Sampling-based network analysis

Tim Grube, P2P Networks, TU Darmstadt

23.01.2014

Outline

- Problem statement
- Our approach
- Background
- Sampling algorithms
- Evaluation
- Conclusion

Problem statement

- Networks grow very large
- Complexity/Size is too high to analyse the complete network
- Reduce the complexity and size: Sampling

- What is the impact of sampling?
  - analysed by many researchers
  - but only a few algorithms and single properties

Our Approach

- Analyse more sampling algorithms with more metrics
- Answer:
  - Impact of sampling?
  - Differences between the sampling techniques?
- Long Term:
  - Properties of sampled network → Properties of the original network

Background

Network models

- Regular network (random topology)
- Random network (Erdös-Rényi)
- Scale-free network (Barabási-Albert)
- Small-world network (Watts-Strogatz)
- Rich-club network (Zhou-Mondragon)
- Special: Ring, Clique

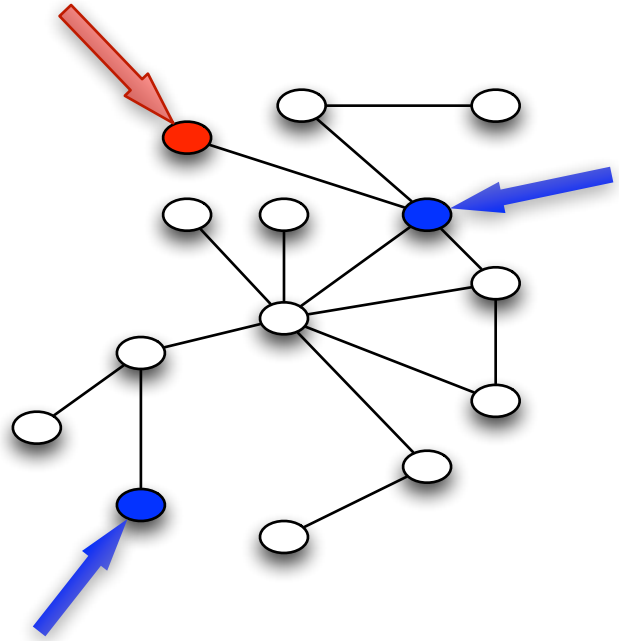
Metrics

- Network properties
  - Degree Distribution, Clustering Coefficient, Assortativity
  - Hop-Plot(SPLD), Diameter, effective Diameter, Characteristic Path Length
  - Betweenness Centrality, Page Rank
- Sampling algorithm properties
  - Selection Bias, Sample Modularity

Sampling Algorithms

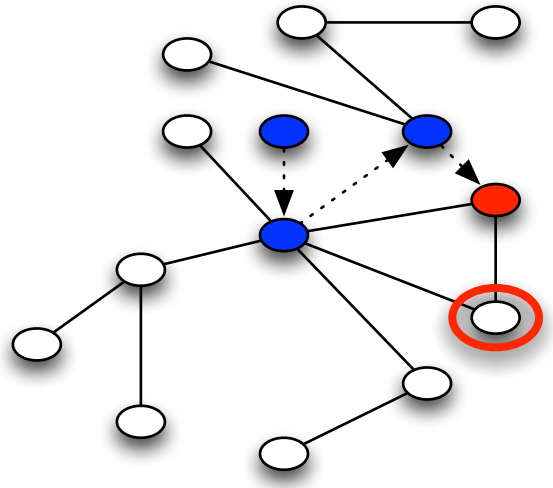
Properties

- Dimension
- Self-awareness
- Start node selection
- Restart procedure
- Walker
- Sampler



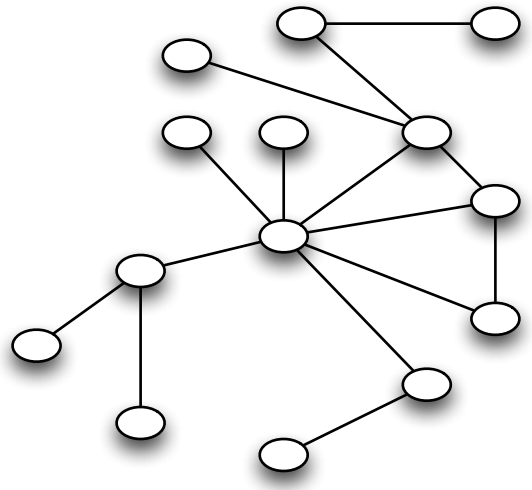
Properties

- Dimension
- Self-awareness
- Start node selection
- Restart procedure
- Walker
- Sampler



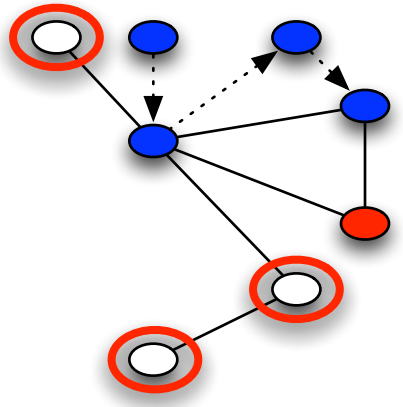
Properties

- Dimension
- Self-awareness
- Start node selection
- Restart procedure
- Walker
- Sampler



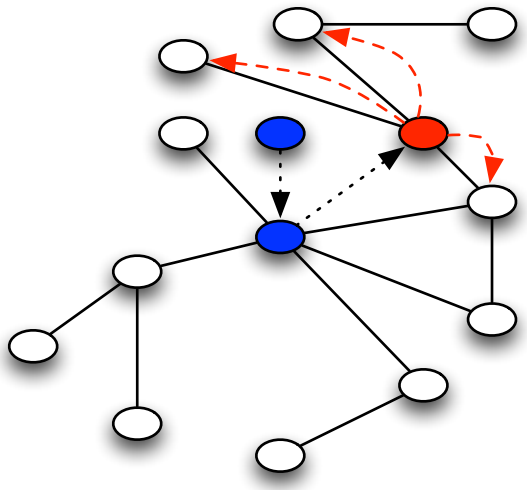
Properties

- Dimension
- Self-awareness
- Start node selection
- Restart procedure
- Walker
- Sampler



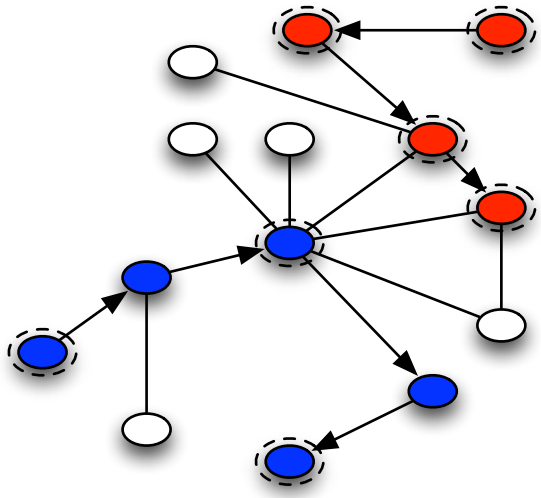
Properties

- Dimension
- Self-awareness
- Start node selection
- Restart procedure
- Walker
- Sampler



Properties

- Dimension
- Self-awareness
- Start node selection
- Restart procedure
- Walker
- Sampler



Sampling algorithms (1/2)

- Uniform/Random Node Selection
  - Uniform Sampling (US)
- Breadth First Sampling (BFS)
  - BFS
  - Snowball Sampling (n of k unvisited neighbours)
  - Respondent-driven Sampling (RDS) (random n of k neighbours)
  - Forest Fire (select neighbours with probability p)
- Depth First Sampling (DFS)
  - DFS

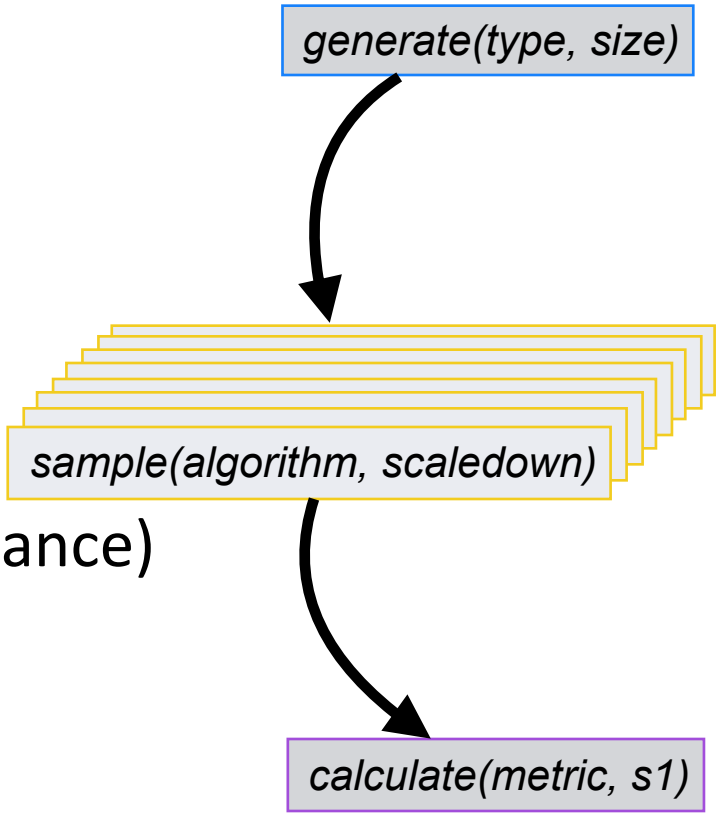
Sampling algorithms (2/2)

- Random Walk (RW)
  - RW
  - RW with degree correction (p depends on neighbour's degree)
  - RW with multiple dimensions (m independent walker)
  - Frontier Sampling (FS) (m dependent walker)
  - Random Stroll (RS) (skip n intermediate nodes)
  - RS with degree correction („RW with degree correction + RS“)
  - Random Jump (RW + jumping to a random node with probability p)

Evaluation

Evaluation Setup: GTNA

- Generate 5 network instances
- Run 16 samples (per instance)
- Calculate and aggregate metrics (per instance)



Evaluation Setup: Networks

- Size: 10.000 Nodes
  - Clique: 1.000 Nodes (due to computational complexity)
- Scaledown: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
  - Sample sizes: 1.000 - 9.000 Nodes

Evaluation Setup: Sampling Algorithms

- Dimension: 1 (except: RW multiple , FS: 5)
- Self-aware: true (except: RDS)
- Startnodes: random
- Restart: random (DFS: backtracking)
- Walker/Sampler: depends on sampling algorithm

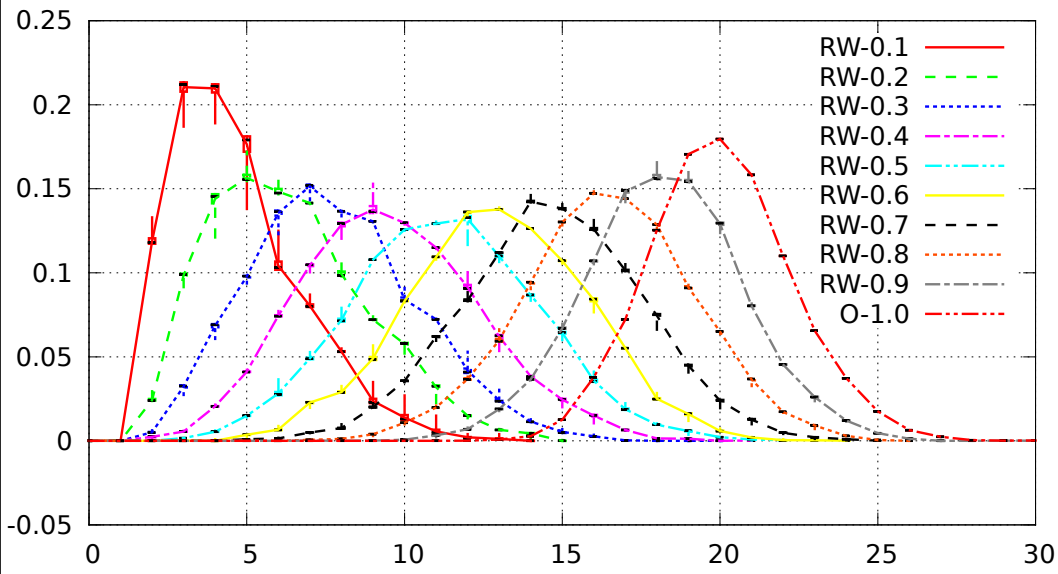
Results

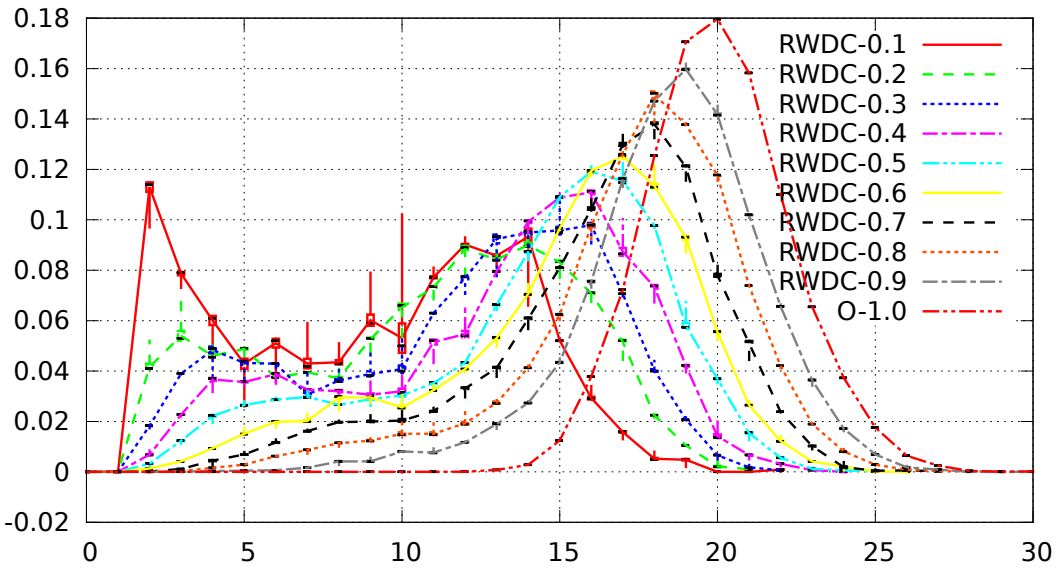
Degree Correcting Algorithms

- Researchers opinion: *„A Degree Distribution has to be preserved well in order to achieve a good sample“*

