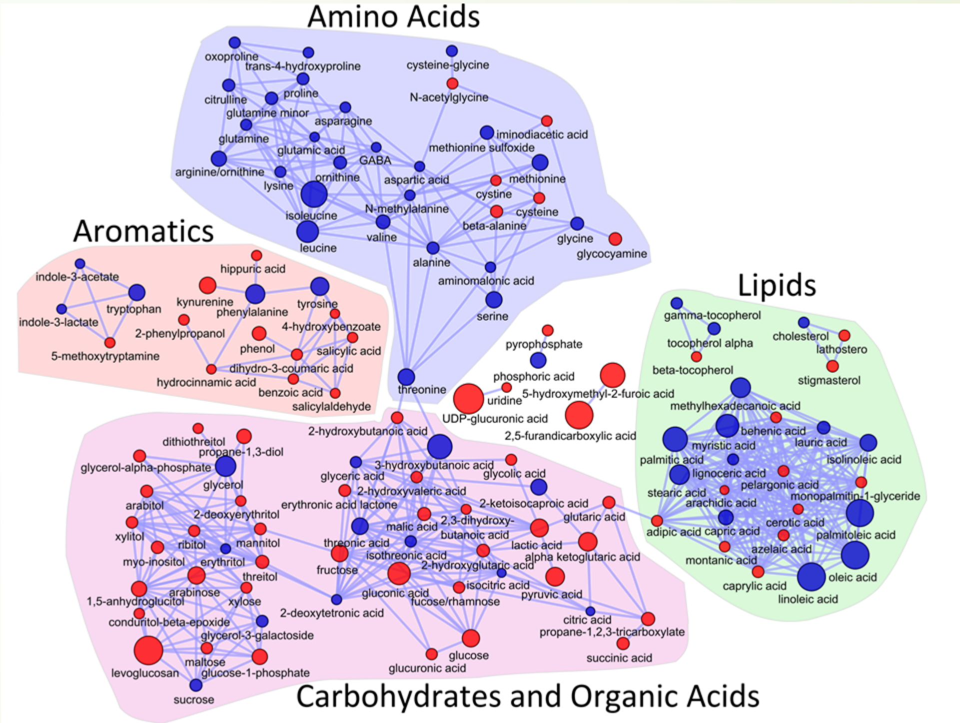
Estimating and Sampling Graphs with Multidimensional Random Walks

Group 2: Mingyan Zhao, Chengle Zhang, Biao Yin, Yuchen Liu

Motivation

Complex Network



Social Network

Biological Network

erence: www.forbe.com imdevsoftware.wordpress.com

Existing Approaches

- Random vertex sampling
- Random edge sampling
- Random walks

Frontier Sampling

- a new *m*-dimensional random walk that uses *m* *dependent* random walkers.

