IEEE SocialCom 2018
December 2018

Comparing Graph Sampling Methods Based on the Number of Queries

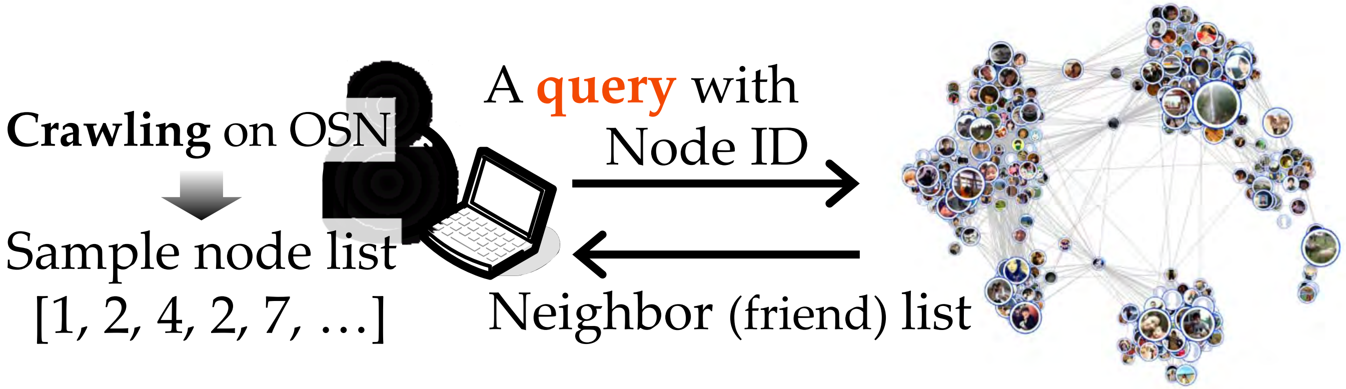
Kenta Iwasaki, **Kazuyuki Shudo**
Tokyo Institute of Technology

岩﨑 謙汰, **首藤 一幸**
東京工業大学

Tokyo Tech

Graph sampling ⊃ Crawling ⊃ Random walk

- They enable estimation of nodal and topological properties of online social networks (OSNs)
  - Effective because the entire network is not available.
  - Properties: Degree distribution, clustering coefficient, …
  - Note: **Crawling** (e.g. random walk) is possible but uniform sampling is not.



Crawling on OSN

Sample node list
[1, 2, 4, 2, 7, …]

A **query** with Node ID

Neighbor (friend) list

- **Query** can be the bottleneck of the sampling performance due to
  - API limits
  - Communication latency is much larger than computation.

Contribution:
Query number standard

- Problem
  - **Sample size** has been the standard to evaluate graph sampling techniques.

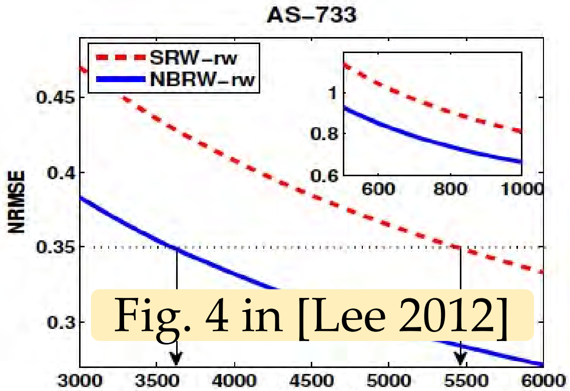


Fig. 4 in [Lee 2012]

# of samples

Standards in studies	
Length of sample node list (walk length)	[Rasti 2009] [Riberio 2010]
Length of sample node list ???	[Lee 2012] [Hardiman 2013]
Number of sample nodes	[Gjoka 2011]

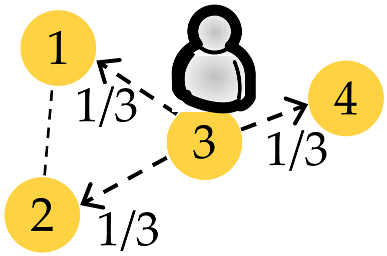
Even not clear !!

