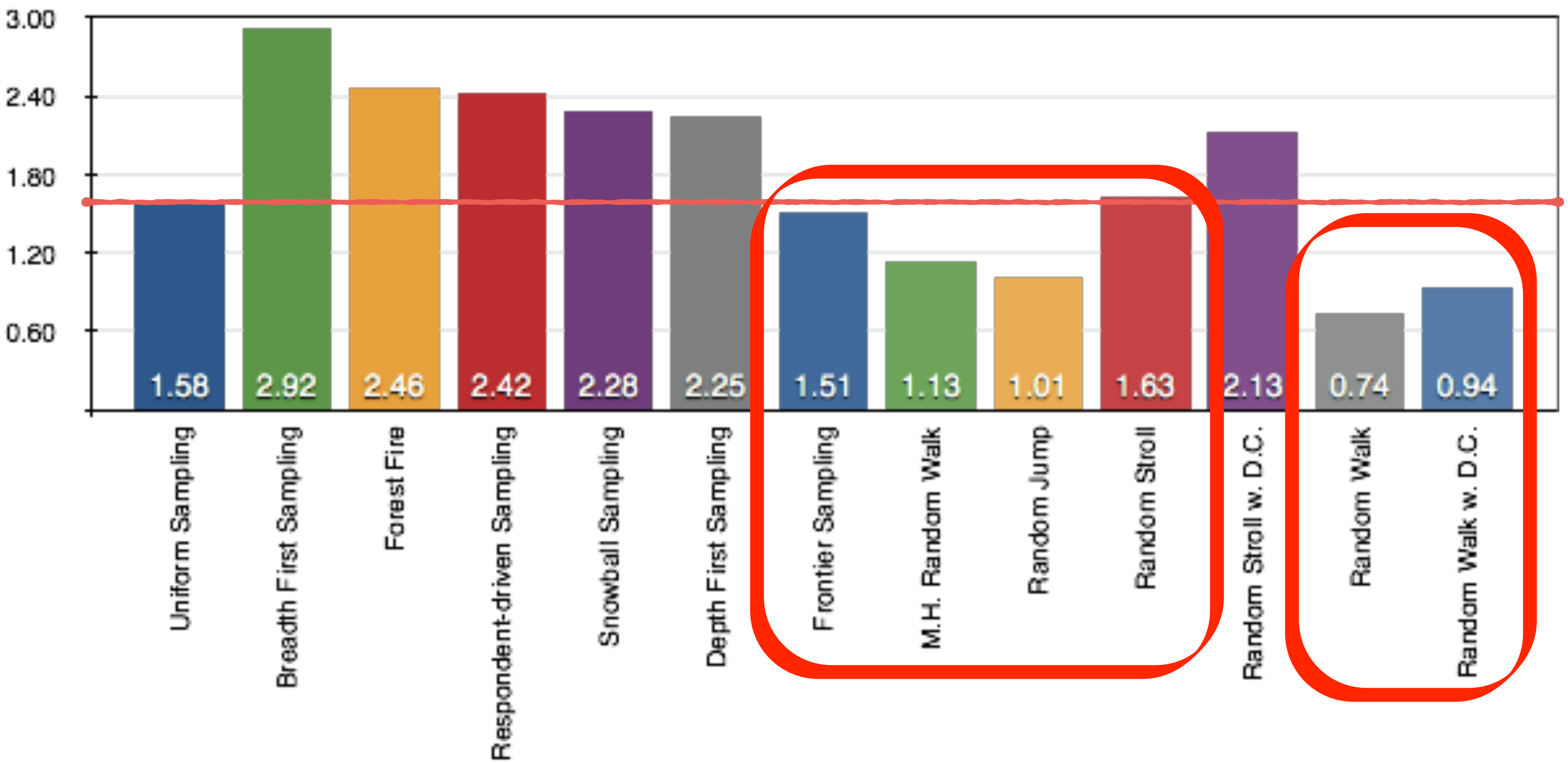
  Uniform Sampling
- Monotonicity

$$\boxed{m = \frac{1}{M} \sum_{M} countChangesInConvergenceDirectionOfProperty(M_i)}$$

[1]: Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/)