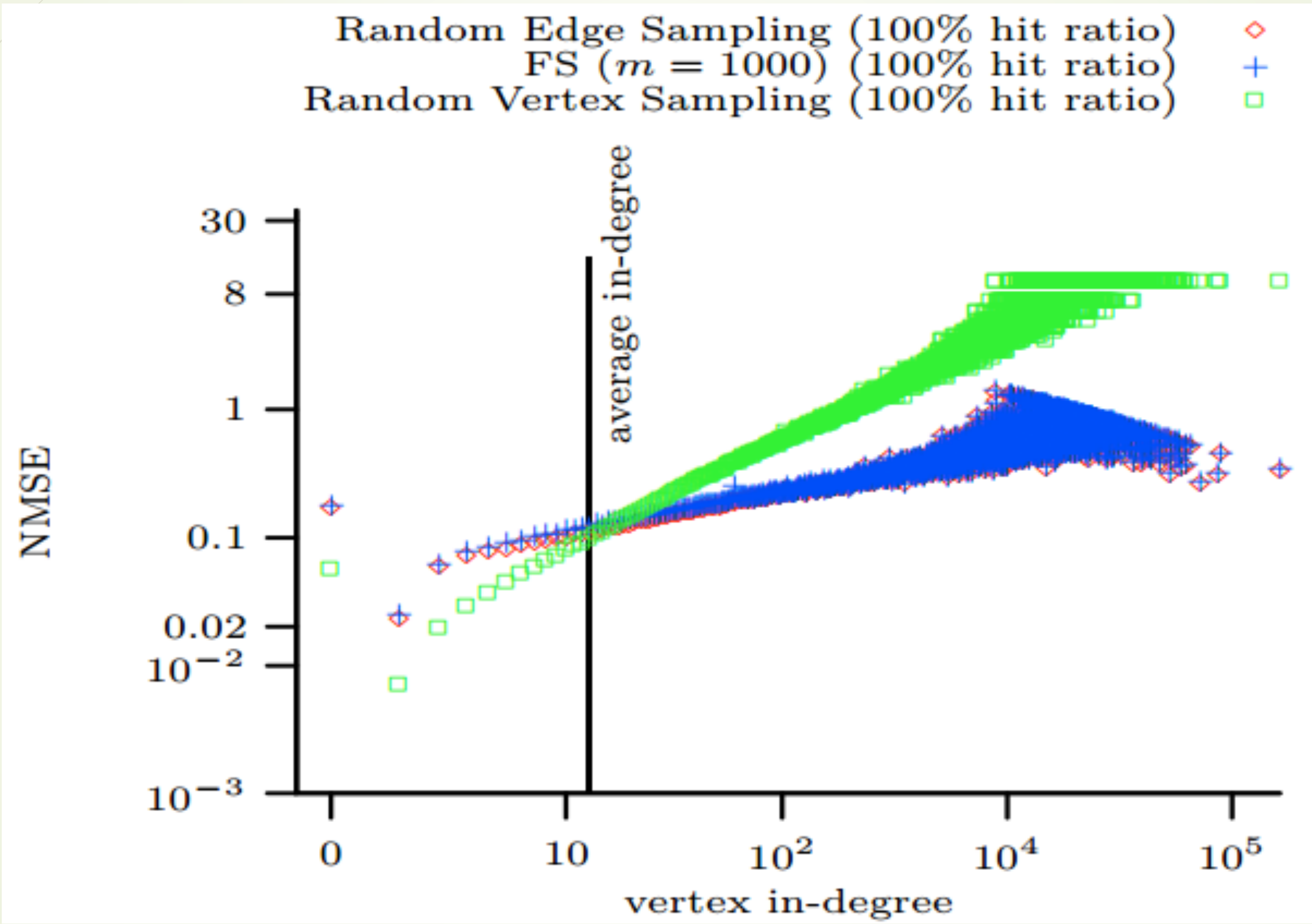
Contribution

- Mitigates the large estimation errors caused by disconnected or loosely connected components.
- Shows that the tail of the degree distribution is better estimated using random edge sampling than random vertex sampling.
- Presents asymptotically unbiased estimators

Definitions

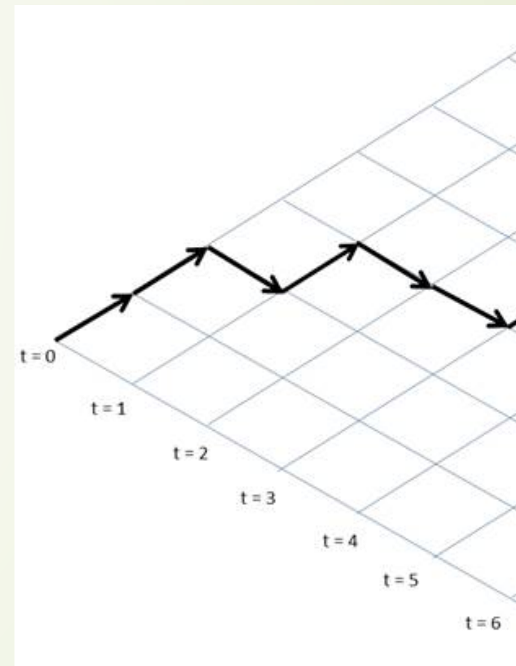
Notation	Meaning
Gd (V, Ed)	A labeled directed grap representing the (origina network graph, where *V* is set of vertices and *Ed* is a s of edges
(u, v)	A connection from *u* to (a.k.a. edges)
$\mathcal{L}$v and $\mathcal{L}$e	Finite set of vertex and edg labels,
$\mathcal{L}_e(u,v) = \emptyset$	Edge (u, v) is unlabeled
$\mathcal{L}_v(v) = \emptyset$	Vertex v is unlabeled

Vertex V.S. Edge Sampling



Section 4

1. Mathematical theories and conductions on Random Walk Sampl
2. Strong Law of Large Numbers
3. Four estimators will be applied in Section 5
4. Deficiency of RW
5. Multiple Independent Random Walkers



Strong Law of Large Numbers

$$\mathrm{m}_{\to\infty} \frac{1}{B^\star(B)} \sum_{i=1}^{B^\star(B)} f(u_i, v_i) \to \frac{1}{|E^\star|} \sum_{\forall (u,v) \in E^\star} f(u, v)$$

inal statistical Thm:

$$\xrightarrow{\text{a.s.}} \mu \qquad \text{when } n \to \infty. \qquad\qquad \Pr\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

k law:

$$\to \infty \qquad \lim_{n\to\infty} P(|X - \mu| \geq \epsilon) = 0.$$

B	Number of RW step
B*(B)	Number of edges in
E*	Total RW Sampled Edges

Estimator 1: Edge Label Density

label edges of interest:

$$\mathbf{1}(l \in \mathcal{L}_e(u,v)) = \begin{cases} 1 & \text{if } l \in \mathcal{L}_e(u,v) \\ 0 & \text{otherwise.} \end{cases}$$

he probability of the labelled edges

$$p_l = \sum_{\forall (u,v) \in E^\star} \frac{\mathbf{1}(l \in \mathcal{L}_e(u,v))}{|E^\star|}$$

estimator based on SLLN

$$\hat{p}_l \equiv \sum_{i=1}^{B^\star(B)} \frac{\mathbf{1}(l \in \mathcal{L}_e(u_i,v_i))}{B^\star(B)}$$