- Contribution
  - **Query number** based comparison shows different relative merits for sampling and estimation techniques.
  - It reflects graph accessing cost better.

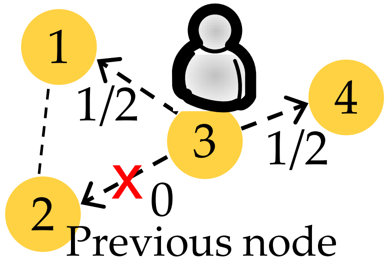
Graph sampling techniques

- Random walk-based techniques are effective for property estimation for OSNs
  - They enable unbiased sampling with Markov chain analysis.
- Our targets
  - SRW-rw : Simple random walk w/ re-weighting
  - NBRW-rw : Non-backtracking random walk w/ re-weighting
  - MHRW : Metropolis-Hastings random walk

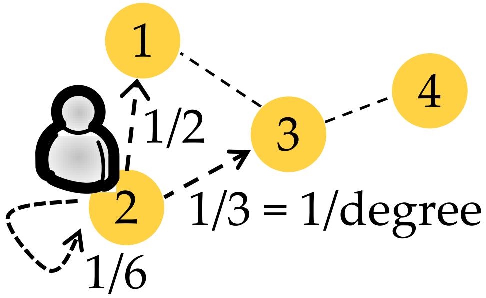
Postprocess to remove bias due to degree



SRW:
Simple
random walk



NBRW:
Non-backtracking
random walk

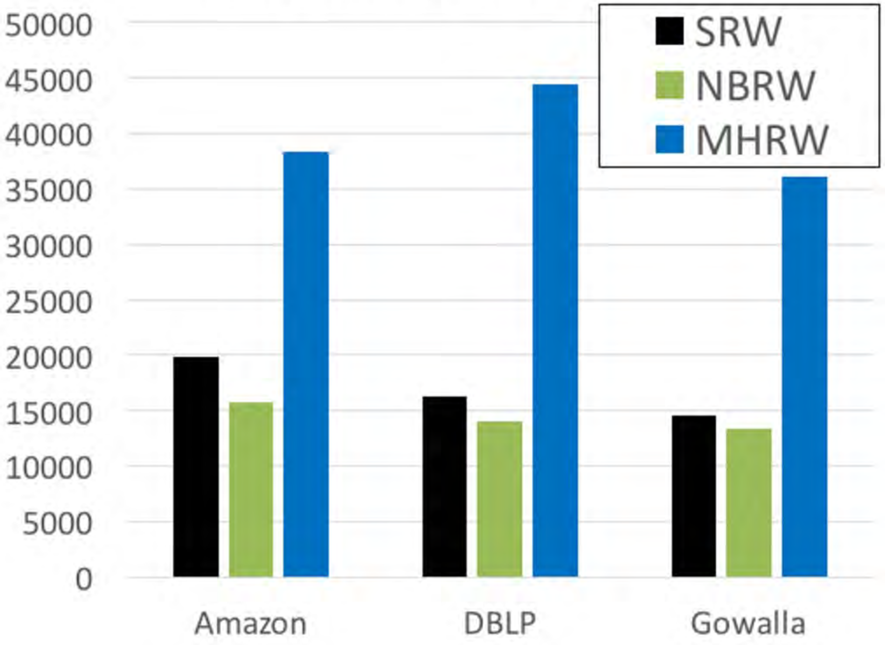


MHRW:
Metropolis-Hastings
random walk

Sample size vs. query number

- Very different

Sample size (length of sample node list) by 10,000 queries



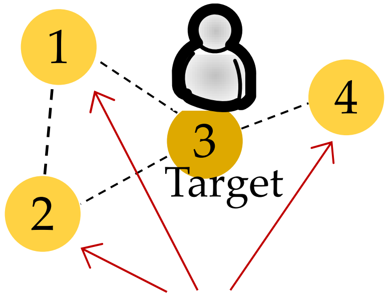
Simple
Non-backtracking
Metropolis-Hastings

Graphs are in
Stanford Large Network
Dataset Collection

- Rationale: MHRW can stay the same node and the length of sample node list grows without a query.
- Note that not only the sample size determines estimation efficiency.
  E.g. NBRW reaches various nodes and it is better with Counting Triangles [Iwasaki 2018].