Degree Correcting: Degree Distribution

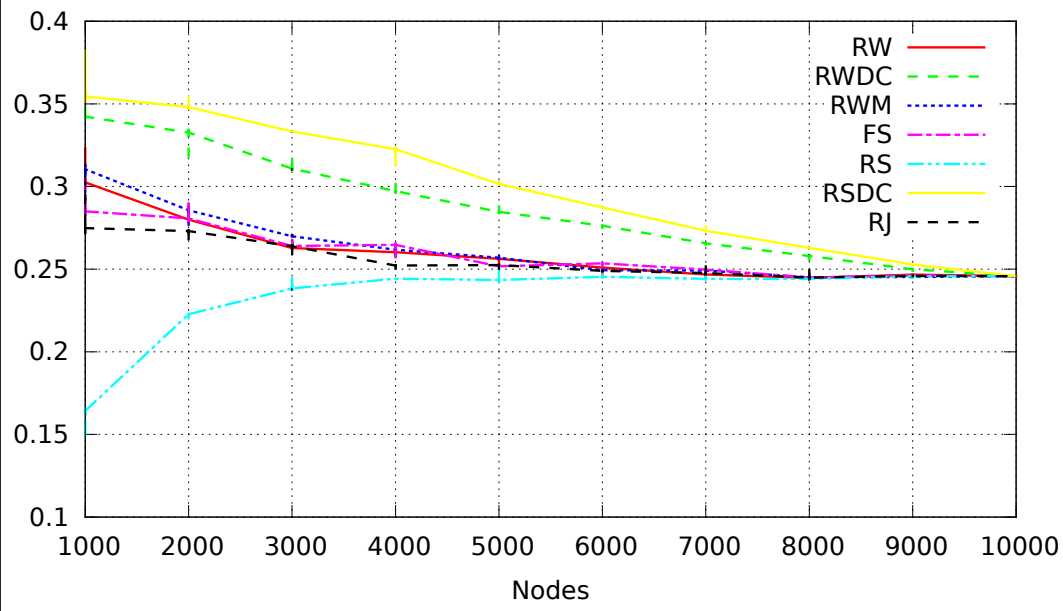
- scaling down a distribution (intuitive)
  - increase variance
  - decrease mean probability



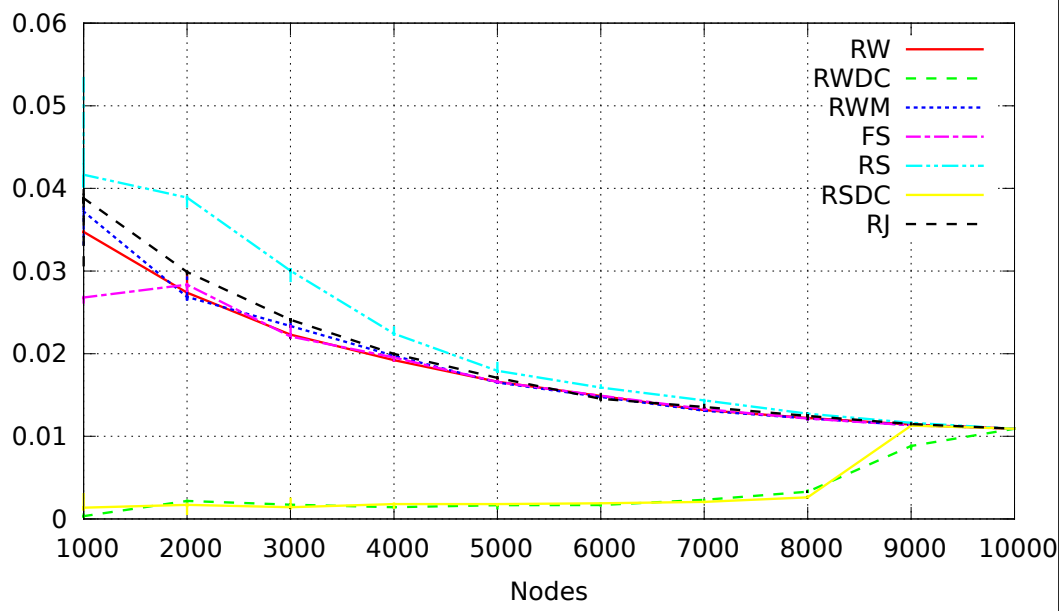


Degree Correcting: Clustering Coefficient

- Small-world sample clustering coefficient higher
- Scale-free sample clustering coefficient lower



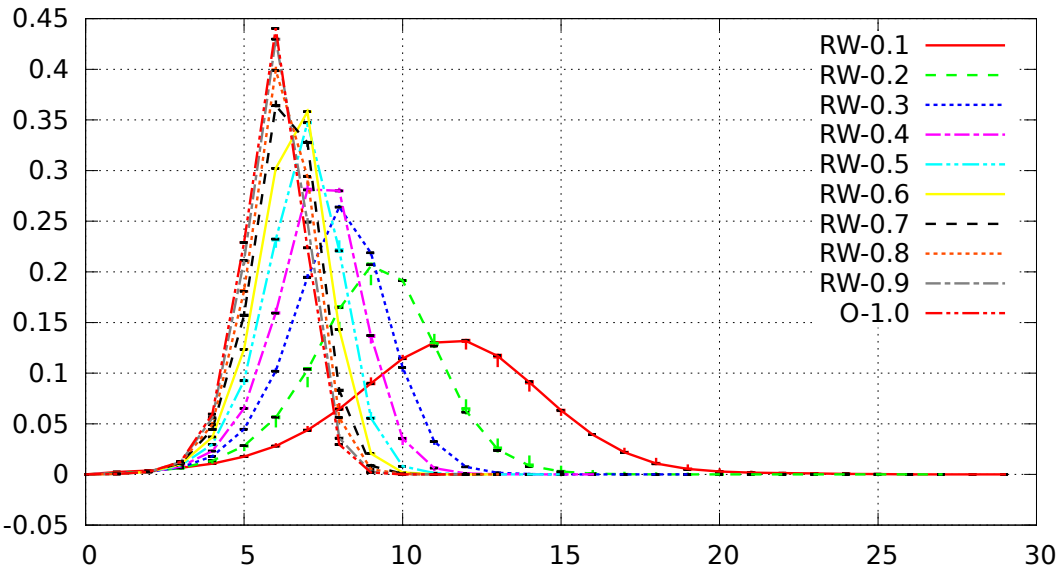
Small-world Network

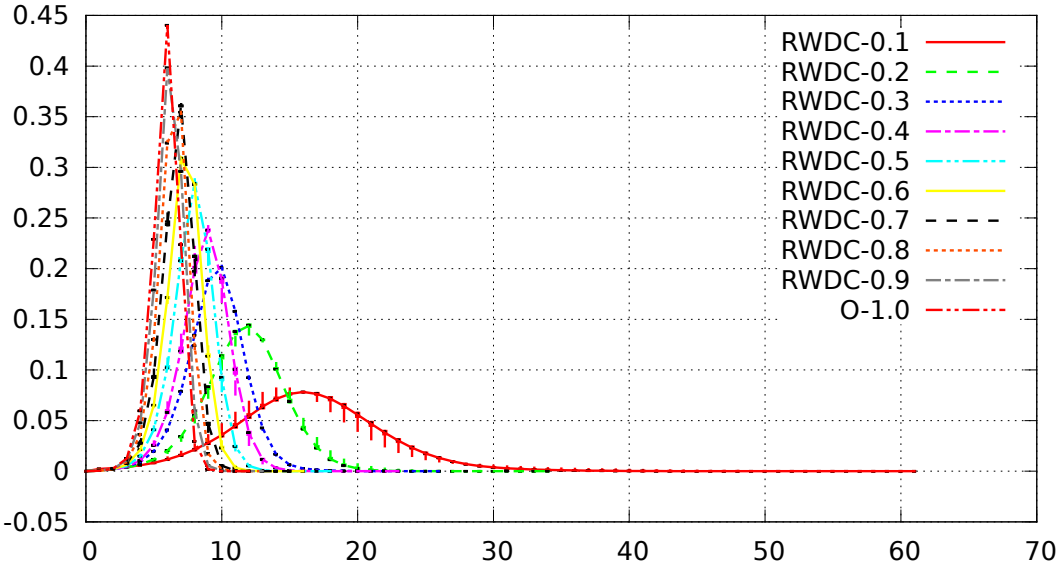


Scale-free Network

Degree Correcting: Shortest Path Length

- Degree correction causes much longer paths





Degree Correcting Algorithms

- *Degree correction helps to preserve the degree distribution*

- but
  - Other network properties are distorted
  - Inconsistencies among the 16 sample runs

Simple or complex Algorithms?

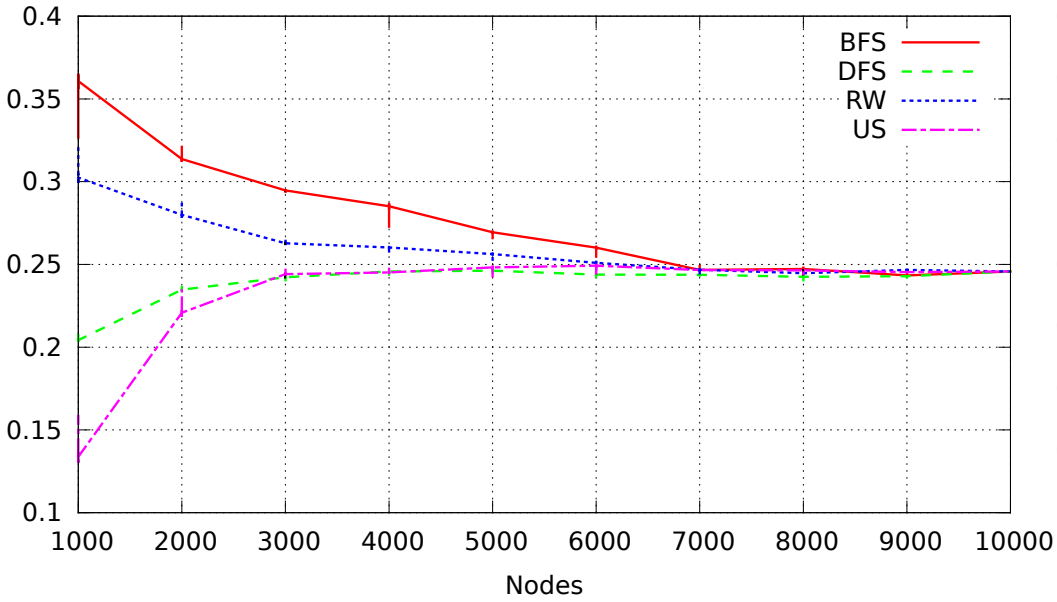
- Many improvements suggested for simple sampling algorithms
- We used simple algorithms like BFS, RW and US as a baseline