Estimator 2: Assortative Mixing Coefficient (AMC)

sidering directed G, an asymptotically unbiased estimator of AMC:

$$\frac{1}{\hat{\sigma}_{\text{in}}\,\hat{\sigma}_{\text{out}}} \underbrace{\sum_{i=0}^{W_{\text{out}}}\sum_{j=0}^{W_{\text{in}}} ij(\hat{p}_{ij} - \hat{q}_i^{\text{out}}\hat{q}_j^{\text{in}})}_{\text{Covariance}}$$

$$\hat{p}_{ij} \equiv \sum_{k=1}^{B^\star(B)} \frac{\mathbf{1}(\text{outdeg}(u_k) = i\,,\ \text{indeg}(v_k) = j)}{B^\star(B)}\ ;$$

$$\hat{q}_i^{\text{out}} \equiv \sum_{k=0}^{W_{\text{in}}} \hat{p}_{ik} \quad ; \quad \hat{q}_j^{\text{in}} \equiv \sum_{k=0}^{W_{\text{out}}} \hat{p}_{kj} \quad ;$$

$$\hat{\sigma}_{\text{in}} = \sqrt{\sum_{i=0}^{W_{\text{out}}} j^2\hat{q}_j^{\text{in}} - \left(\sum_{i=0}^{W_{\text{out}}} j\hat{q}_j^{\text{in}}\right)^2} \quad ; \quad \text{and}$$

$$\hat{\sigma}_{\text{out}} = \sqrt{\sum_{i=0}^{W_{\text{in}}} i^2\hat{q}_i^{\text{in}} - \left(\sum_{i=0}^{W_{\text{in}}} i\hat{q}_i^{\text{out}}\right)^2}$$

ch two are highly correlated?

Estimator 3:Vertex label Density

onstruct an asymptotically unbiased estimator:

$$\hat{\theta}_l \equiv \frac{1}{S\,B}\sum_{i=1}^{B}\frac{\mathbf{1}(l \in \mathcal{L}_v(v_i))}{\deg(v_i)}$$

ce:

$$\frac{1}{B}\sum_{i=1}^{B}\frac{\mathbf{1}(l \in \mathcal{L}_v(v_i))}{\deg(v_i)} \rightarrow \frac{1}{|E|}\sum_{\forall (u,v)\in E}\frac{\mathbf{1}(l \in \mathcal{L}_v(v))}{\deg(v)}$$

$$S = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{\deg(v_i)}$$

$$\lim_{B\rightarrow\infty} S \rightarrow |V^{\star}|/|E|$$

, this estimator converges to:

$$\theta_l = \frac{1}{|V|}\sum_{\forall (u,v)\in E}\frac{\mathbf{1}(l \in \mathcal{L}_v(v))}{\deg(v)}$$

Estimator 4: Global Clustering Coefficient

$$C \equiv \frac{1}{|V^\star|} \sum_{\forall v \in V} c(v) \,,$$

ere

$$c(v) = \begin{cases} \Delta(v)/\binom{\deg(v)}{2} & \text{if } \deg(v) \geq 2 \\ 0 & \text{otherwise}\,, \end{cases}$$

ere $\Delta(v) = |\{(u,w) \in E \,:\, (v,u) \in E \text{ and } (v,w) \in E\}$

unbiased estimator by SLLN

$$\hat{C} \equiv \frac{1}{S\,B} \sum_{i=1}^{B} \frac{f(v_i, u_i)}{\binom{\deg(v_i)}{2}} \frac{1}{\deg(v_i)}$$

$$S = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{\deg(v_i)}$$

e:

$$\lim_{B\to\infty} \frac{1}{B} \sum_{i=1}^{B} \frac{f(v_i,u_i)}{\binom{\deg(v_i)}{2}} \frac{1}{\deg(v_i)} \to \frac{1}{|E|} \sum_{\forall (v,u) \in E} \frac{f(v,u)}{\binom{\deg(v)}{2}}$$

$$\lim_{B\to\infty} S \to |V^\star|/|E|$$

Deficiency of RW from one point

1. **"Trapped" inside a subgraph (MSE)**
2. **Start from non-stationary (non-steady) state (MSE, Bias)**

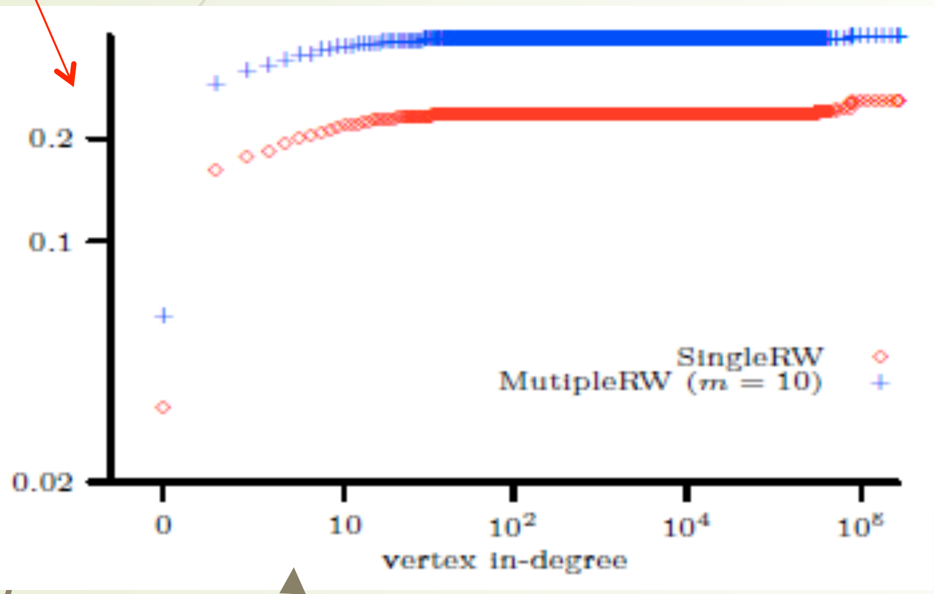
Burn-in period: Discard the non-stationary samples

