IEEE SocialCom 2019
December 16th - 18th 2019

Metropolis-Hastings Random Walk with a Reduced Number of Self-Loops

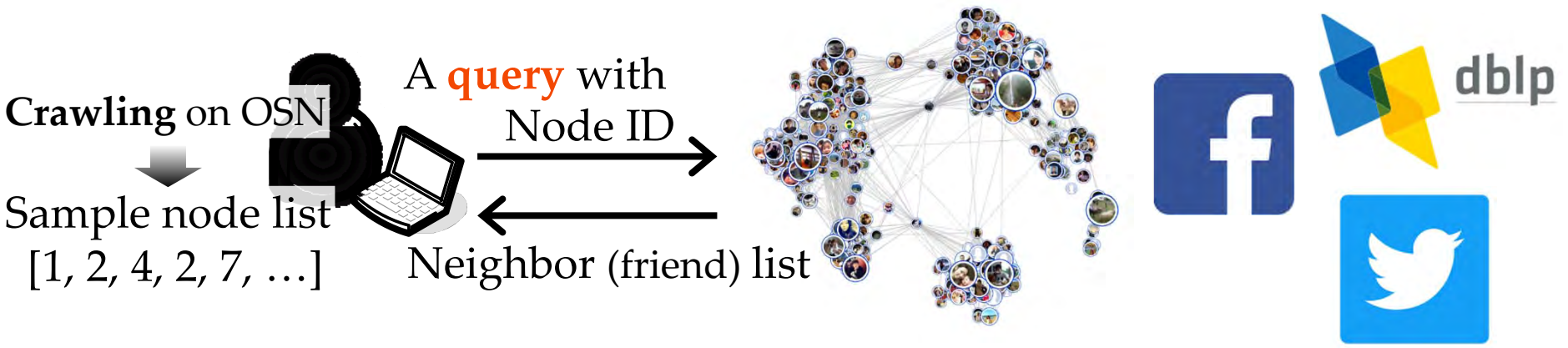
Toshiki Matsumura, **Kazuyuki Shudo**
Tokyo Institute of Technology

松村 俊樹, **首藤 一幸**
東京工業大学

Tokyo Tech

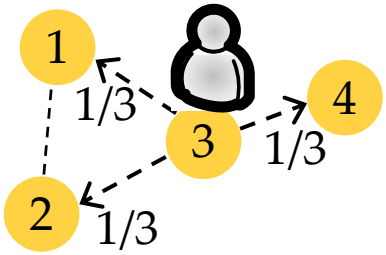
Graph sampling ⊃ Crawling ⊃ Random walk

- 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.

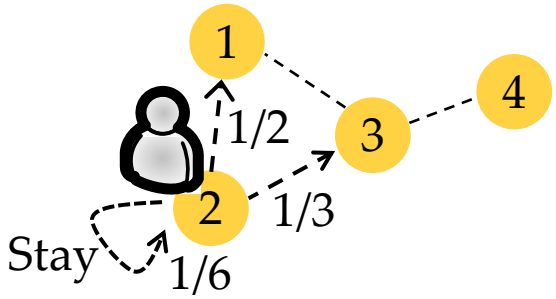


Random walk-based techniques

- They are effective for property estimation for OSNs
  - They enable unbiased sampling with Markov chain analysis.



Simple random walk (SRW)



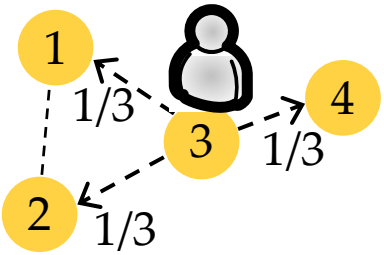
Metropolis-Hastings random walk (MHRW)

Target of this research

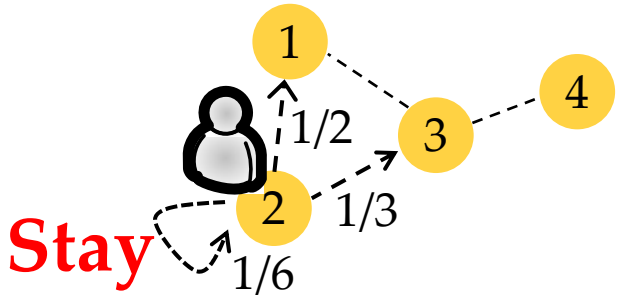
- Stationary distribution of each node
  - SRW – proportional to the degree
    - requires postprocess to remove the bias ☹
  - MHRW – uniform
    - Ready-to-use ☺ without postprocess