Breadth First Sampling

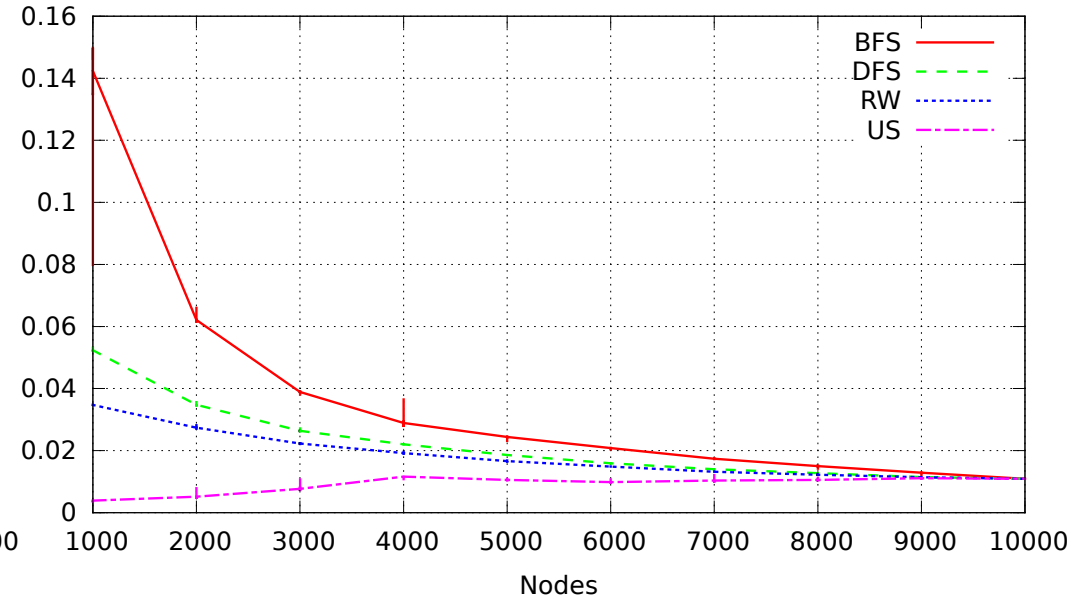
- related Work:
  - highly biased
  - bad sample quality

BFS: Clustering Coefficient

- „higher by definition“
- slight inconsistencies



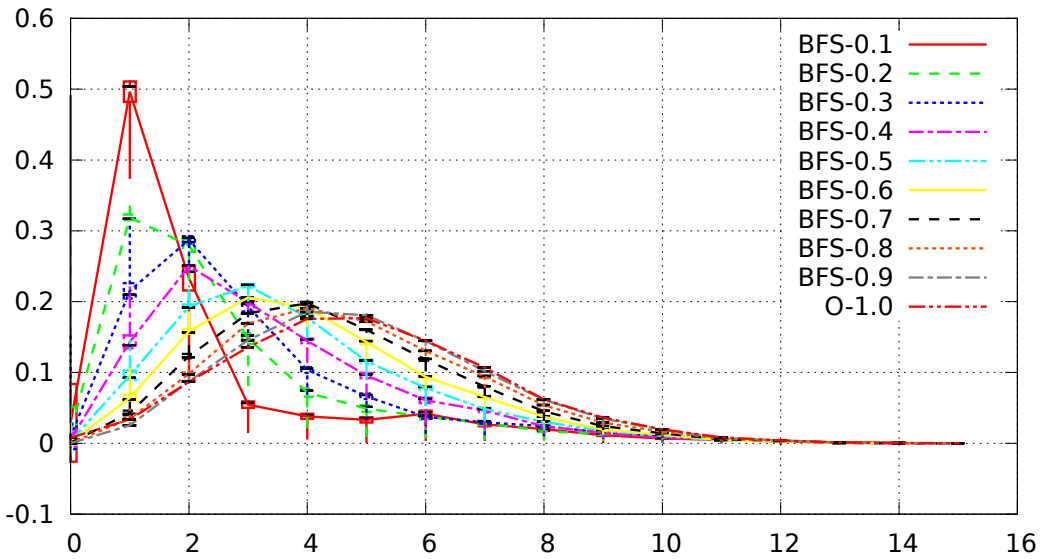
Small-world network

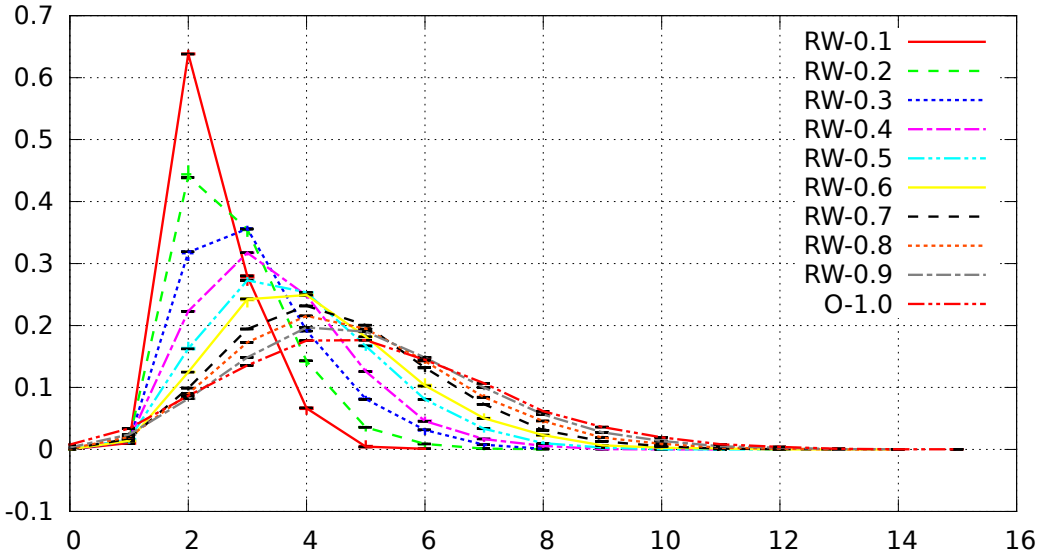


Scale-free network

BFS: Degree Distribution

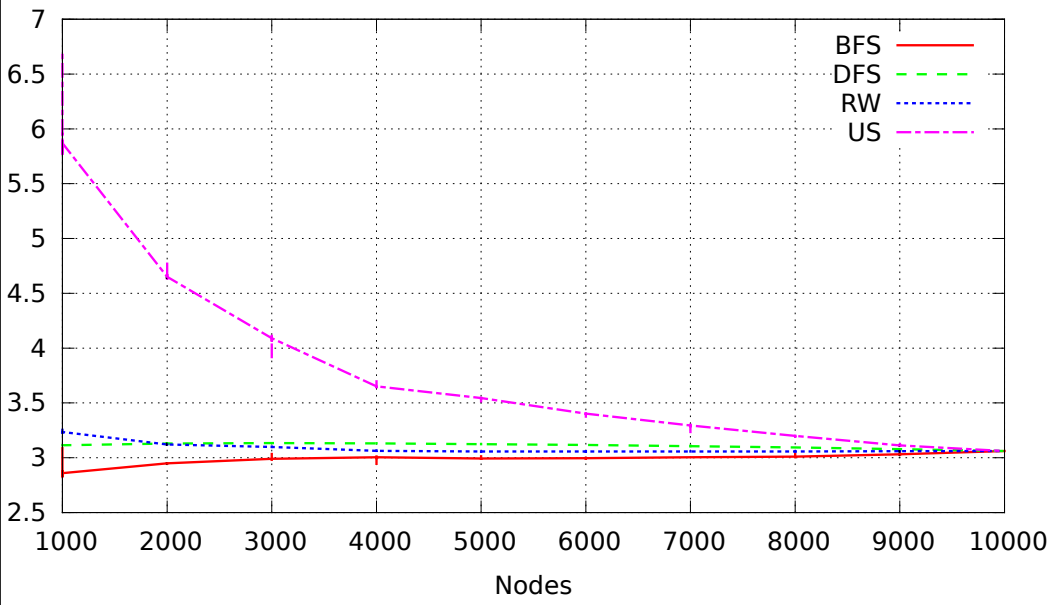
- slightly larger inconsistencies
- similar progression



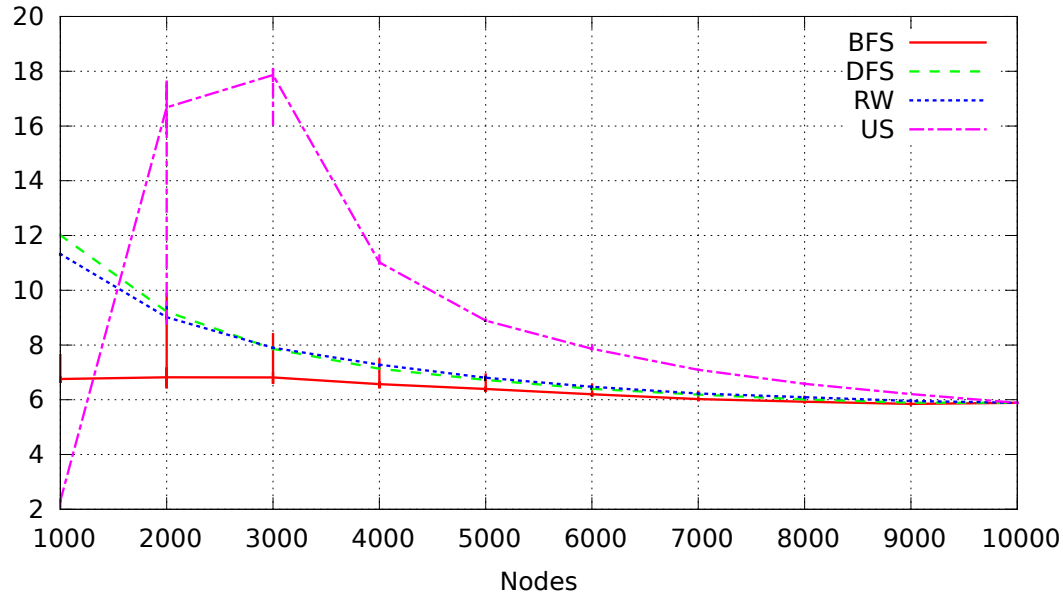


BFS: Characteristic Path Length

- Characteristic Path Length similar or better than RW samples



Scale-free network



Random network

BFS: not as bad as expected

- *related Work prompts BFS as „bad“ sampling algorithm*

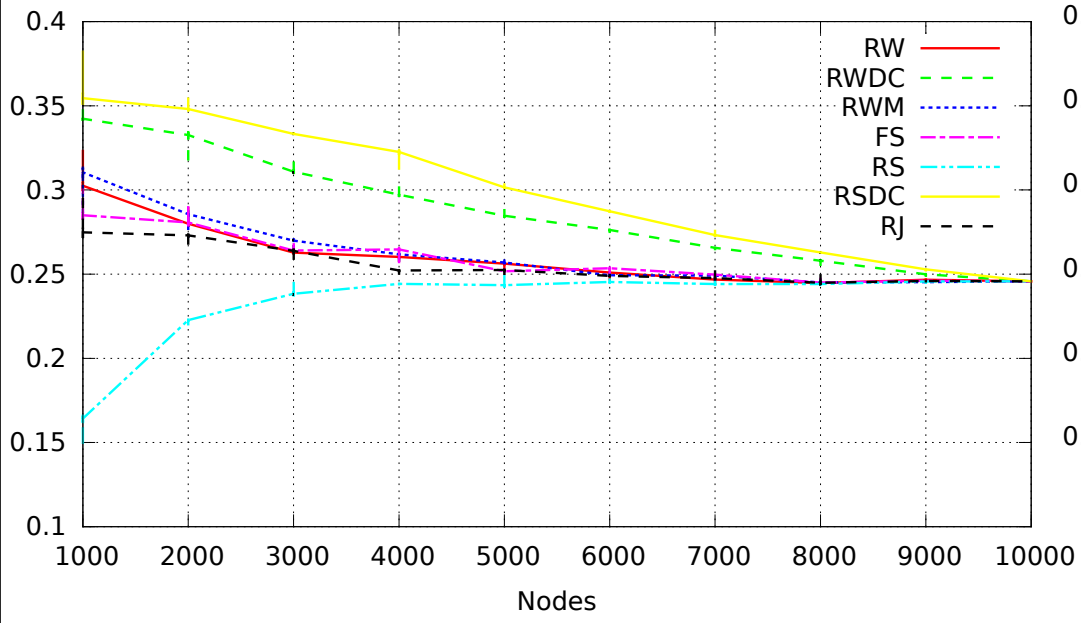
- but:
  - (slight) inconsistencies
  - different by construction, but not unusable bad!
  - only slightly higher biased than other sampling algorithms

Random Walk

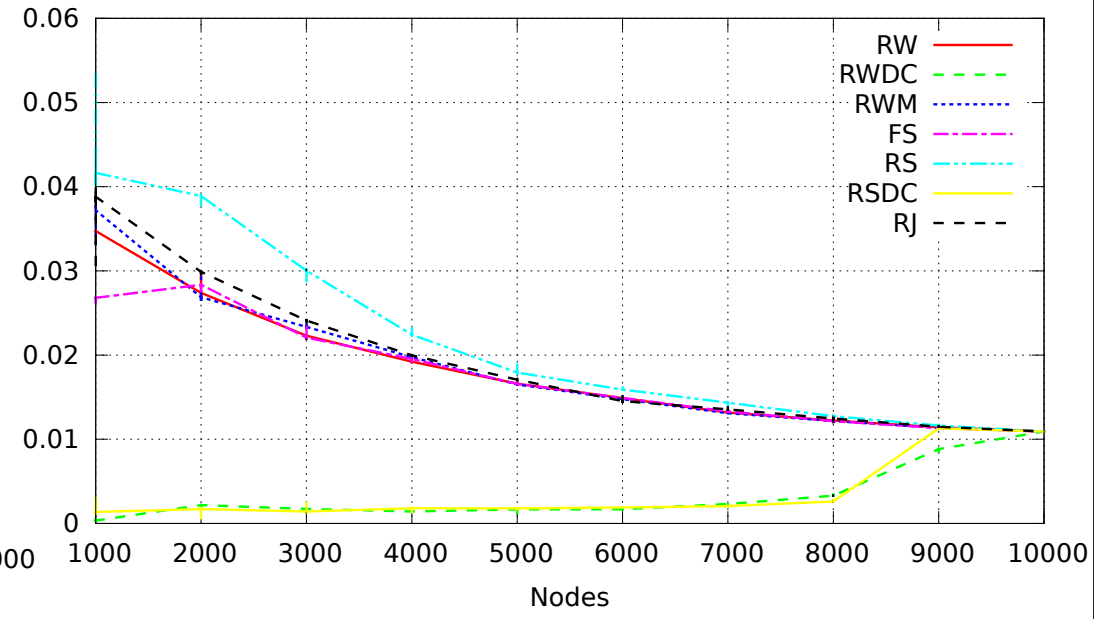
- many improvements for the classical RW:
  - degree correction
  - multiple (dependent) walker
  - skip intermediate nodes
  - jump with a certain probability