1. Just decrease error with non-stationary one
2. Discarding in a small sample is not ideal

Multiple Independent Random Walkers cor

Single RW and Multiple RW

Still very high



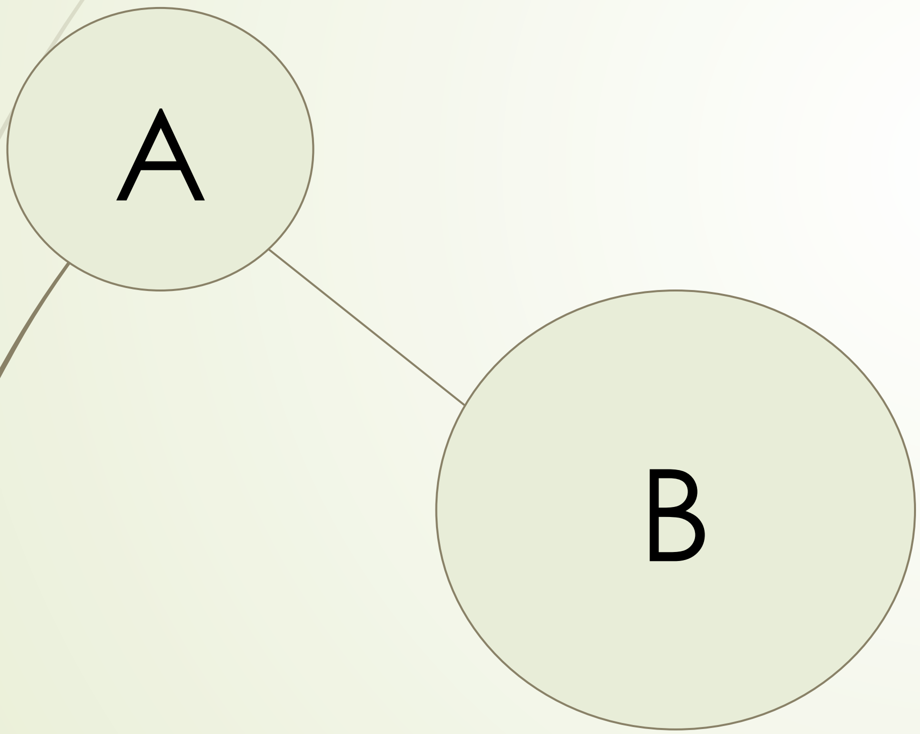
Log-log plot!

Single RW is depend on sample sizes from the estimators provided.

Mutiple RW will split the sample sizes into each path.

So, error in the total CNMSE

Why not M-independent RW ?



MIRW is hard to sample m independent vertex with p proportional to their degrees.

$$\deg(v)/\mathrm{vol}(V)$$

Motivation

- We want an m-dimensional random walk that, in steady state, samples edges uniformly at random but, unlike MultipleRW, can benefit from starting its walkers at uniformly sampled vertices.

Frontier Sampling

orithm 1: Frontier Sampling (FS).

$n \leftarrow 0$ {n is the number of steps}
Initialize $L = (v_1, \ldots, v_m)$ with m randomly chosen vertices (uniformly)
repeat
  Select $u \in L$ with probability $\deg(u)/\sum_{\forall v \in L} \deg(v)$
  Select an outgoing edge of u, (u, v), uniformly at random
  Replace u by v in L and add (u, v) to sequence of sampled edges
  $n \leftarrow n + 1$
until $n \geq B - mc$

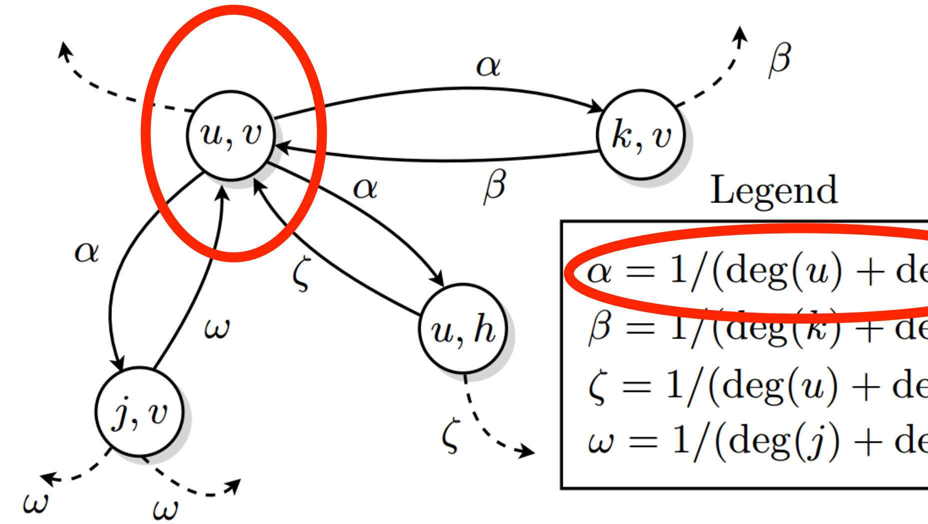


Figure 2: Illustration of the Markov chain associated to tier sampler with dimension $m = 2$.

= deg(u)/∑L↑▒deg($\dfrac{1}{\sum_{\forall v \in L_n} \deg(v)}$.