Contribution:
Faster spread of the crawler

- Pros and cons of MHRW
  - Pros: Ready-to-use ☺
  - Cons: Slow spread of the crawler ☹ due to "stays"



Simple random walk (SRW)



Metropolis-Hastings random walk (MHRW)

- Contribution: **Faster spread of the crawler**
  - Not only based on # of steps, but also # of queries

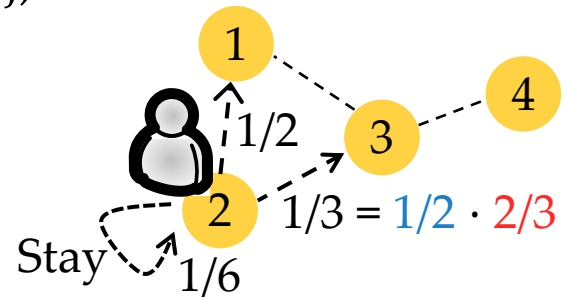
Proposed technique:
Regulate probabilities

- A single step of MHRW
  - (1) Choose a candidate of the next hop with Q_{ij}
  - (2) Decide whether or not to accept it with A_{ij}

    "Stay" if rejected

- Transition probability p_{ij} (from i to j) of MHRW

$$p_{ij} = \begin{cases} \frac{1}{d_i} \cdot \min\left(1, \frac{d_i}{d_j}\right) = Q_{ij}A_{ij} & j \in N(i) \\ 1 - \sum_{k \in N(i)} p_{ik} & i = j \\ 0 & \text{otherwise} \end{cases}$$



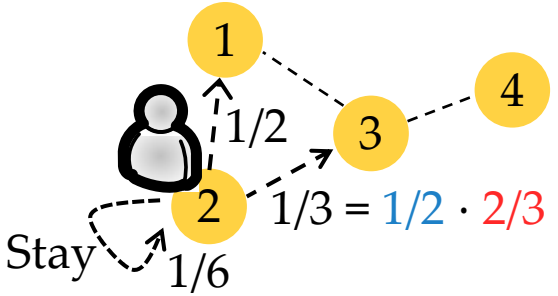
Metropolis-Hastings random walk (MHRW)

–Product of
- Choice probability Q_{ij} and
- Acception probability A_{ij}

Proposed technique:
Regulate probabilities

- Transition probability p_{ij} of MHRW

$$p_{ij} = \begin{cases} \frac{1}{d_i} \cdot \min\left(1, \frac{d_i}{d_j}\right) = Q_{ij} A_{ij} & j \in N(i) \\ 1 - \sum_{k \in N(i)} p_{ik} & i = j \\ 0 & \text{otherwise} \end{cases}$$



Metropolis-Hastings random walk (MHRW)

- Problem
  - Low acception probability A_{ij} when transiting to a high-degree node
- Solution
  - Lower choice probability Q_{ij} to high-degree nodes while keeping the same transition probability p_{ij}
    - Rejections decrease.