RW: Clustering Coefficient

- small or none inconsistencies
- tightly surrounded by the „improved“ algorithms



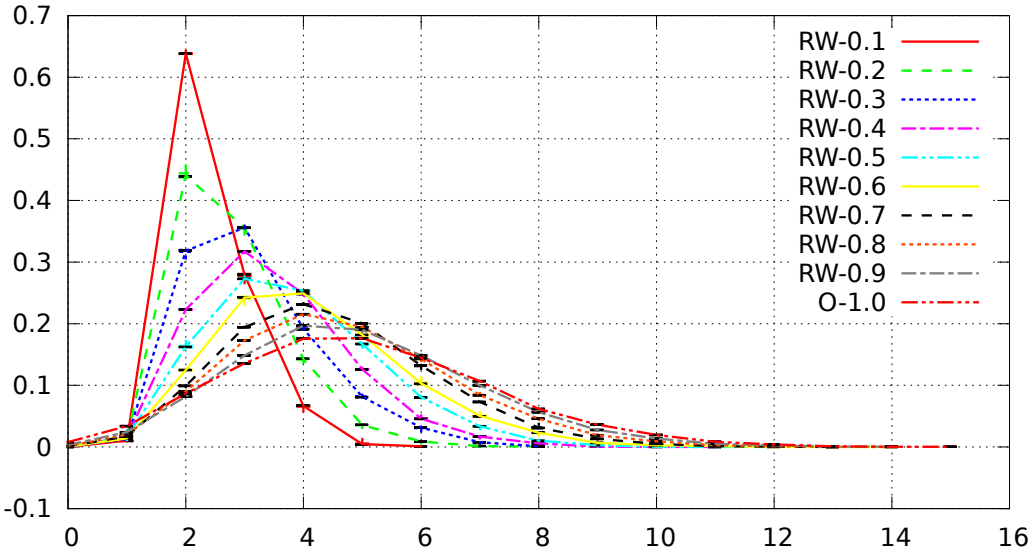
Small-world network

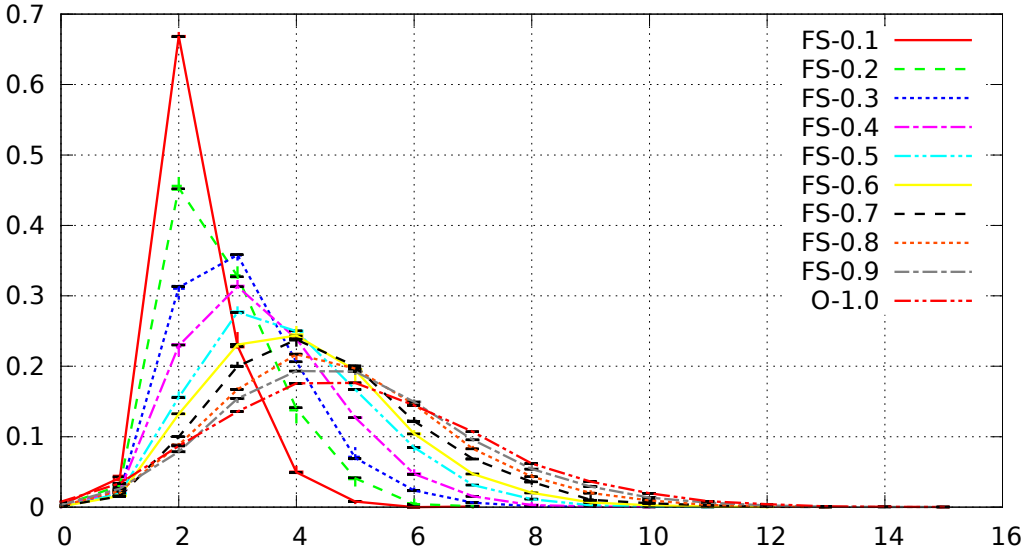


Scale-free network

RW: Degree Distribution

- similar progression
  - advanced algorithms showing larger inconsistencies





RW

- *many improvements for the classical RW available*

- but classical RW has
  - … less inconsistencies
  - … the same behaviour as the improved versions

Simple or complex Algorithms?

- *many improvements suggested for simple sampling algorithms*

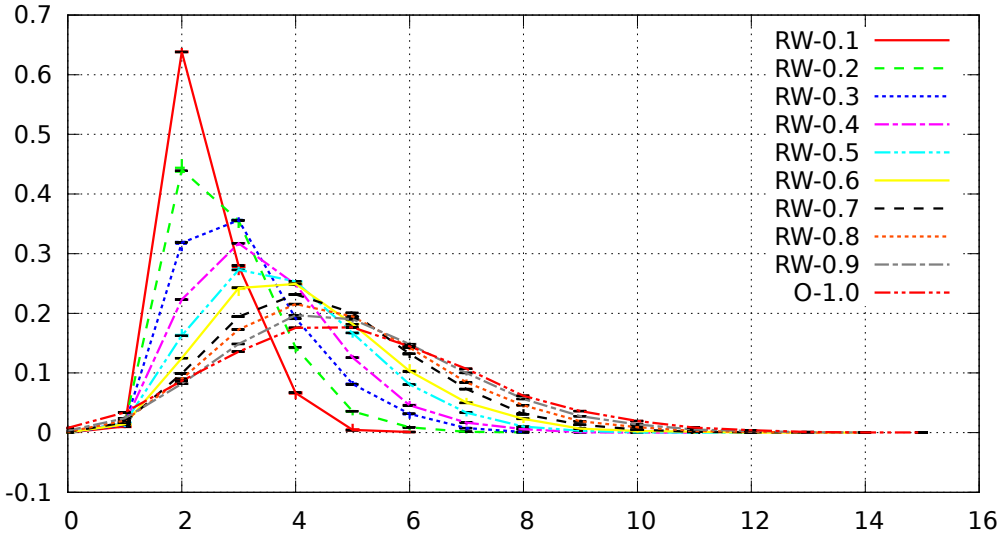
- but
  - the simple algorithms are not really „worse“

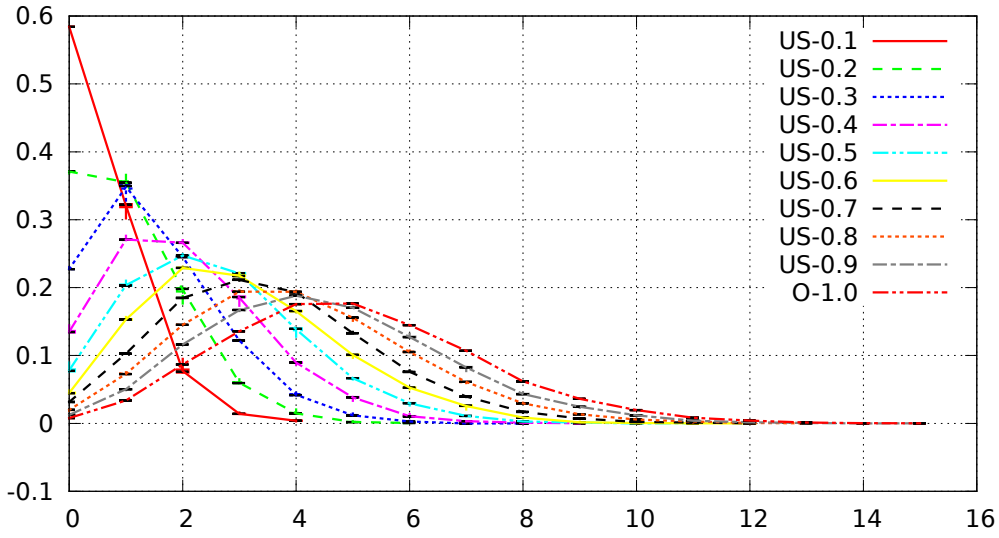
Random Stroll

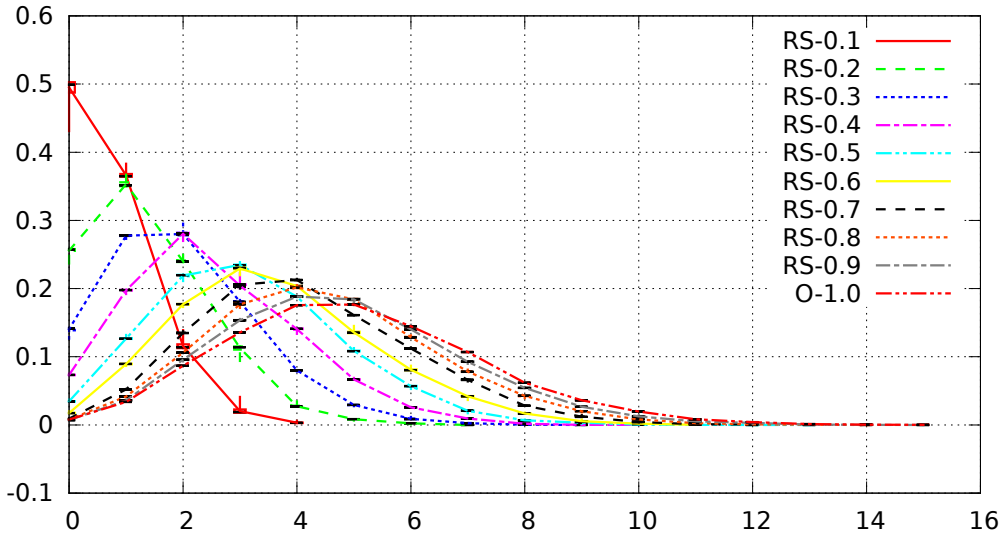
- Developed to randomise the Random Walk