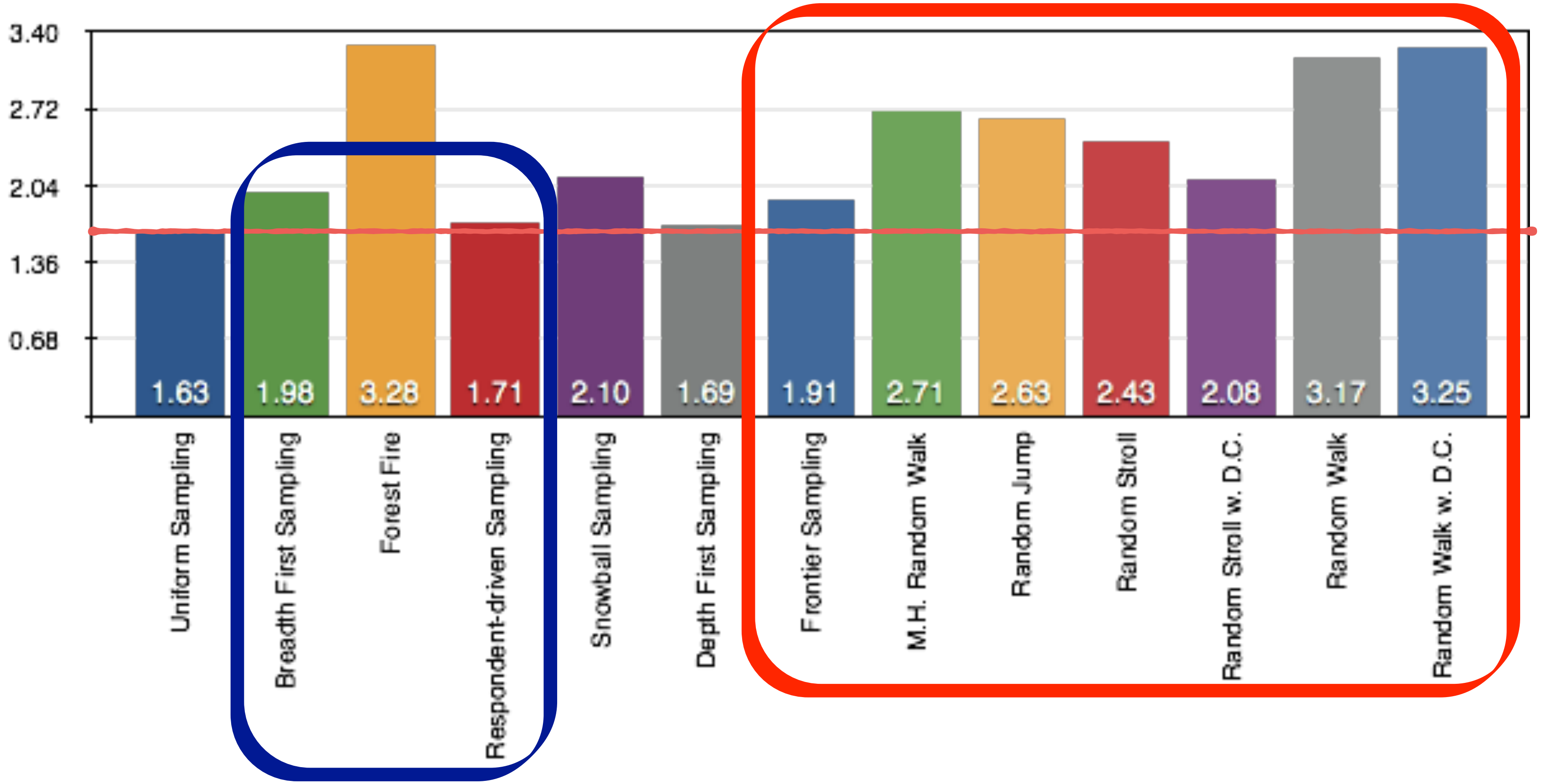
Results: Ranking w.r.t. to Monotonicity (1/4)