Choice probabilities

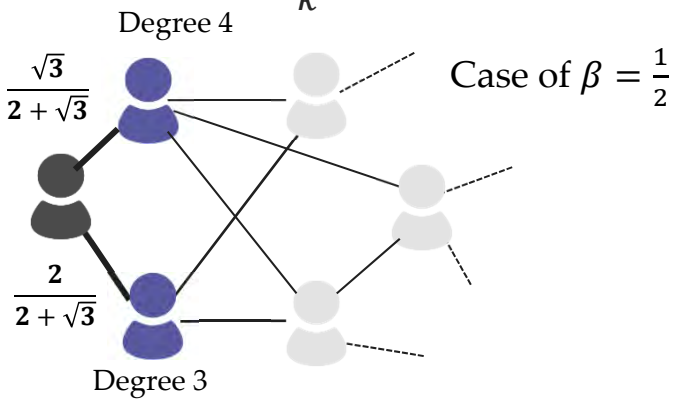
- Two kinds of Q_{ij} tried

cf. $p_{ij} = Q_{ij} \cdot A_{ij}$

  - giving lower probabilities to high-degree nodes

β-MHRW

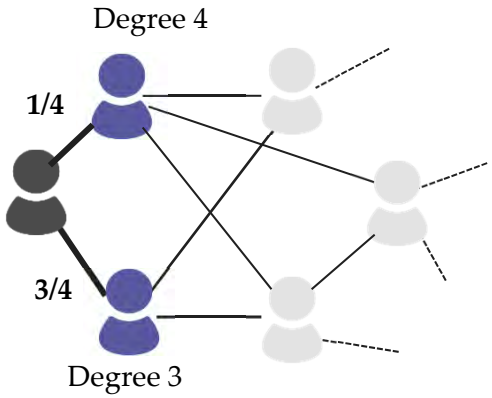
$$Q_{ij} = \frac{\frac{1}{d_j^{\beta}}}{\sum_{k \in N(i)} \frac{1}{d_k^{\beta}}}$$



Case of $\beta = \frac{1}{2}$

Reselecting-MHRW

$$Q_{ij} = \frac{2 \cdot \#\{d_j < d_k\}_{k \in N(i)} + \#\{d_j = d_k\}_{k \in N(i)}}{d_i^2}$$



Choice probability of
β-Random Walk [Hosaka 2012]

Evaluation

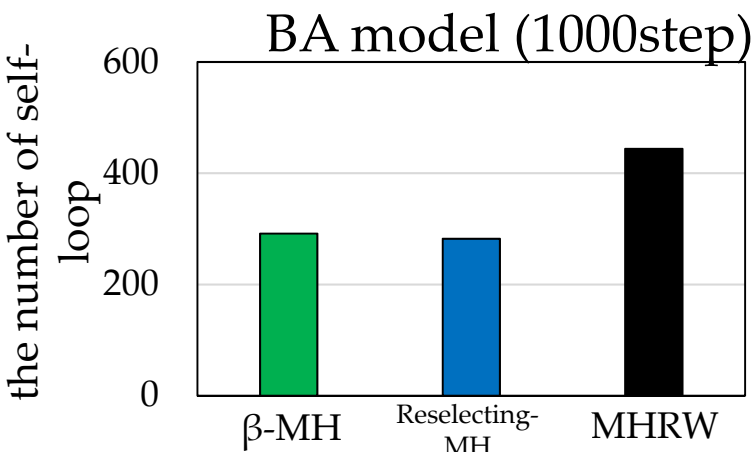
- Target networks

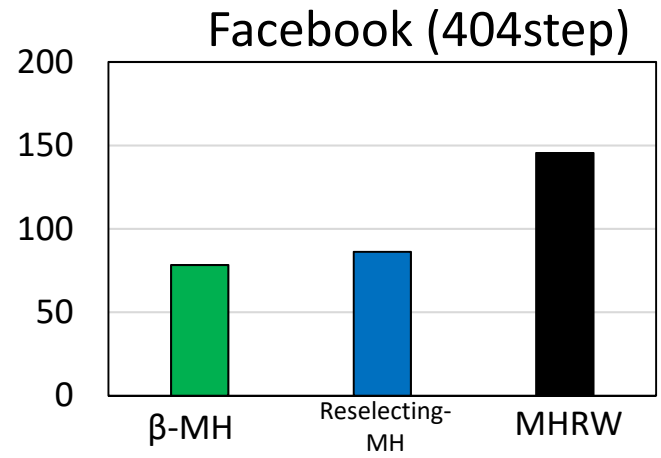
Network	# of nodes	# of vertices	Average degree
Barabasi-Albert model	10,000	29,991	5.998
Facebook (*)	4,039	88,234	43.691
DBLP (*)	317,080	1,049,866	6.622
Amazon (*)	334,863	925,872	5.530