RS: Degree Distribution

- Degree Distribution like the US

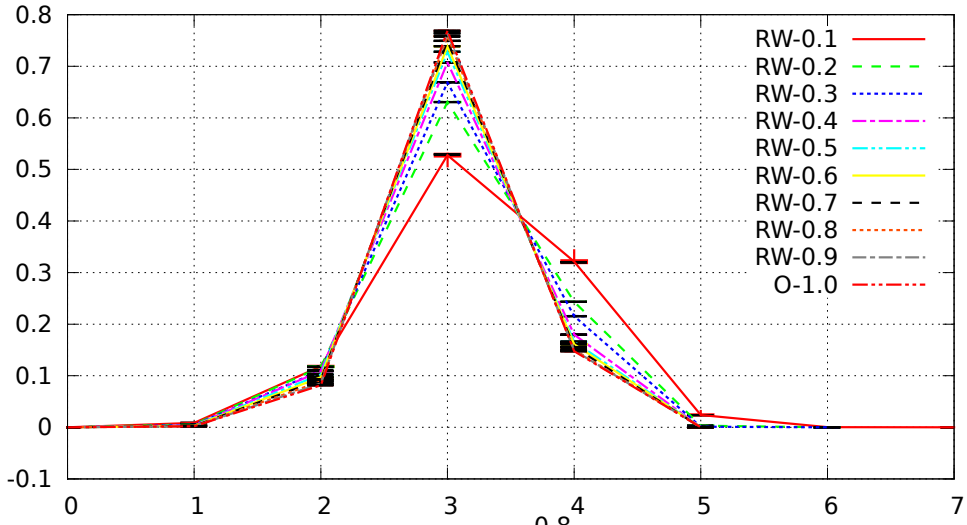


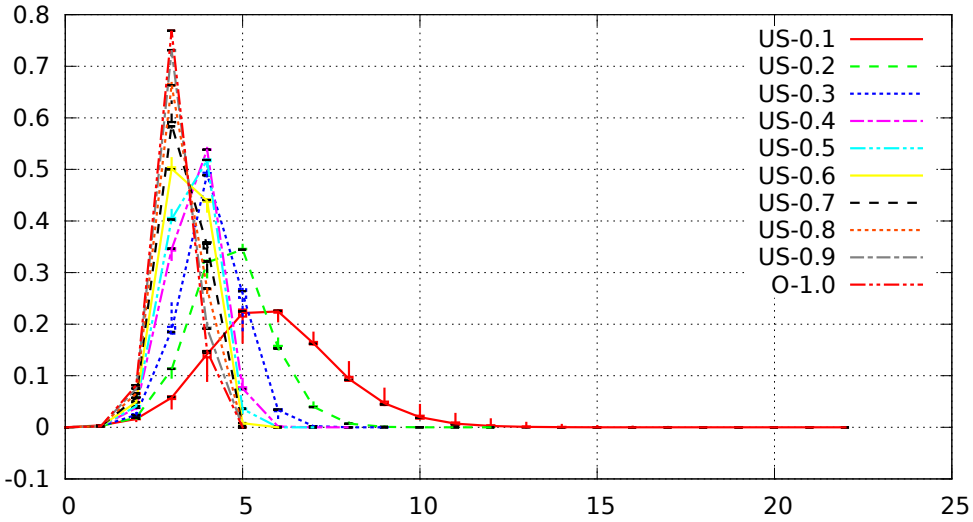


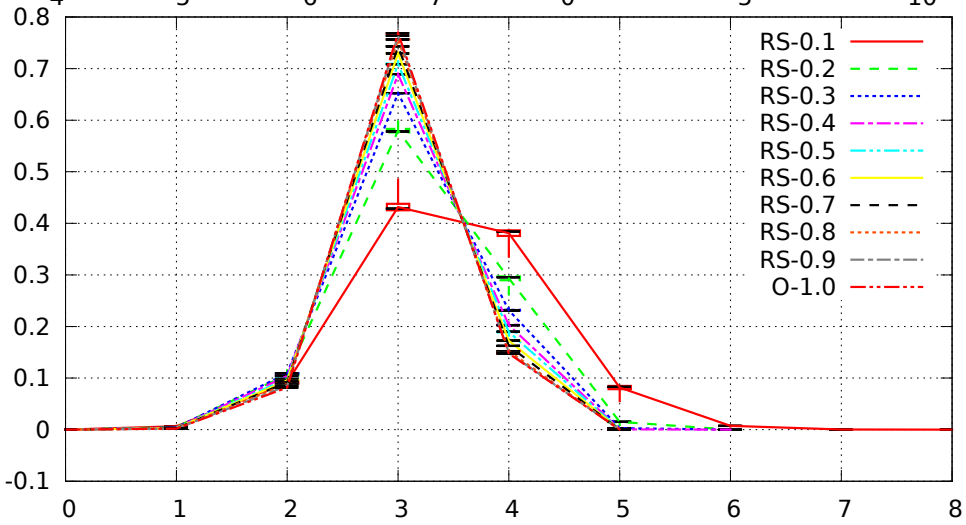


RS: Path Length (1/2)

- Shortest paths like the RW

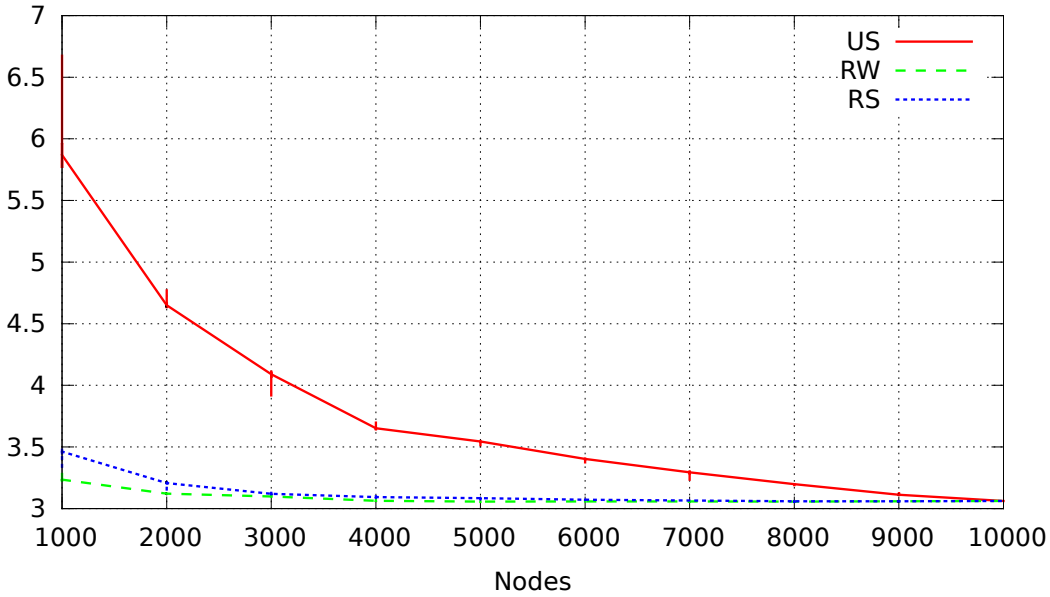


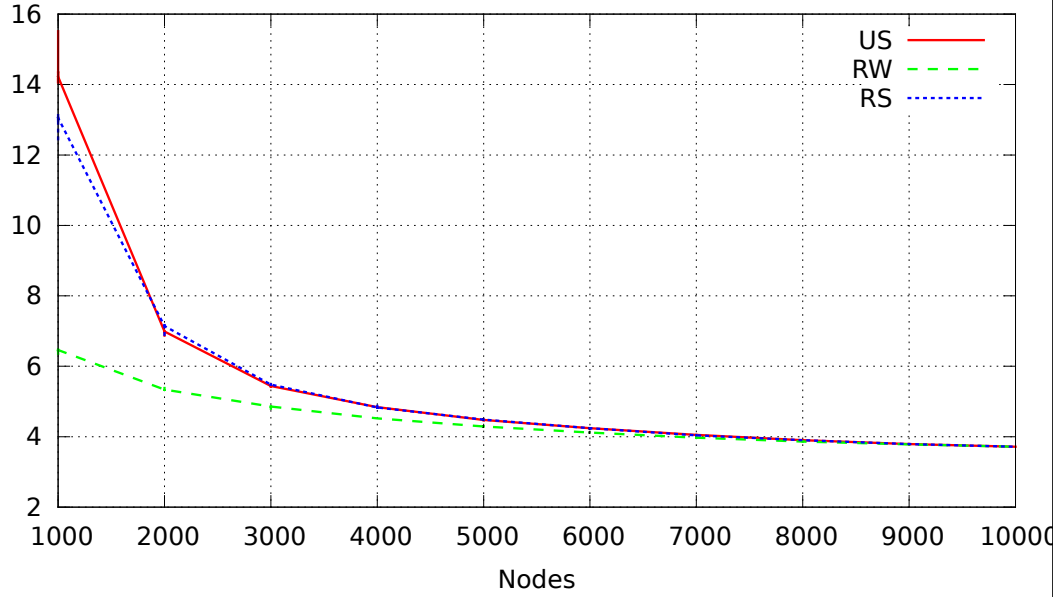




RS: Path Length (2/2)

- Similarity to RW or US is highly depending on:
  - underlying network
  - number of skipped nodes





Random Stroll: between RW and US

- *Developed to randomise the Random Walk*

- „Applicable“ form of „Random Node Selection“ (US) sampling
- Randomness depends on the number of skipped nodes

Results

- Degree correcting algorithms are not sampling „better“
- Complex algorithms are not necessarily better than simple ones
  - some are even worse than the simple ones
- Random Stroll
  - applicable uniform sampling
  - adjustable between Random Walk and Uniform sampling

- Regular networks are hard to sample
- DFS ≈ RW

Conclusion & Future Work

- Simple Algorithms are good enough for the most applications
- Complex algorithms offering special properties
- Consistent sampling is possible

- Open Questions:
  - Similar results on real world networks?
  - Metrics on sampled networks scalable to original size networks?

backup slides

Data base

- 5 networks (ring, clique: 1)
- 16 sampling runs per network instance
- aggregate 16 runs and compute metrics per network instance

Implementation: Components

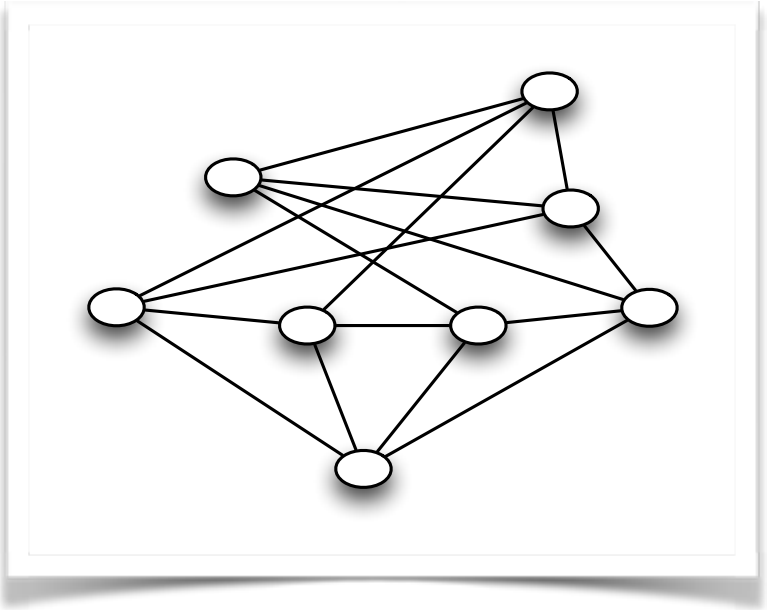
- SamplingController: Coordinate sampling process
- WalkerController: Control the Set of Walkers
- Walker: Walking through the network
- Sampler: Collection nodes to the sample
- CandidateFilter: Filter candidate nodes (self-aware, …)
- StartNodeSelector: Select proper Startnodes

Network models

- Regular network (random topology)
- Random network (Erdös-Rényi)
- Scale-free network (Barabási-Albert)
- Small-world network (Watts-Strogatz)
- Rich-club network (Zhou-Mondragon)
- Special: Ring, Clique

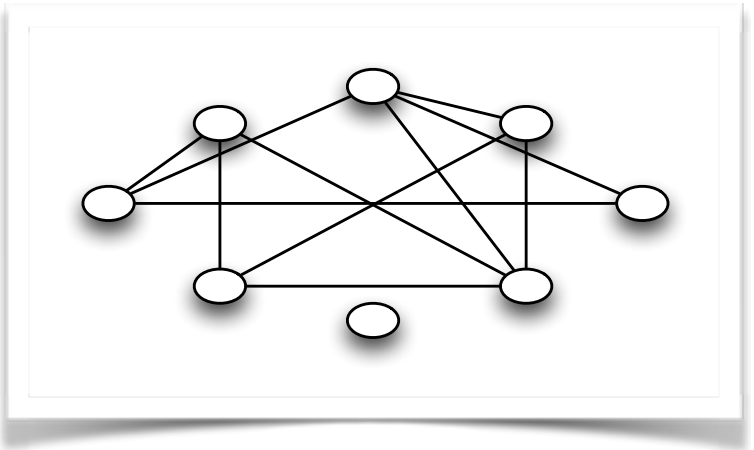
Network: Regular

- Constant Node Degree
- Randomly connected



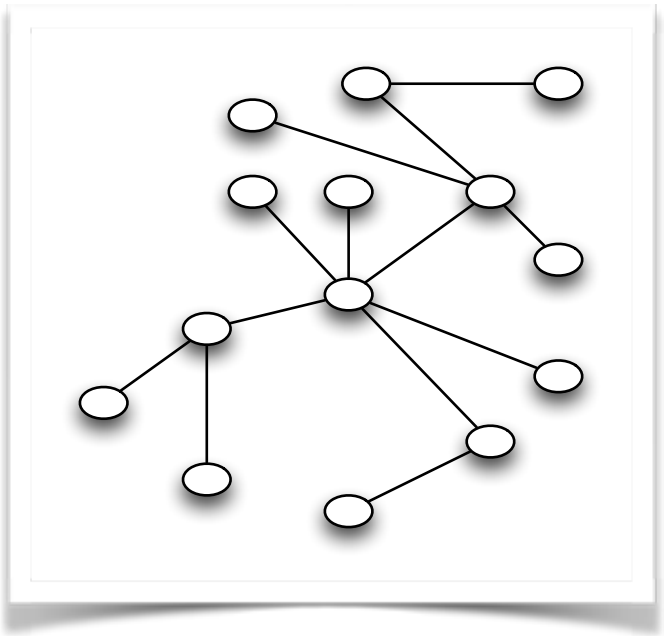
Network: Random

- adapted Erdös-Rényi model
- predefined average degree



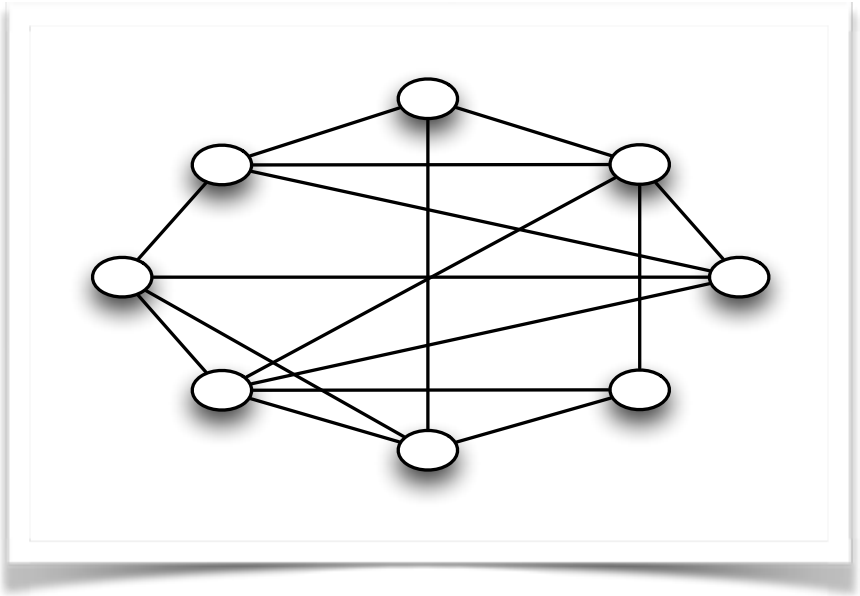
Network: Scale-free

- growth: add nodes iteratively
- preferential attachment: attach likely to other good connected nodes