Frontier Sampling: A m-dimensional Random Wa

- The frontier sampling process is equivalent to the sampling process of a single random walker over G^m. (Lemma 5.1)
  - P(selecting a vertex and its outgoing edge in FS) = P(randomly sampling an edge from e(L_n) in single random walker over G^m).

$$p = \frac{1}{|e(L_n)|} = \frac{1}{\sum_{\forall v \in L_n} \deg(v)}.$$

FS Steady State v.s Uniform Distribution

- $K_{fs}(m)$ be a random variable that denotes the number of random walkers in V_A in steady state.
- Let $K_{un}(m)$ be a random variable that denotes the number of sampled vertices, out of m uniformly sampled vertices from V, that belong to V_A.

$$\lim_{m\to\infty} P[K_{fs}(m) = k] = \lim_{m\to\infty} P[K_{un}(m) = k],\ \forall k \geq 0. \quad (9)$$

- Proving this to be true indicates that the FS algorithm starting with m random walkers at m uniformly sampled vertices approaches the steady state distribution. This means FS benefits from starting its walkers at uniformly sampled vertices by reducing transient of RW.

FS Steady State V.S. Uniform Distribution

- By definition $P[K_{un}(m) = k] = \binom{m}{k} p^k (1-p)^{m-k},$
- Theorem 5.2 $$P[L = (v_1, \ldots, v_m)] = \frac{\sum_{i=1}^{m} \deg(v_i)}{m|V|^{m-1}\mathrm{vol}(V)}.$$
- Lemma 5.3 $$P[K_{fs}(m) = k] = \frac{1}{md}\binom{m}{k} p^k (1-p)^{m-k} (k\, d_A + (m-k) d_B),$$ *where* $p = |V_A|/|V|$ *and* $0 \leq k \leq m.$
- Theorem 5.4 $\lim_{m\to\infty} P[K_{fs}(m) = k] = \lim_{m\to\infty} P[K_{un}(m) = k], \forall k \geq 0.$

MultipleRW Steady State V.S. Uniform Distributio

- $K_{mw}(m)$ be a random variable that denotes the steady state number of MultipleRW random walkers in V_A.

$$E[K_{mw}(m)] = \frac{m\,|V_A|d_A}{|V|d}$$

$$E[K_{un}(m)] = \frac{m\,|V_A|}{|V|}$$

$$\alpha_A = E[K_{mw}(m)]/E[K_{un}(m)] = d_A/d.$$

Note: d_A (average degree of vertices in V_A)

Conclusion: If we initialize m random walkers with uniformly sampled vertices, FS starts closer to steady state than MultipleRW.

Distributed Frontier Sampling

- Frontier Sampling can also be parallelized.
- A MultipleRW sampling process where the cost of sampling a vertex v is an exponentially distributed random variable with parameter deg(v) is equivalent to a FS process. (Theorem 5)

Experiment and Result

- Data:
  - “Flickr”, “Livejournal”, and “YouTube”
  - Barabási-Albert [5] graph
- Goal:
  - Compare FS to MultipleRW, SingleRM
  - Compare FS on random vertex and edge sampling
- Result: FS is constantly more accurate

Assortative Mixing Coefficient

$$\hat{r} \equiv \frac{1}{\hat{\sigma}_{\text{in}}\,\hat{\sigma}_{\text{out}}} \sum_{i=0}^{W_{\text{out}}} \sum_{j=0}^{W_{\text{in}}} ij(\hat{p}_{ij} - \hat{q}_i^{\text{out}}\,\hat{q}_j^{\text{in}}),$$

Graph	r	FS		MultipleRW		SingleRW	
		Bias	|NMSE|	Bias	|NMSE|	Bias	|NMSE|
Flickr	0.007	8%	1.08	752%	7.65	−619%	27.32
LiveJournal	0.07	−0.5%	0.11	−12%	0.16	1%	0.17
Internet RLT	0.17	3%	0.33	2%	0.32	17%	0.44
Youtube	−0.03	0.001%	0.02	2%	0.03	−1%	0.1
G_{AB}	0.08	0.01%	0.12	70%	0.72	100%	1.00

In-degree Distribution Estimates

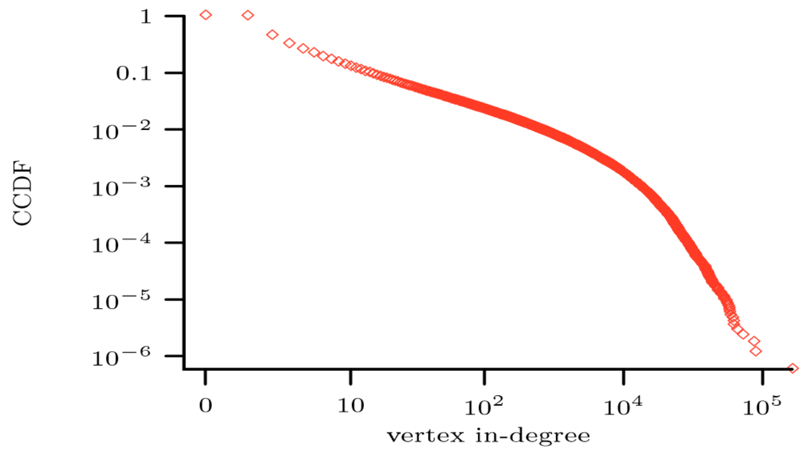


Figure 3: (Flickr) Log-log plot of the in-degree CCDF.