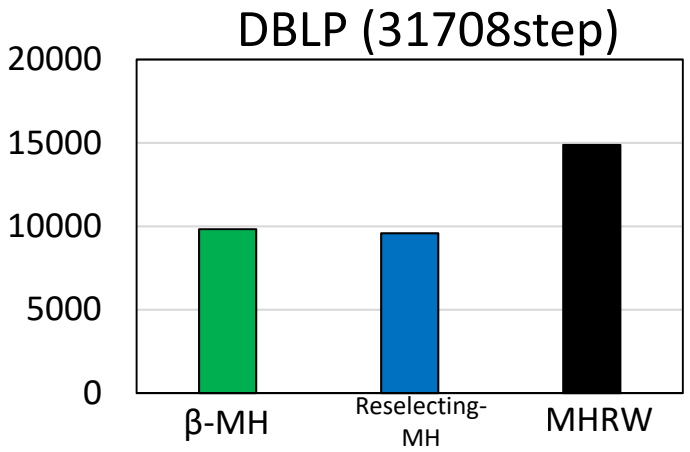
(*) SNAP: https://snap.stanford.edu/data/

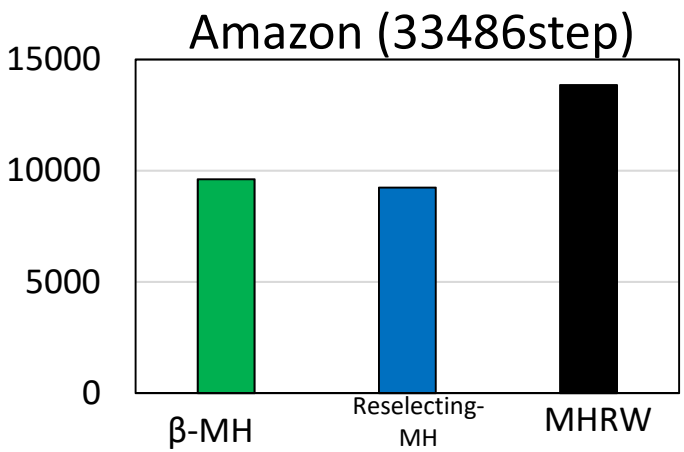
- Metrics
  1. Number of "stays"
  2. Uniformity of the sampled node list
     - L^1 distance

Metrics 1 / 2:
Number of "stays"