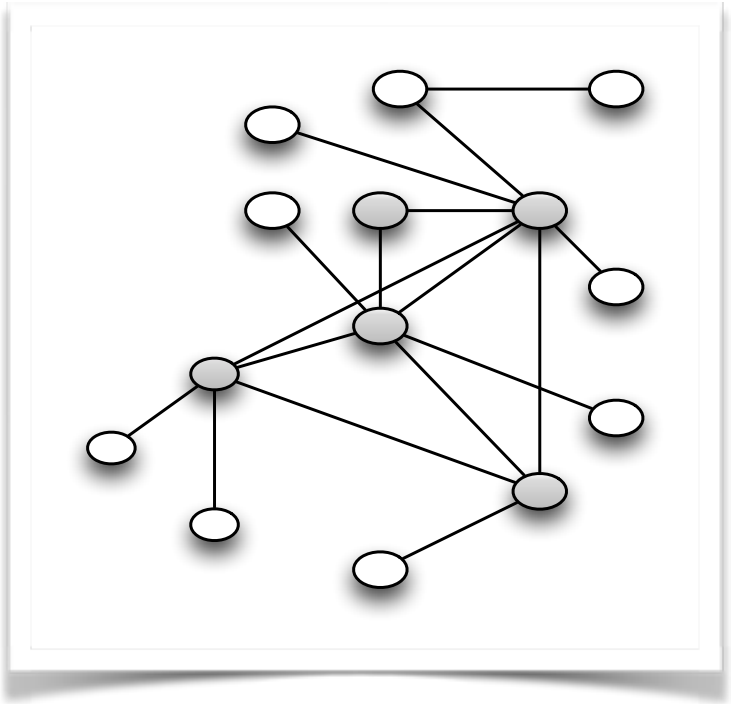
Network: Small-world

- regular ring, rewired edges with probability p



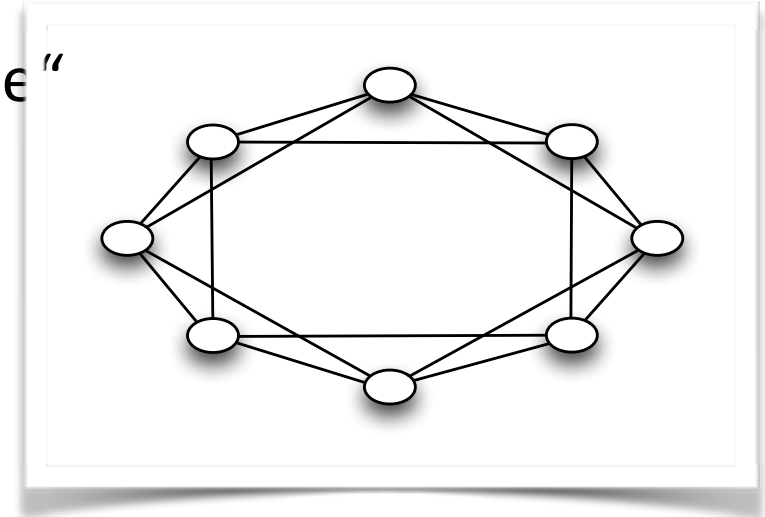
Network: Rich-club

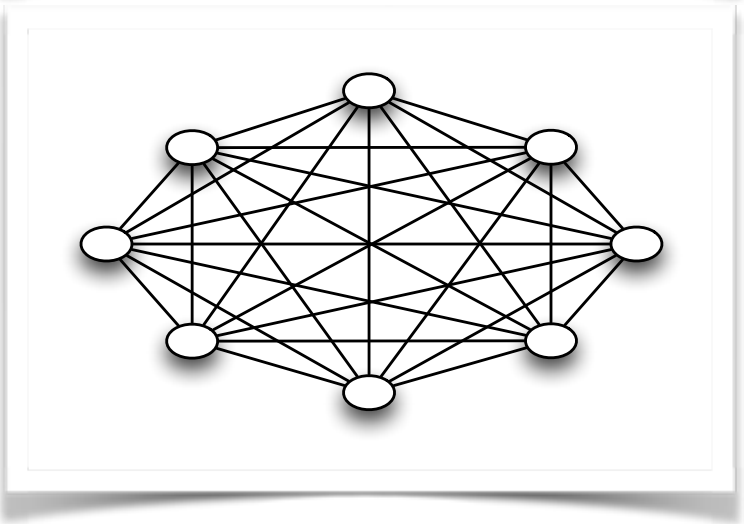
- Rich-club: Set of highest degree nodes
- Maybe scale-free
- Zhou-Mondragon adapted Barabasi-Albert model



Network: Ring, Clique

- Ring is a special regular network topology
- Connected to the k successors
- a fully connected network is „Clique“





Evaluation Setup: Networks

- Regular: k=10
- Random: avgdegree=10
- Scale-free: edgespernode=10
- Small-world: successors=10
- Rich-club: edgespernode=10, p=0.3
- Ring: k=10
- Clique: size=1.000 Nodes (due to computational complexity)

- Size: 10.000 Nodes
- Scaledown: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

Metrics

- Network properties
  - Neighborhood: Degree Distribution, Clustering Coefficient, Assortativity
  - Reach: Hop-Plot, Diameter, effective Diameter, Characteristic Path Length
  - Centrality: Betweenness Centrality, Page Rank
- Sampling algorithm properties
  - Selection Bias, Sample Modularity, Revisit Frequency

Neighborhood/Density

- Degree Distribution
  - Indicates the density of the network
- Average Clustering Coefficient / Transitivity
  - measure for „my friends know each other“
- Assortativity
  - assortative mixing = nodes tend to connect to similar nodes
  - disassortative mixing = nodes tend to connect to unsimilar nodes

Reach

- Hop-Plot
  - measures the reachability within a given number of hops
- Diameter
  - maximum shortest path length, can be used as upper border for some measures
- effective Diameter (90% quantile)
  - similar to the Diameter but not prone to degenerated structures as the 90% quantile of reachable nodes is enough. (shortest path length when 90% of the connected nodes are reachable)
- Characteristic Path Length
  - average shortest path length, can be used as a expectation of hops to reach another node

Centrality

- Betweenness Centrality
  - measures the the centrality of a node by counting the fraction of shortest path including the node
- PageRank Distribution
  - measures the importance of a node by distributing „importance“ through the network

Sampling Algorithm Properties

- Selection Bias
  - shows biases towards „some“ nodes in repeated sampling runs
- Sample Modularity
  - indicates the strength of the division of the sample and the remaining network (adapted modularity metric)
- Revisit Frequency
  - records the number of revisits of nodes in a single sampling run (not analysed as only useful for revisiting sampling algorithms)

Simple or complex Algorithms?

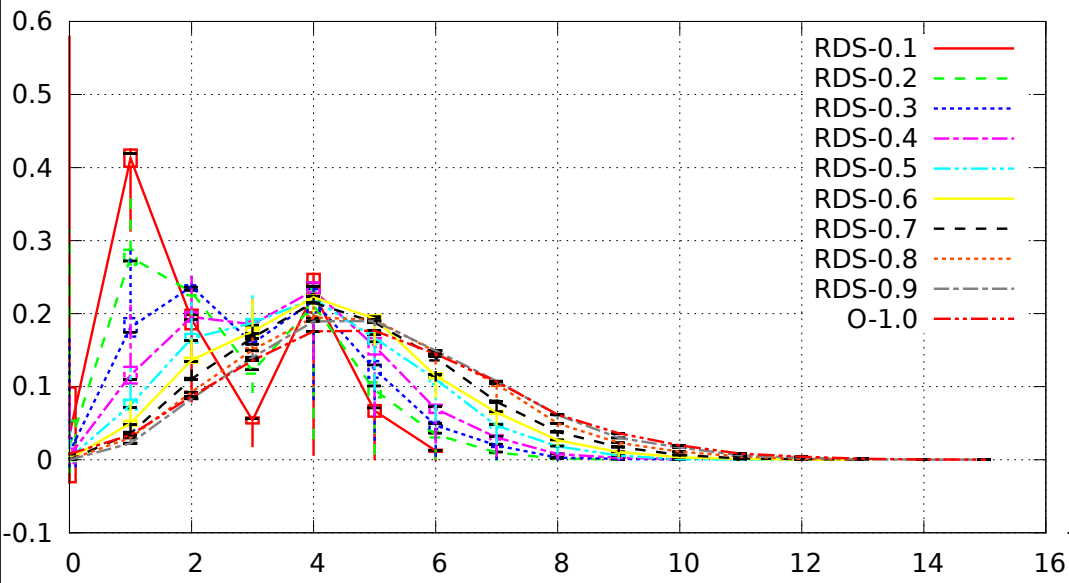
- many improvements suggested for simple sampling algorithms
- We used simple algorithms like RW, BFS and US as a baseline

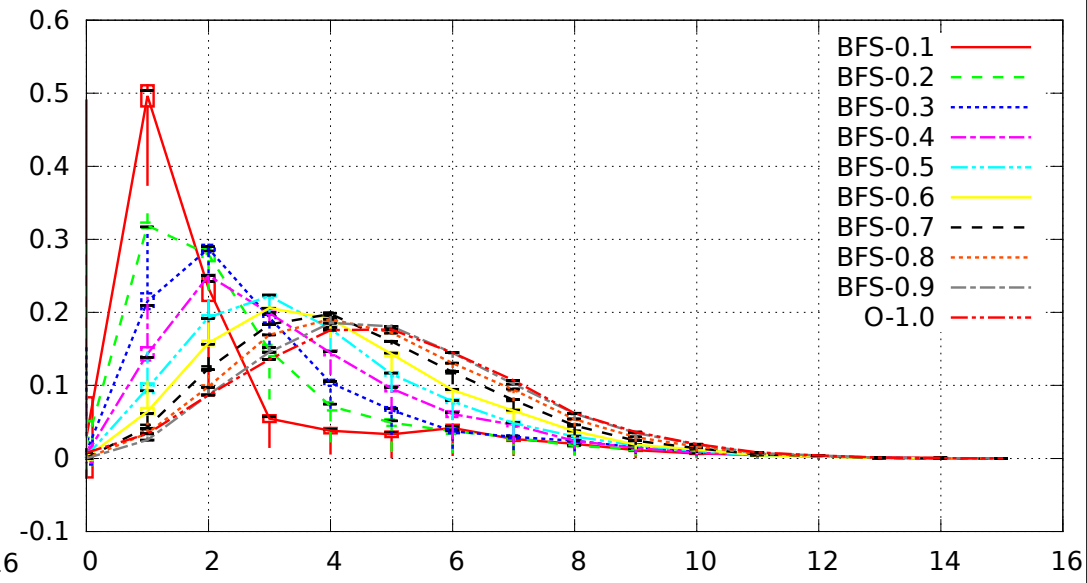
RDS, SS

- Advanced BFS variants
  - select n of k neighbours
  - RDS: select n random neighbours (not self-aware)
  - SS: select the first n unvisited neighbours

RDS, SS: (In) Degree Distribution

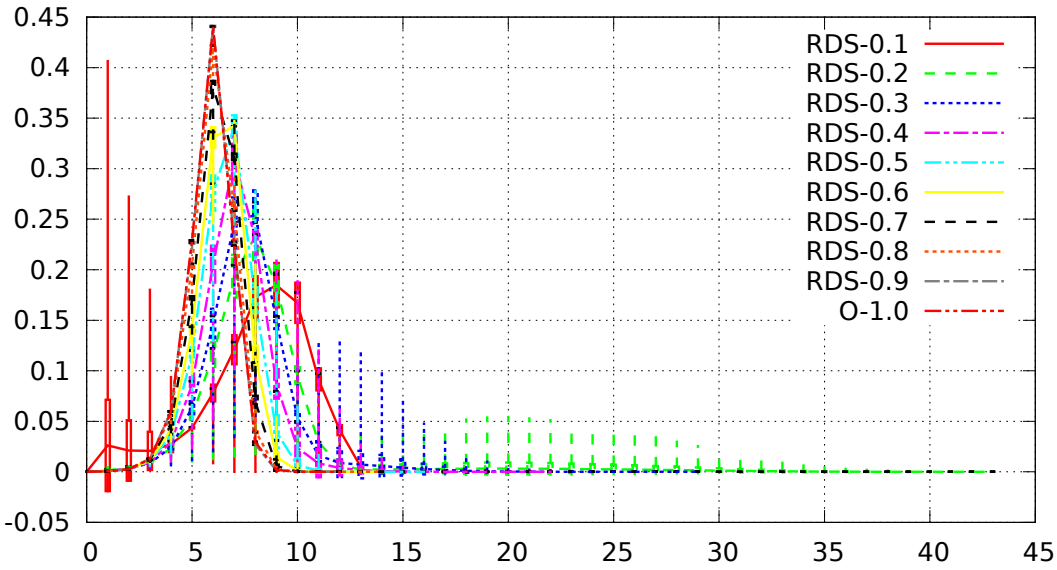
- larger inconsistencies
- additional peaks