- Directed Social Networks



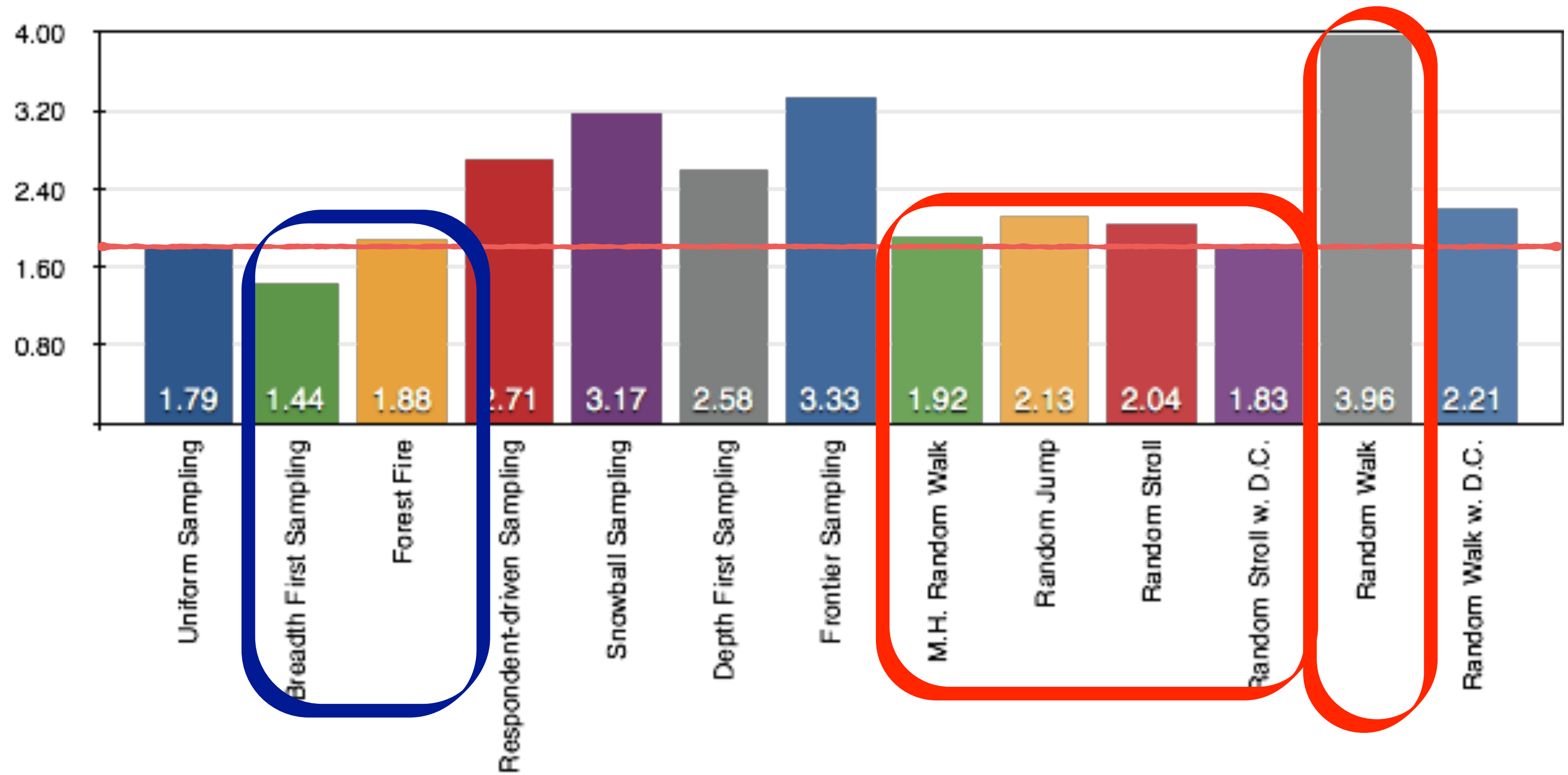
Results: Ranking w.r.t. to Monotonicity (2/4)