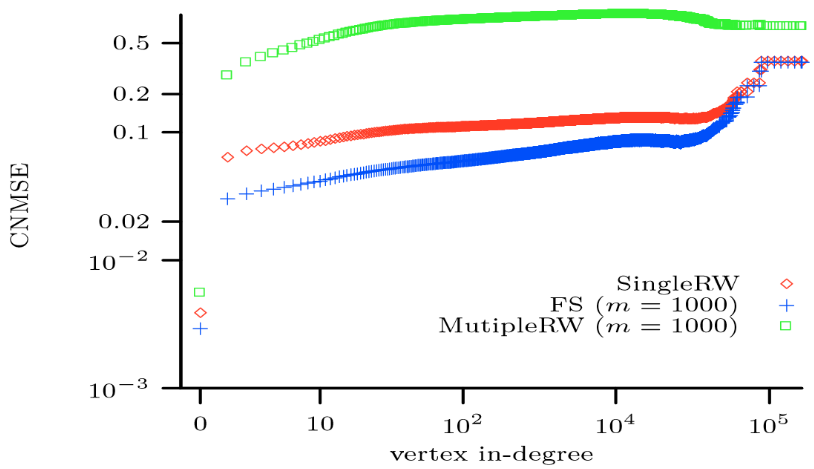


Figure 4: (LCC of Flickr) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $B = |V|/100$.

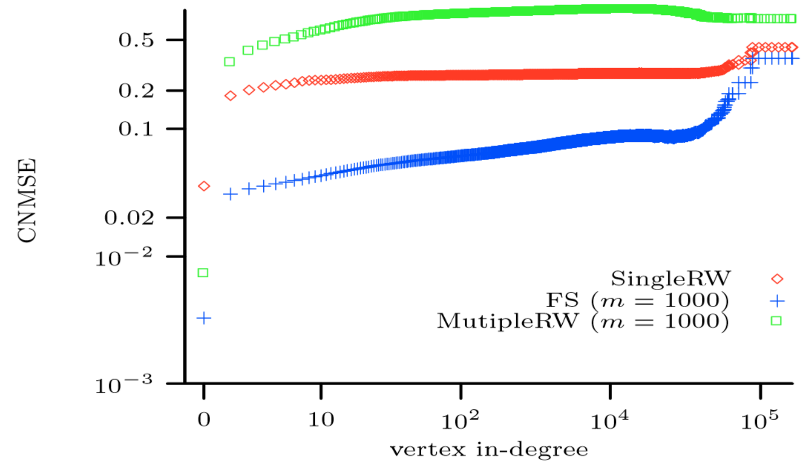


Figure 5: (Flickr) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $B = |V|/100$.