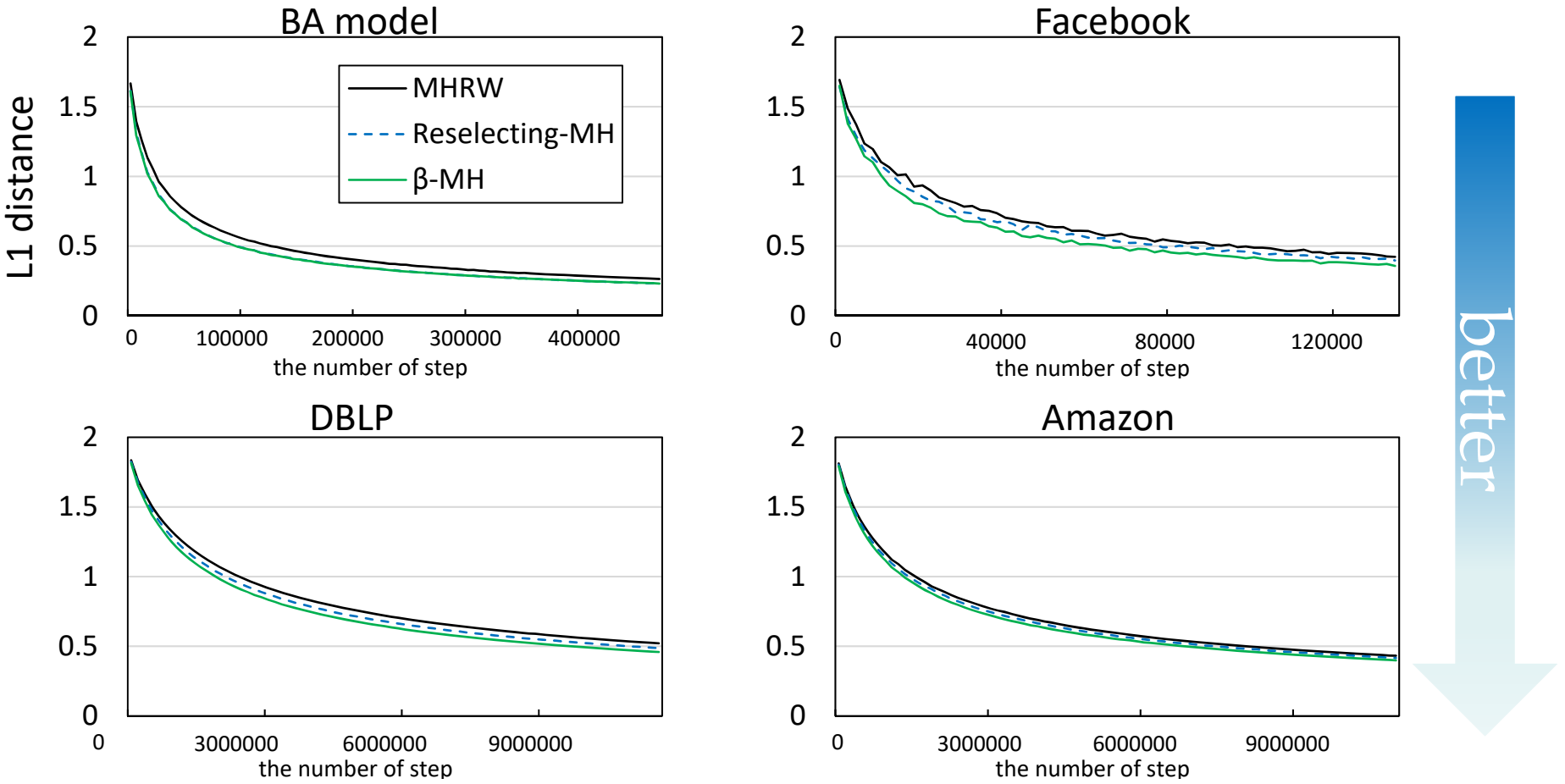






- Reduced ☺

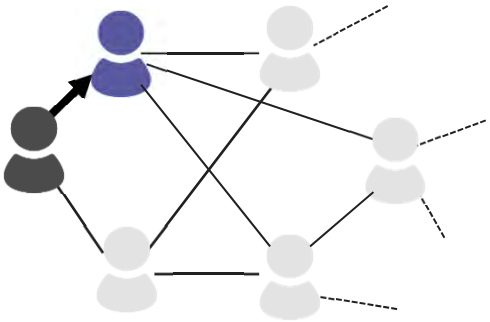
Metrics 2 / 2:
Uniformity of the samples



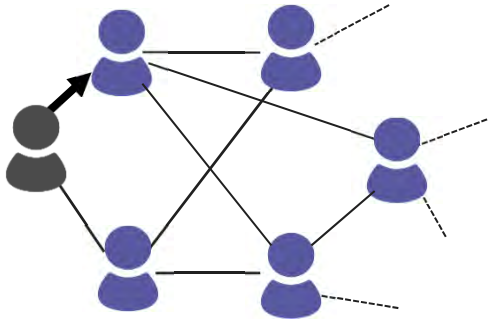
- Better ☺

Wait …

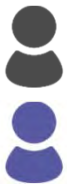
- The proposed technique requires more information.
  - More queries to online social networks, or accesses to storage, ...
- Nodes whose # of degree required for a single step
  - MHRW: 1 candidate of the next hop
  - Proposed tech: all the 2-hop further nodes !!!
    - involves higher querying cost



MHRW



Proposed technique



Current node

Nodes whose # of degree required

Cost reduction technique

- Estimate average degree by running normal MHRW, and use it as degrees of unknown nodes.
  - Parameter φ : ratio of nodes obtained by normal MHRW

Sampled node list :