Query issuing timings

1. For random walk
   - When getting neighbor (friend) list of the next hop ☺
2. For property estimation
   - Depends on each estimation technique
   - E.g. When getting neighbor (friend) list of multiple neighbor nodes ☹ of a node to calculate clustering coefficient of the node naively.



It is necessary to know how the neighbor nodes connected each other to calculate cluster coefficient.

Experiments

with sample size and query number standards

- Clustering coefficient estimated
- Estimation efficiency (precision / cost) compared on

1. Estimation techniques:
   Naïve method vs. **Counting Triangles** [Hardiman 2013]
   Counting Triangle does not require additional queries for property estimation.
2. Sampling (random walk) techniques:
   SRW vs. **NBRW** vs. **MHRW**

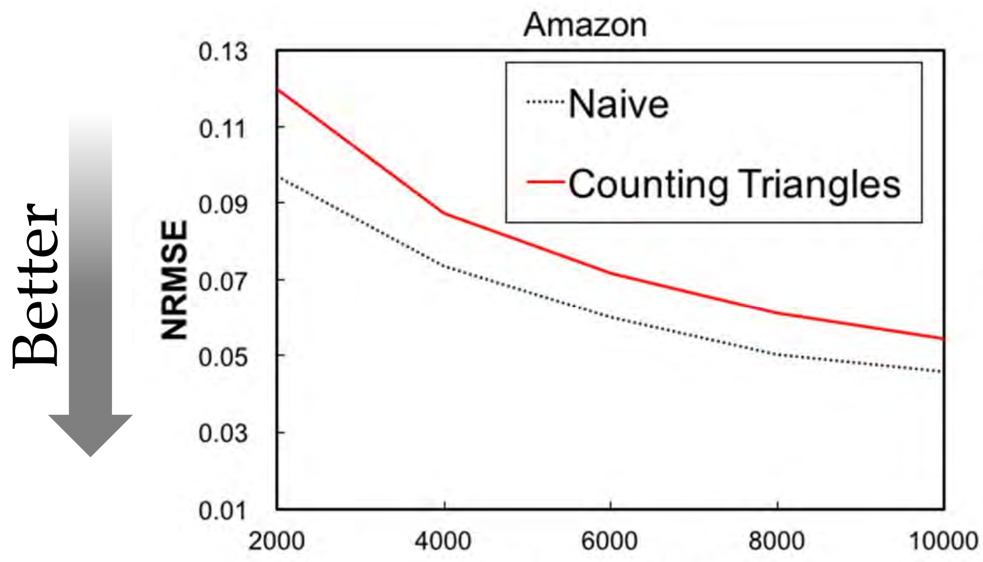
Graph	# of nodes	Average degree	Average Clust. Coeff.
Amazon	334,863	5.530	0.3967
DBLP	317,080	6.622	0.6324
Gowalla	196,591	9.668	0.2367