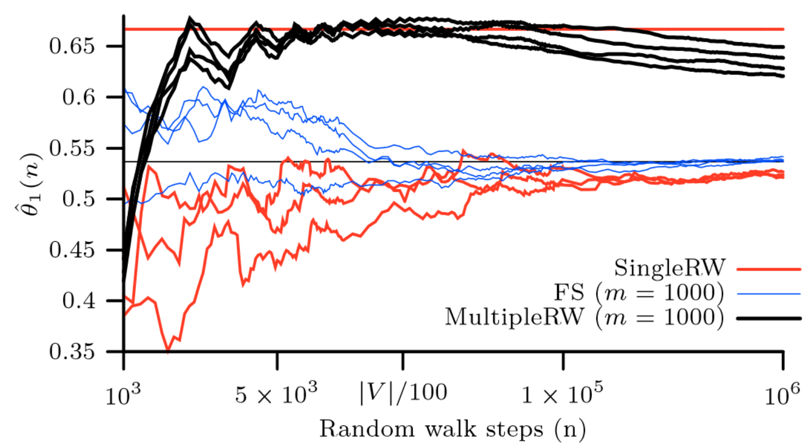
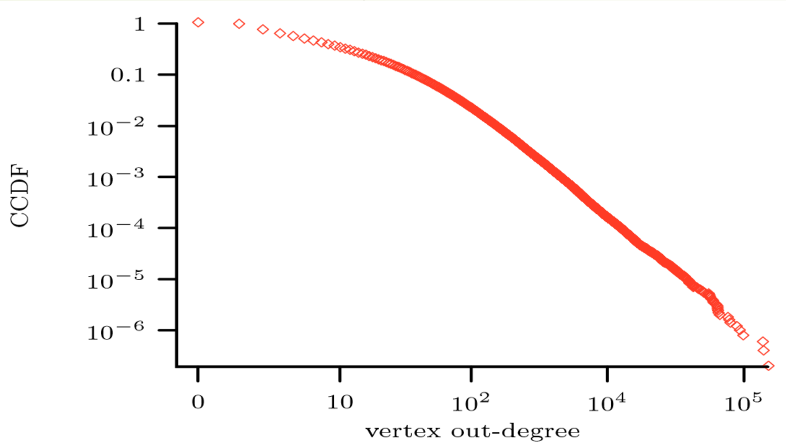
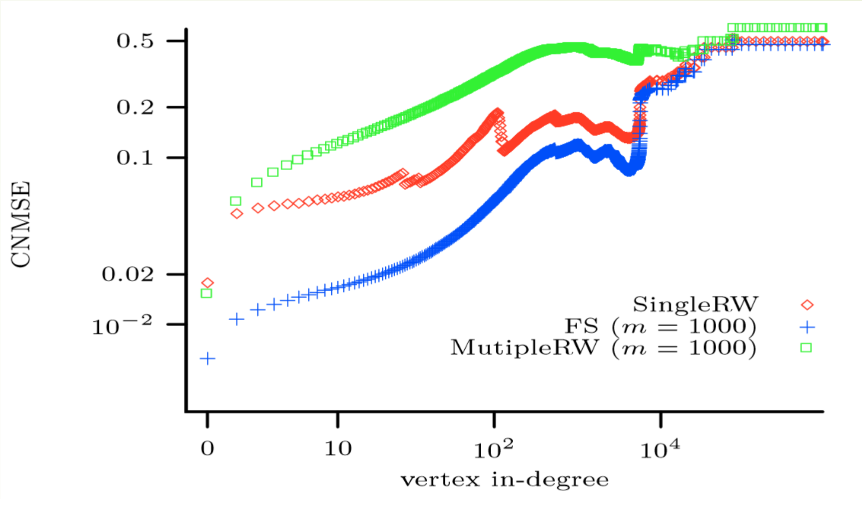
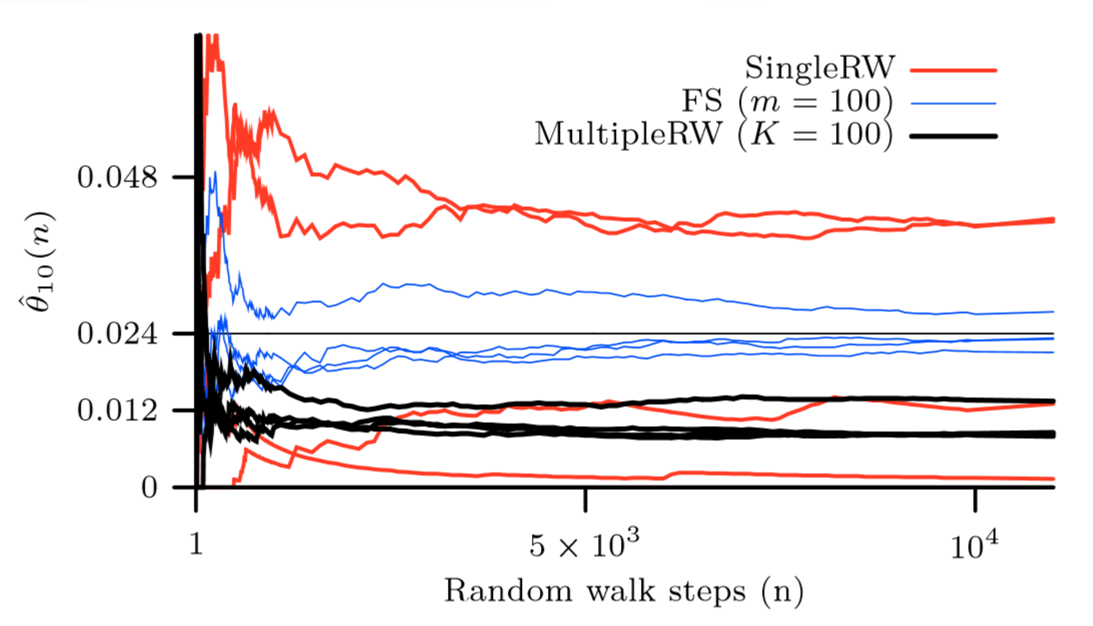


Figure 6: (LCC of Flickr) Four sample paths of $\hat{\theta}_1$ ($\theta_1 = 0.53$) as a function of the number of steps n (horizontal axis in log scale).