{A, B, A, A, C, D, B, D, A, A, ..., D}

by normal MHRW by proposed technique

φ : ratio to the entire list

↓ ↑

Estimated average degree

  - Of course, count the queries for the initial normal MHRW

Evaluation based on # of queries

- Target networks

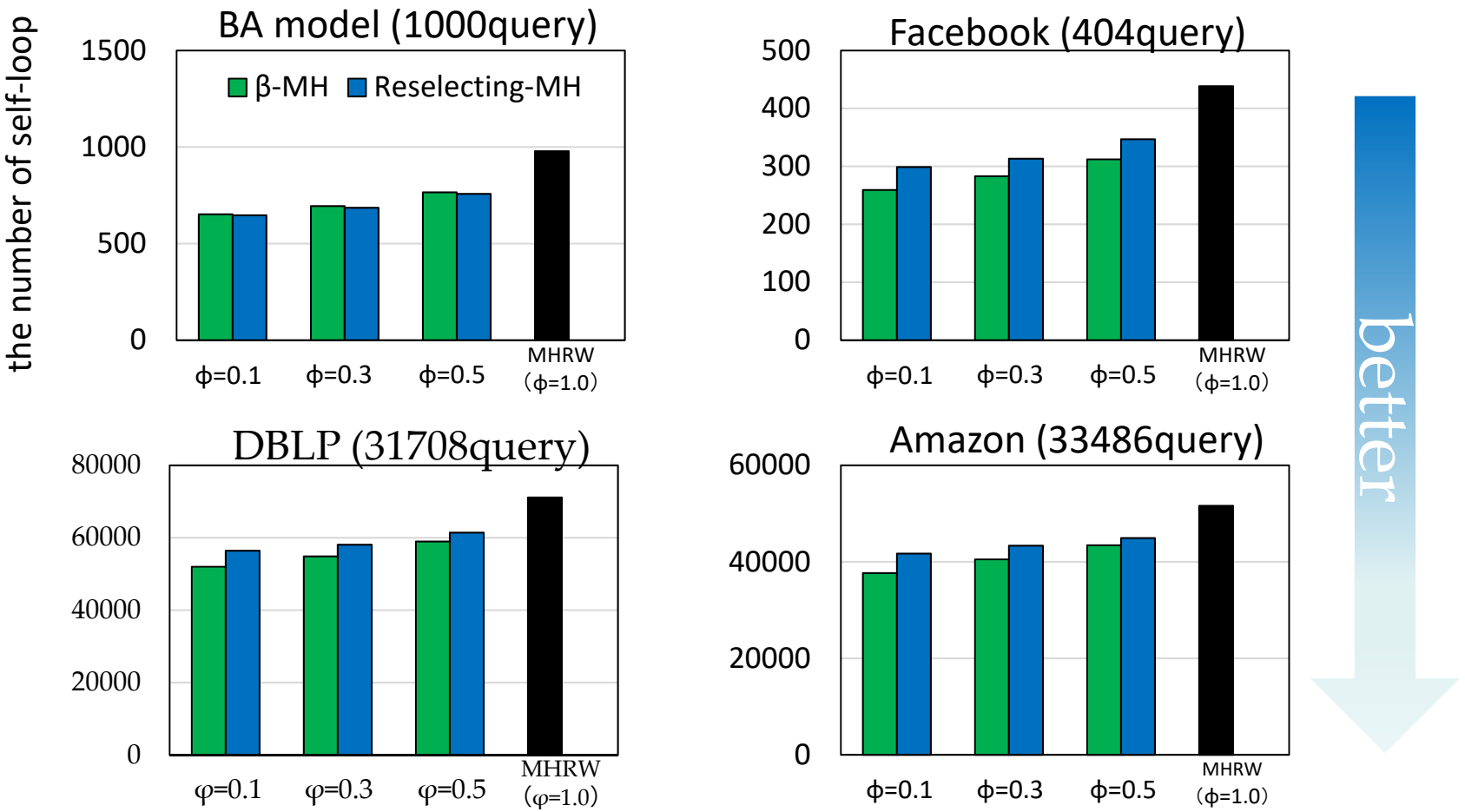
Updated based on p.8

Network	# of nodes	# of vertices	Average degree
Barabasi-Albert model	10,000	29,991	5.998
Facebook (*)	4,039	88,234	43.691
DBLP (*)	317,080	1,049,866	6.622
Amazon (*)	334,863	925,872	5.530

(*) SNAP: https://snap.stanford.edu/data/

- Metrics
  1. Number of "stays"
  2. Number of unique nodes
     - To evaluate cost (# of queries) performance (# of unique nodes)
  3. Uniformity of the sampled node list
     - L^1 distance