in Stanford Large Network Dataset Collection

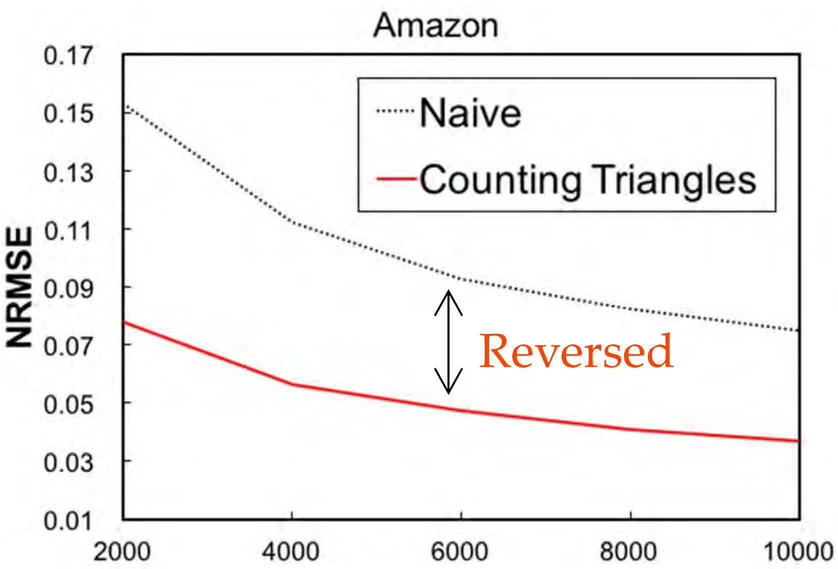
Naïve method vs. Counting Triangles

[Hardiman 2013]

- Sampling with **simple random walk (SRW)**
- Relative merits are reversed.
  - The similar results shown with the other networks.



Sample size



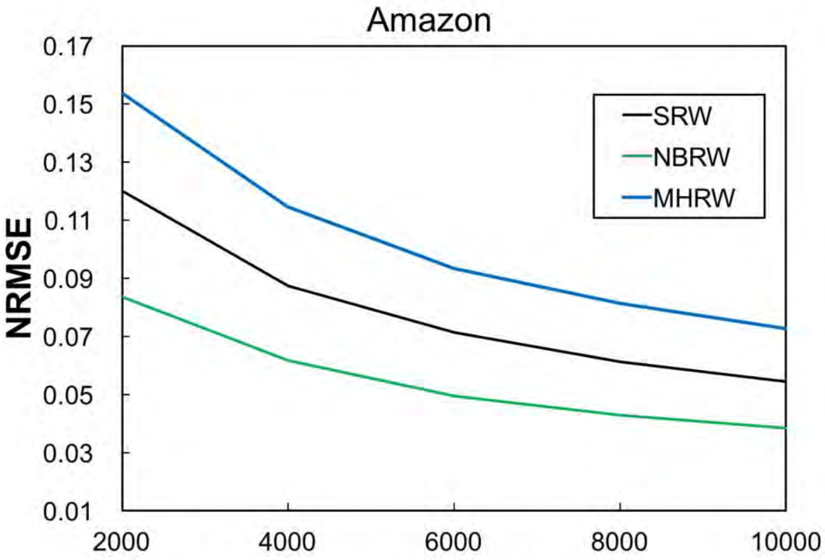
Reversed

Query number

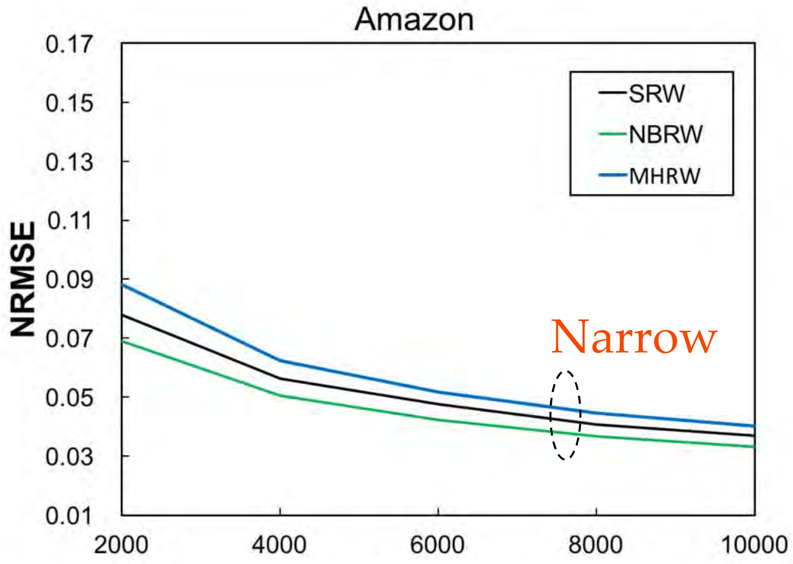
SRW vs. NBRW vs. MHRW

- Estimating with **Counting Triangles**
- Margins are much narrowed.