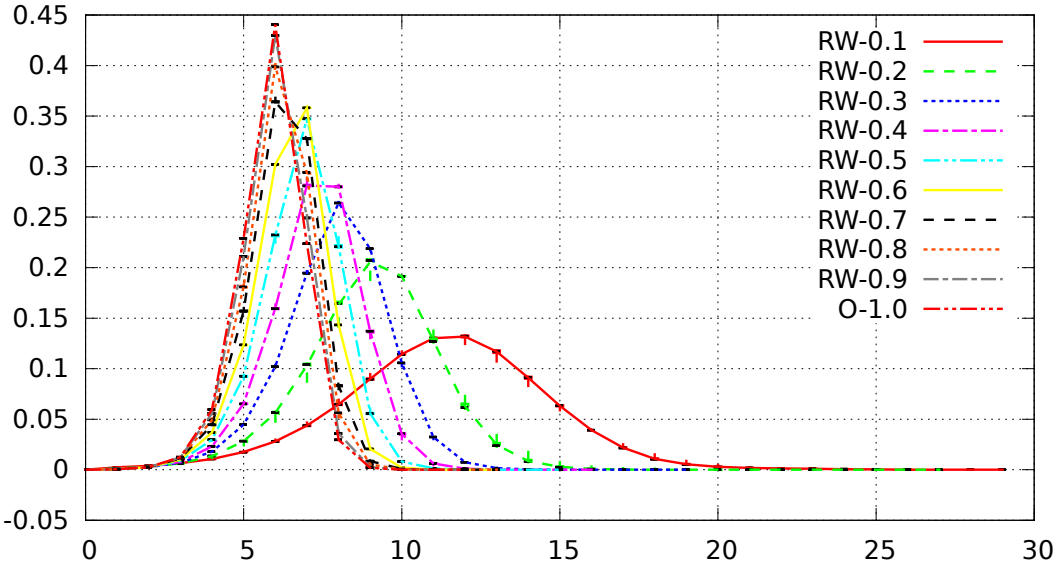




RDS, SS: Shortest Path Lengths

- very large inconsistencies
- longer shortest path lengths





RDS, SS

- Advanced BFS variants

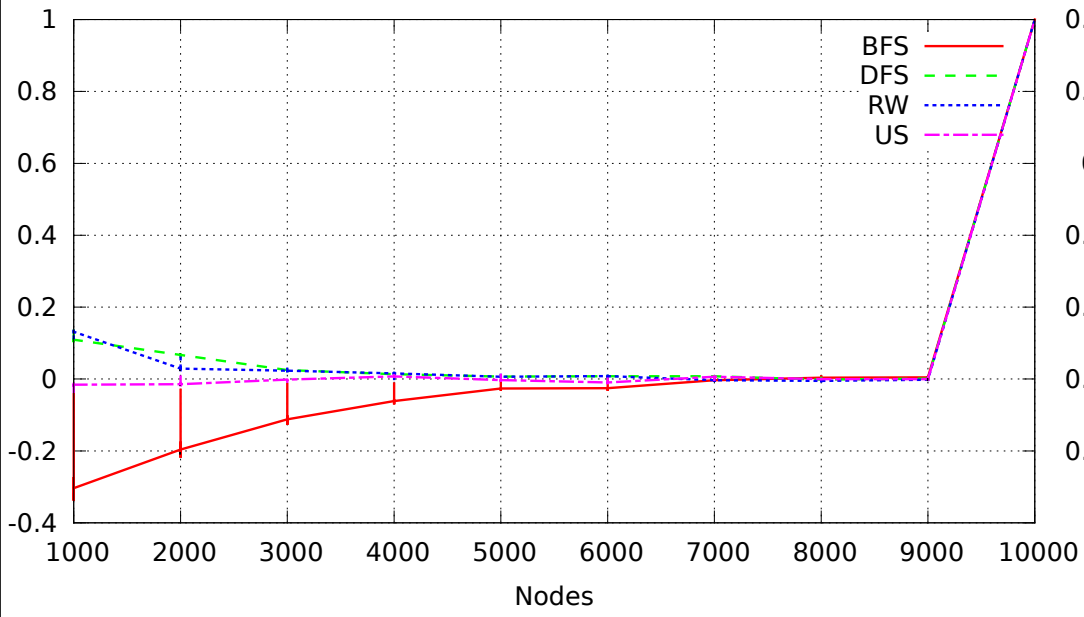
- large inconsistencies
- worse results
  - additional peaks
  - longer paths

Regular Network

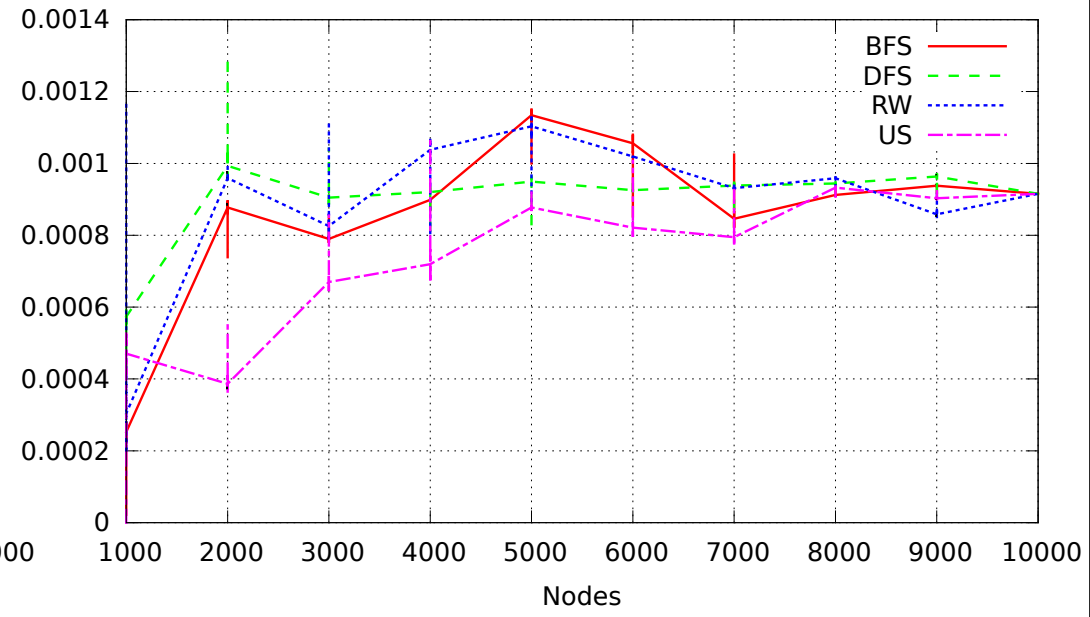
- „special“ structure
  - all nodes have the same degree
  - randomly connected

Regular Network: Assortativity, Clustering Coefficient

- Assortativity coefficient is never converging
- Clustering Coefficient
  - „random“ for small scaledowns
  - but only a small interval of values



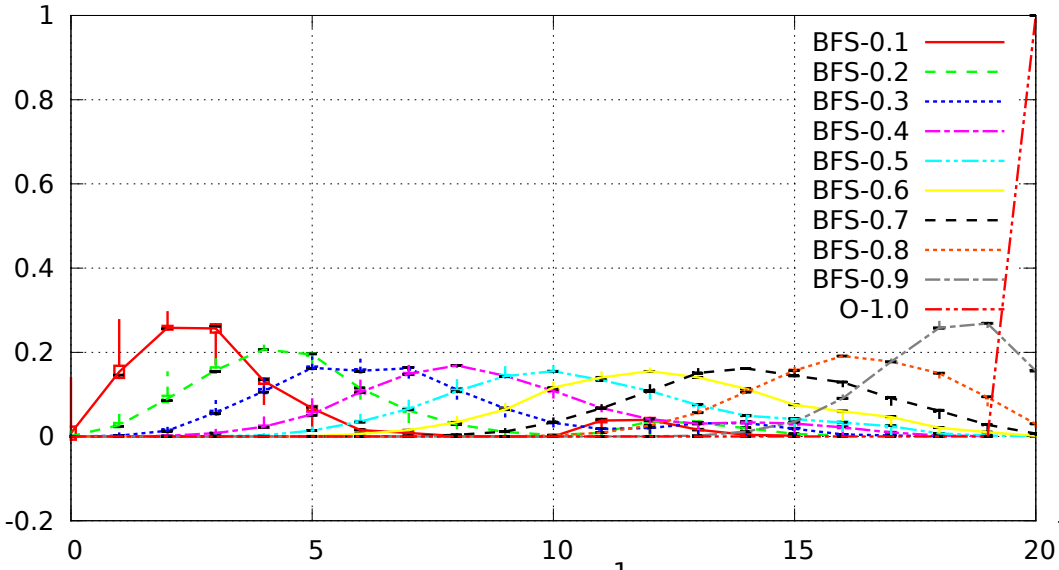
Assortativity Coefficient

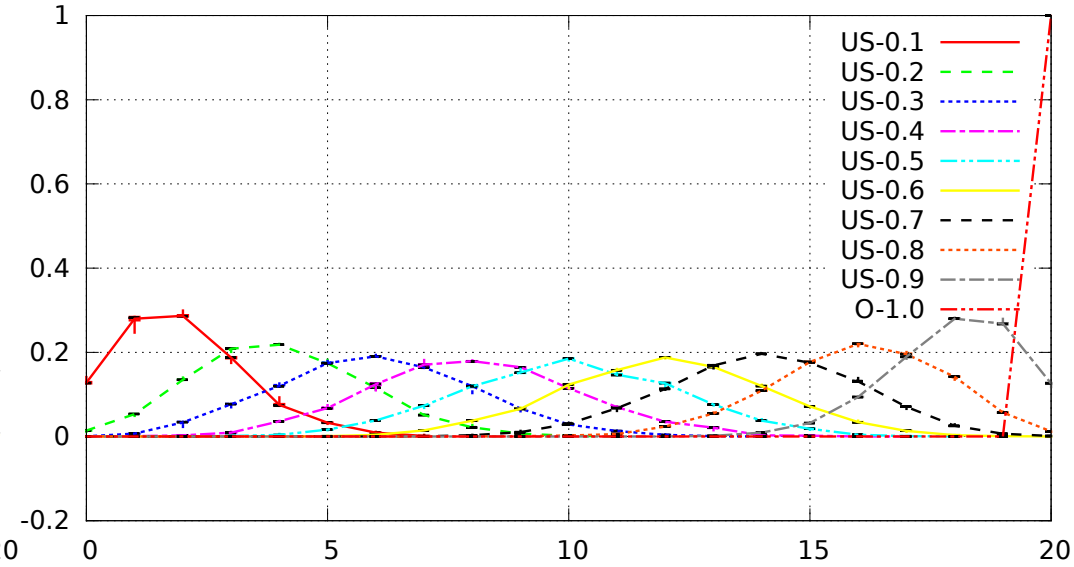


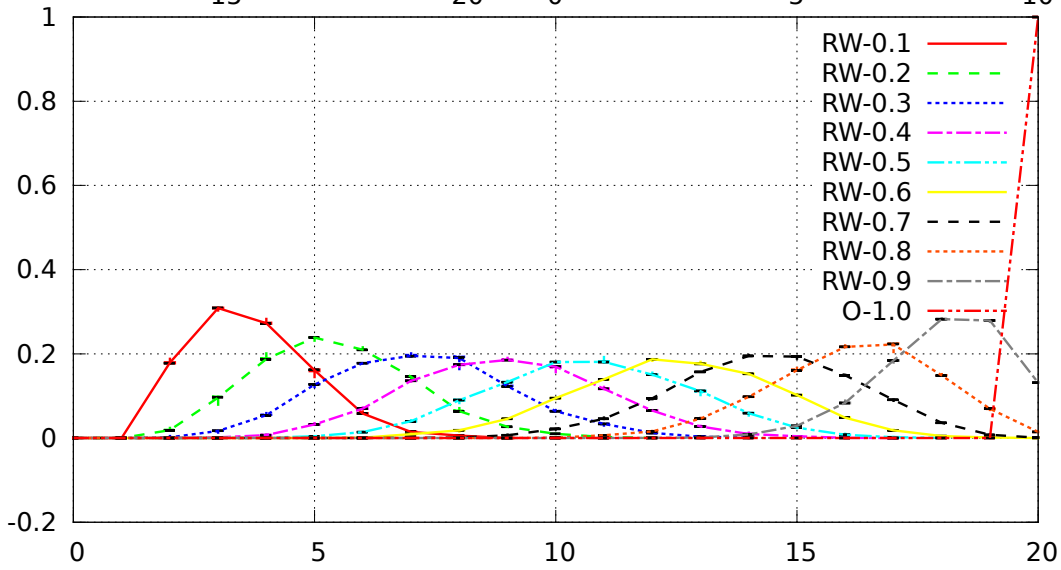
Clustering Coefficient

Regular Network: Degree Distribution

- none of the algorithms is converging







Regular Network

- *„special“ structure*
  - *all nodes have the same degree*
  - *randomly connected*