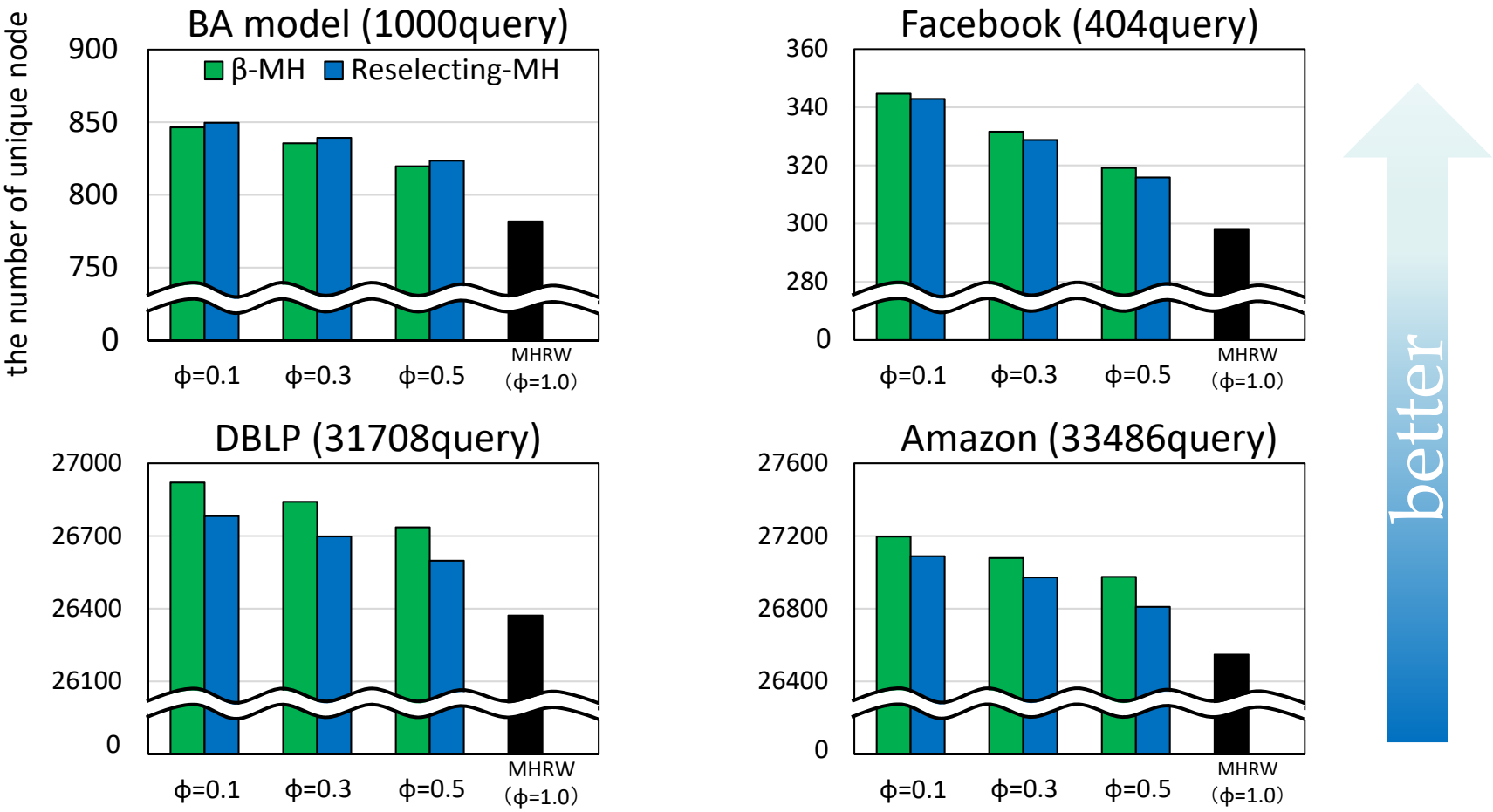
Metrics 1 / 3:
Number of "stays"