Better



Sample size



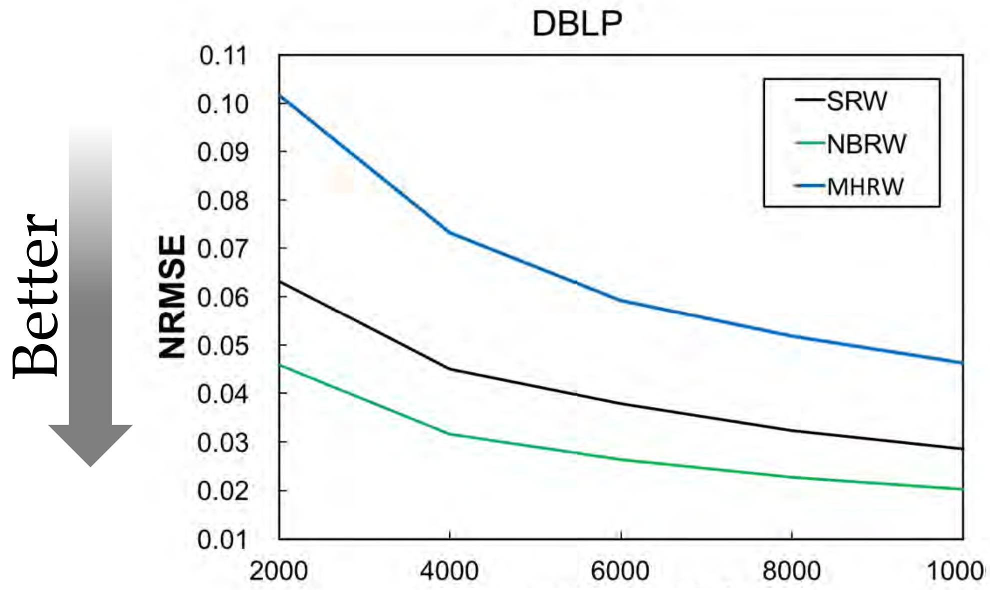
Narrow

Query number

- Note: Our contribution includes Counting Triangles with MHRW.

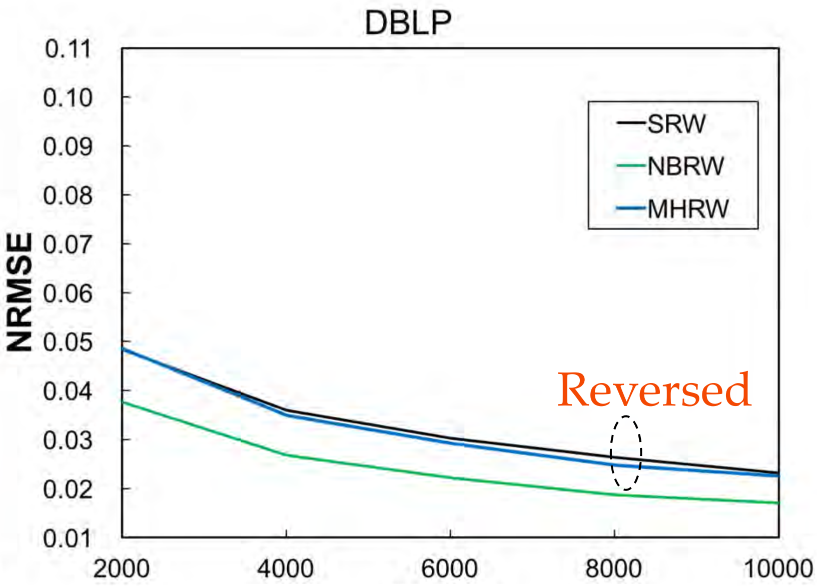
SRW vs. NBRW vs. MHRW

- Estimating with **Counting Triangles**
- Relative merits are reversed for DBLP graph.



Better

Sample size



Reversed

Query number

Summary

- **Query number** standard Cf. sample size standard
  - for comparing graph sampling techniques
  - for comparing property estimation techniques
  - It reflects graph accessing cost better.
    - Accessing online social networks
    - Accessing a graph on storage and memory

- The two standards showed different relative merits for techniques.

Tokyo Tech