- Undirected Social Networks



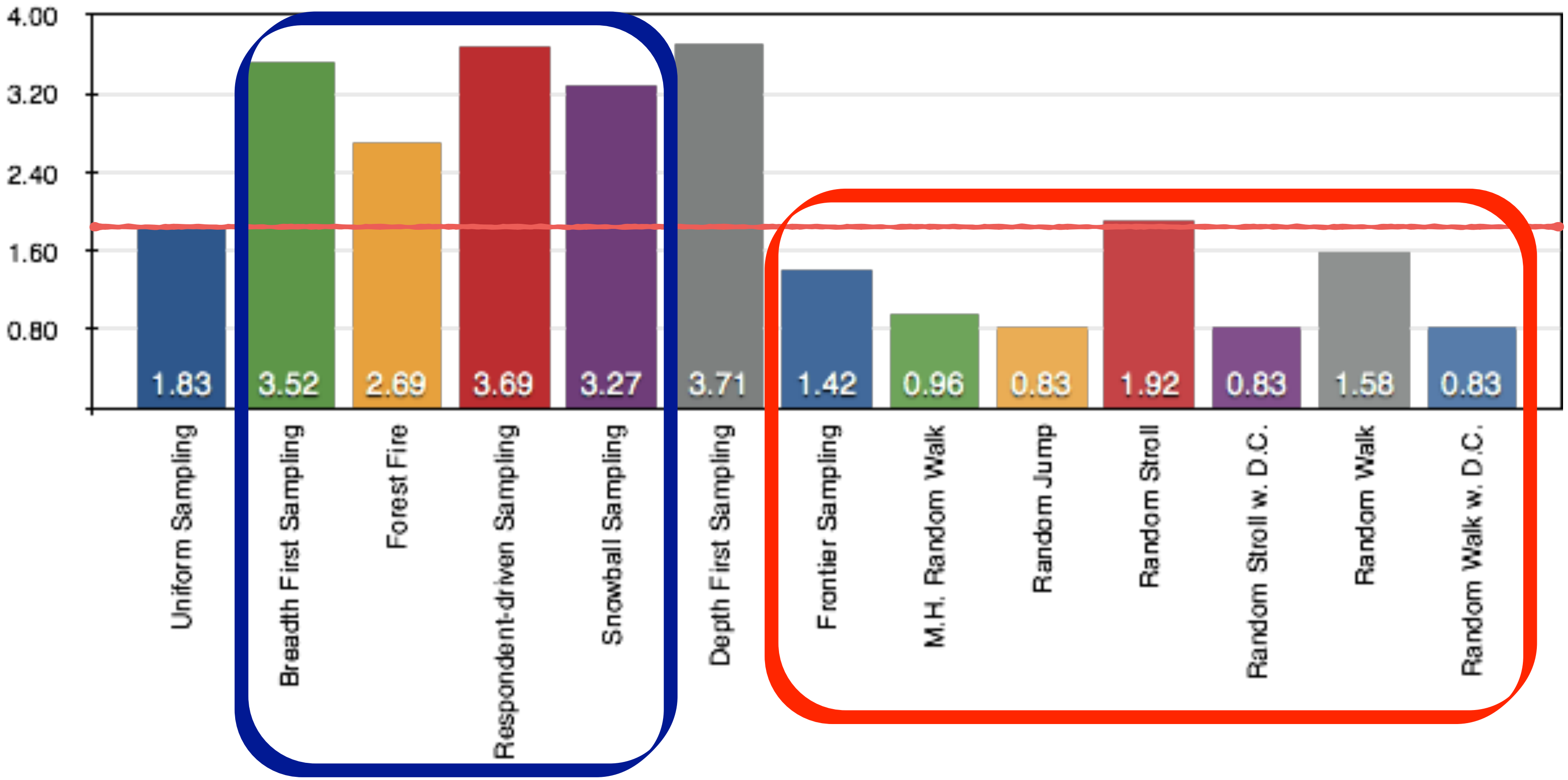
Results: Ranking w.r.t. to Monotonicity (3/4)

- Peer-to-Peer networks



Results: Ranking w.r.t. to Monotonicity (4/4)

- Webgraph



Conclusion & Future Work

- Improved Algorithms not necessarily better
- Basic Algorithms performing good
- Random Walks mostly outperforming BFS
- Interesting: Uniform Sampling always in the better half
- Future Work
  - Analyze the monotonicity of distributions
  - Understand the reasons for the differences of the monotonocity
  - Develop estimators for network properties