- Reduced ☺

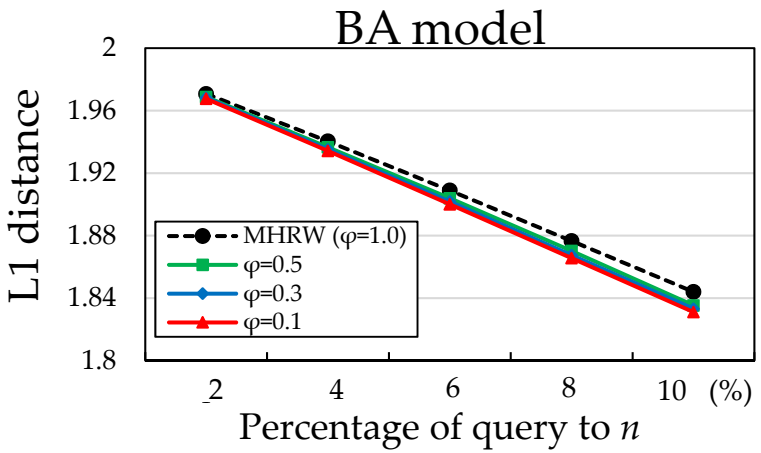
Metrics 2 / 3:

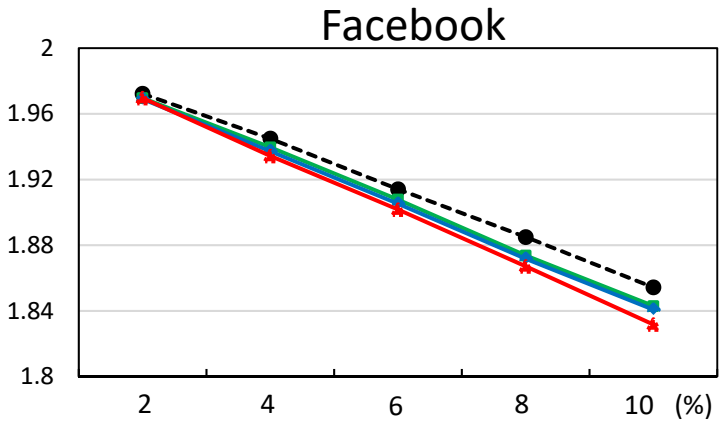
Number of unique nodes

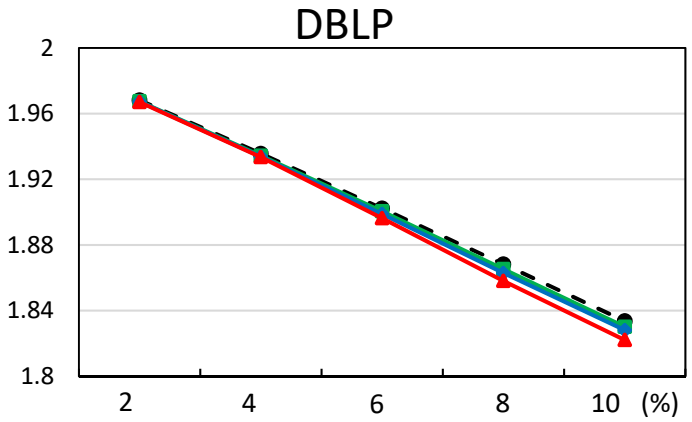


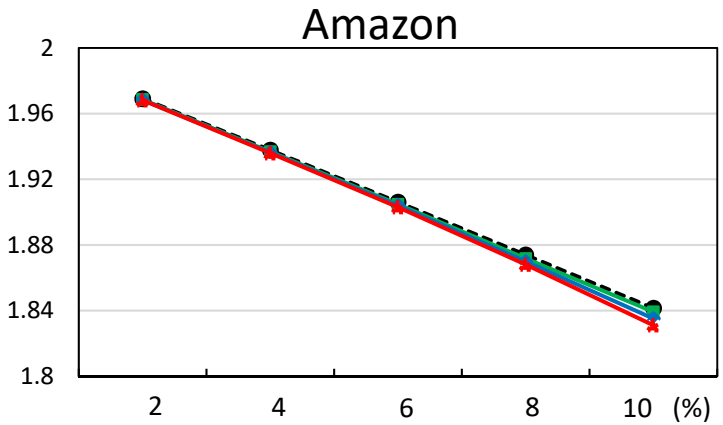
- Increased ☺

Metrics 3 / 3:
Uniformity of the samples









- Better ☺

Summary

- Contributions
  - A random walk with less "stays" while keeping ready-to-use property designed
    - Choice and adoption probabilities regulated
  - A querying cost reduction technique proposed
- Resulted in
  - Less "stays" and more uniform samples
  - Even based on # of queries, not only # of steps
- Future work
  - Better probabilities
  - Better estimation of average degree

Tokyo Tech