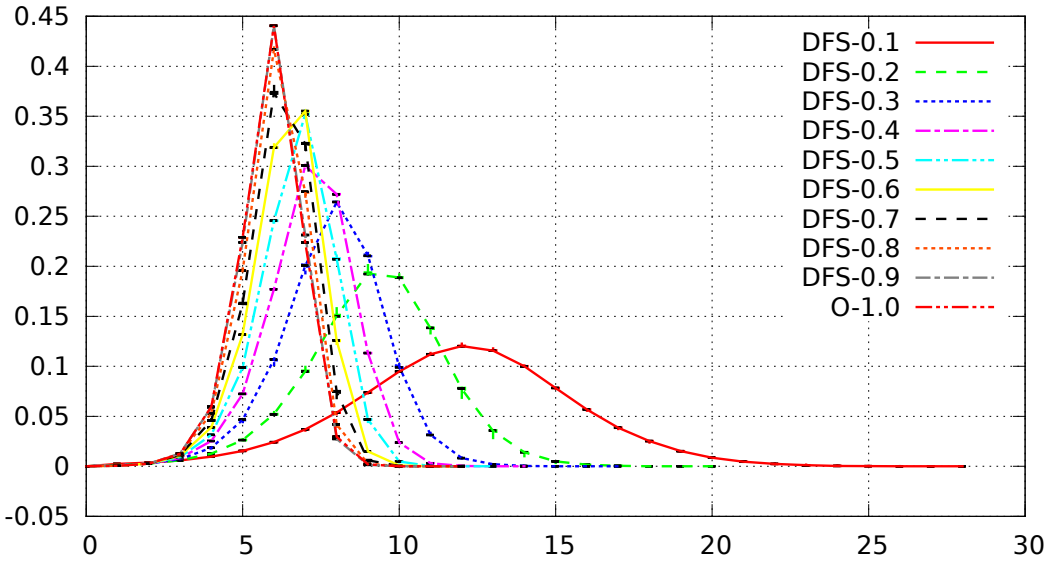
- hard to sample

DFS ≈ RW?

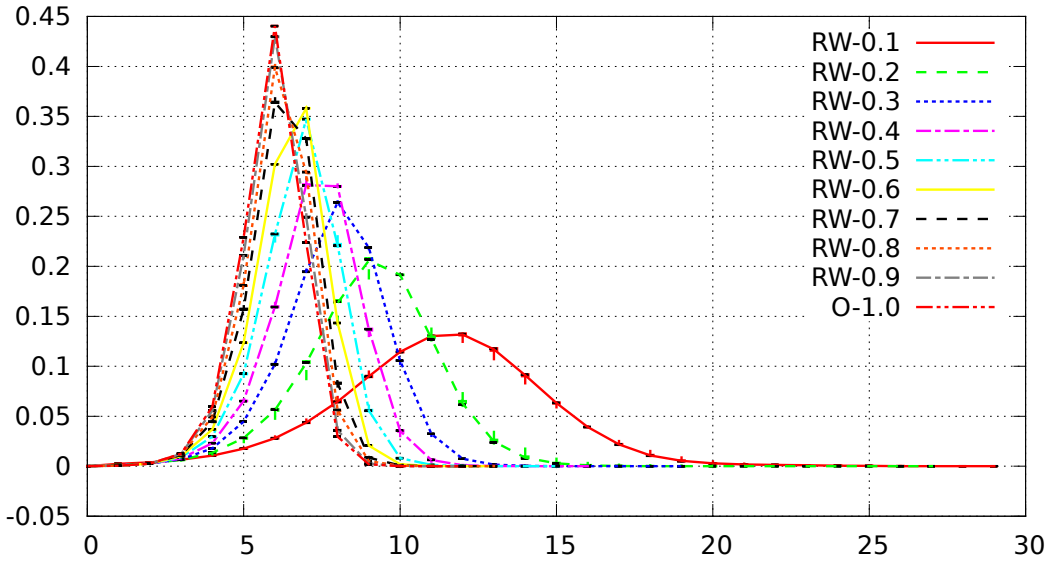
- theoretical analysis shows: DFS and RW are very similar

DFS ≈ RW

- Random network
- similar progressions
- similar confidence intervals



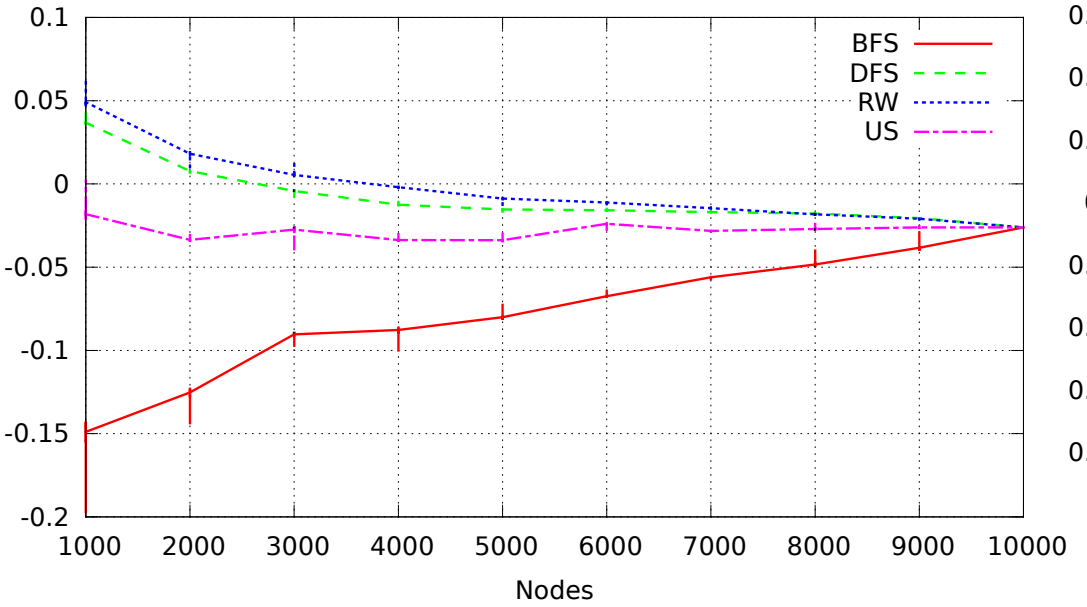
DFS



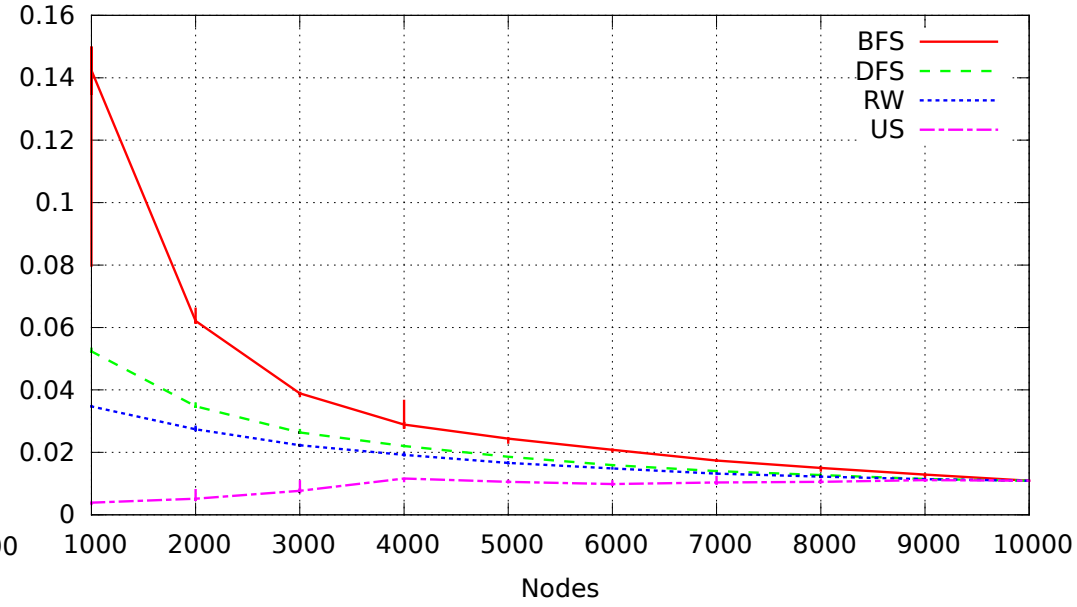
Random Network

DFS ≈ RW: Assortativity, Clustering Coefficient

- Scale-free network
- similar progressions
- similar confidence intervals



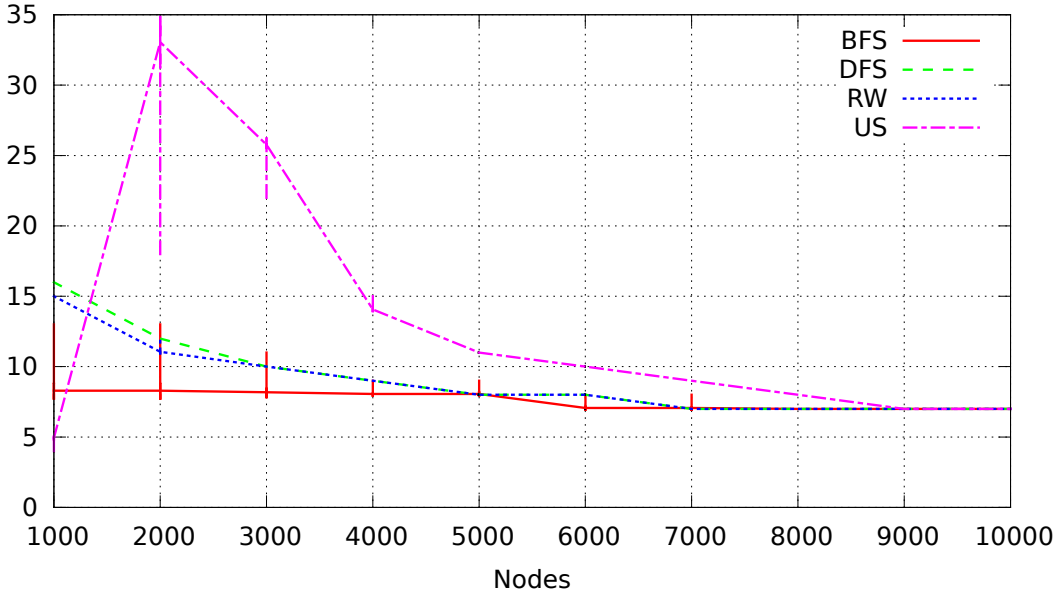
Assortativity Coefficient



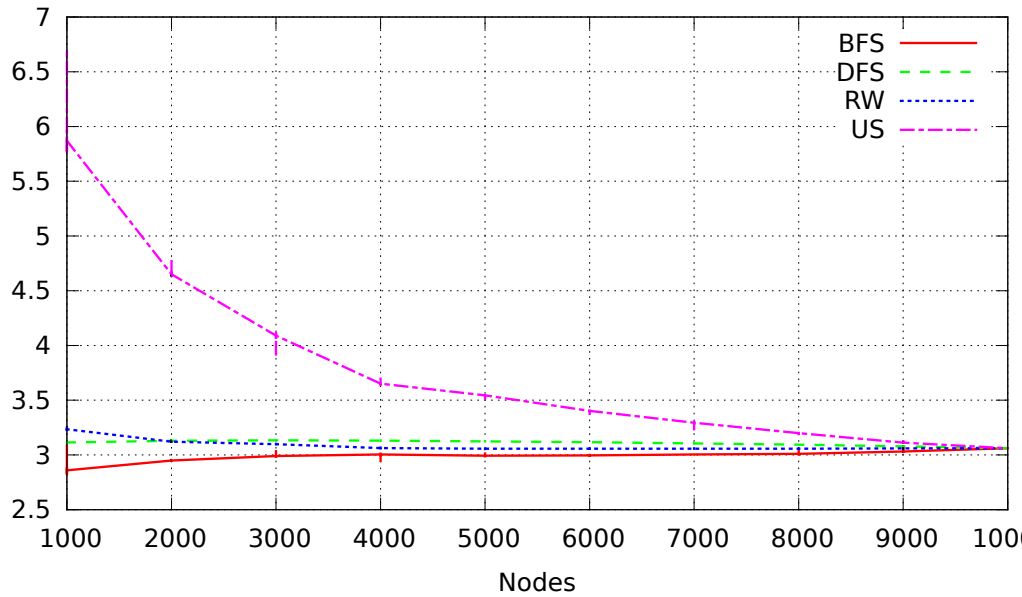
Clustering Coefficient

DFS ≈ RW: Path Lengths

- effective Diameter on a random network
  - nearly equally sampled
- Characteristic path length on a scale-free network
  - nearly equally sampled



effective Diameter (Random)



Characteristic Path Length (Scale-free)

DFS ≈ RW!

- *theoretical analysis shows: DFS and RW are very similar*

- sampling results are nearly equal
- differences are caused by the sorted adjacency lists in GTNA
  - DFS walks towards „small-id“ nodes / communities
  - adjacency lists are not necessarily sorted in reality