Out-degree Distribution Estimates

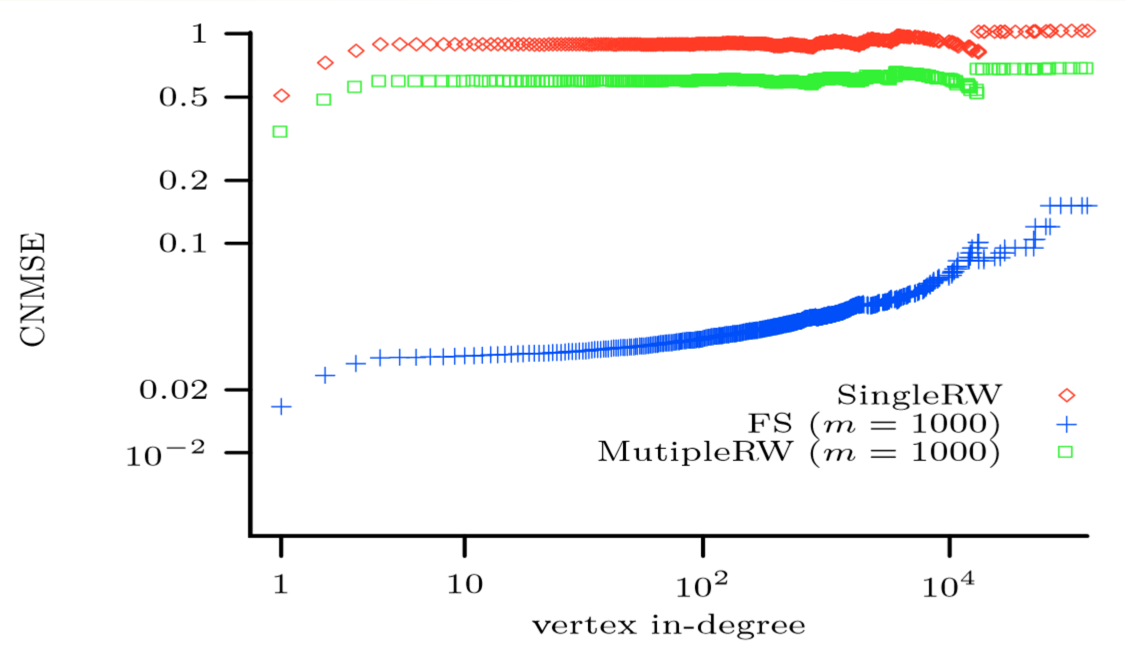






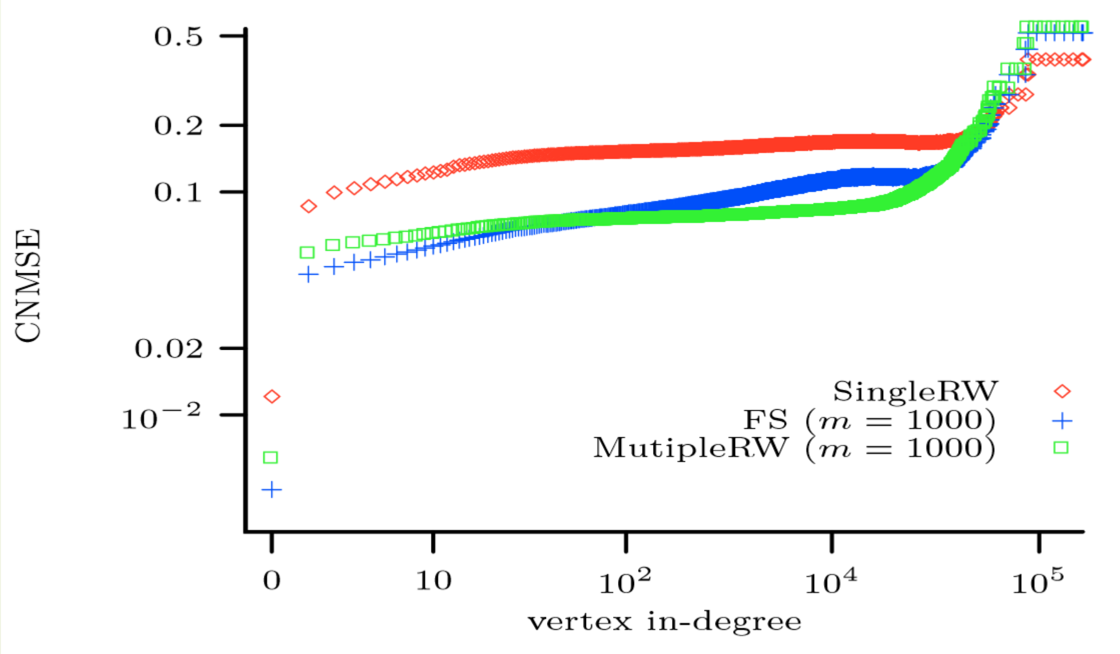
In-degree Distribution
loosely connected components

Barabási-Albert Graph

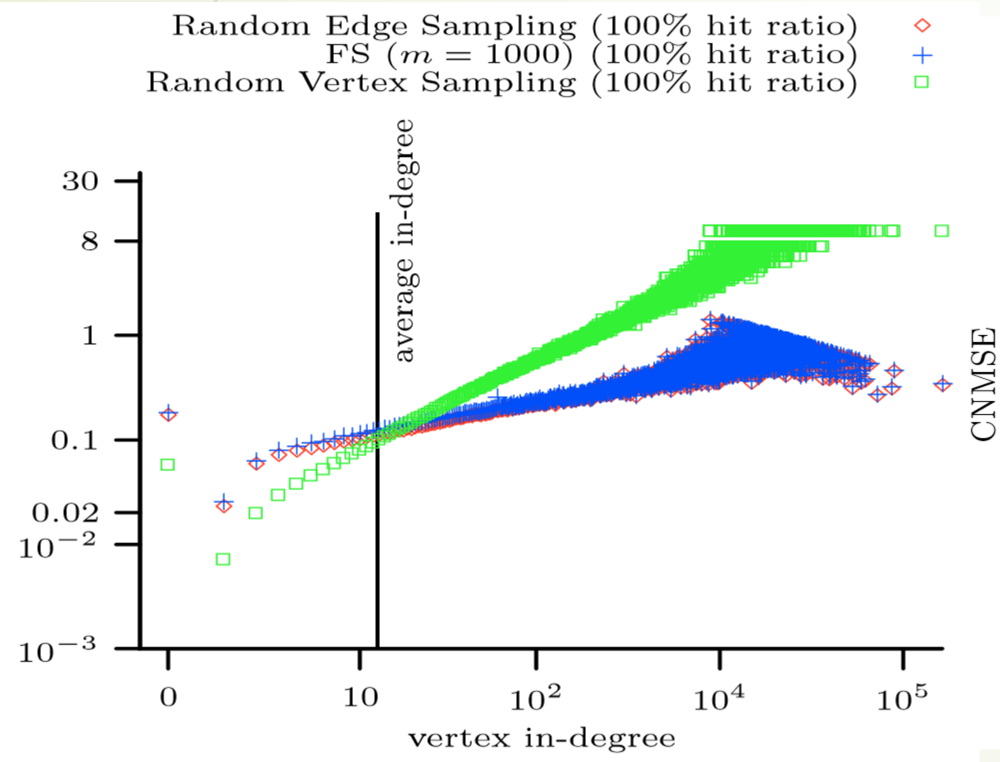


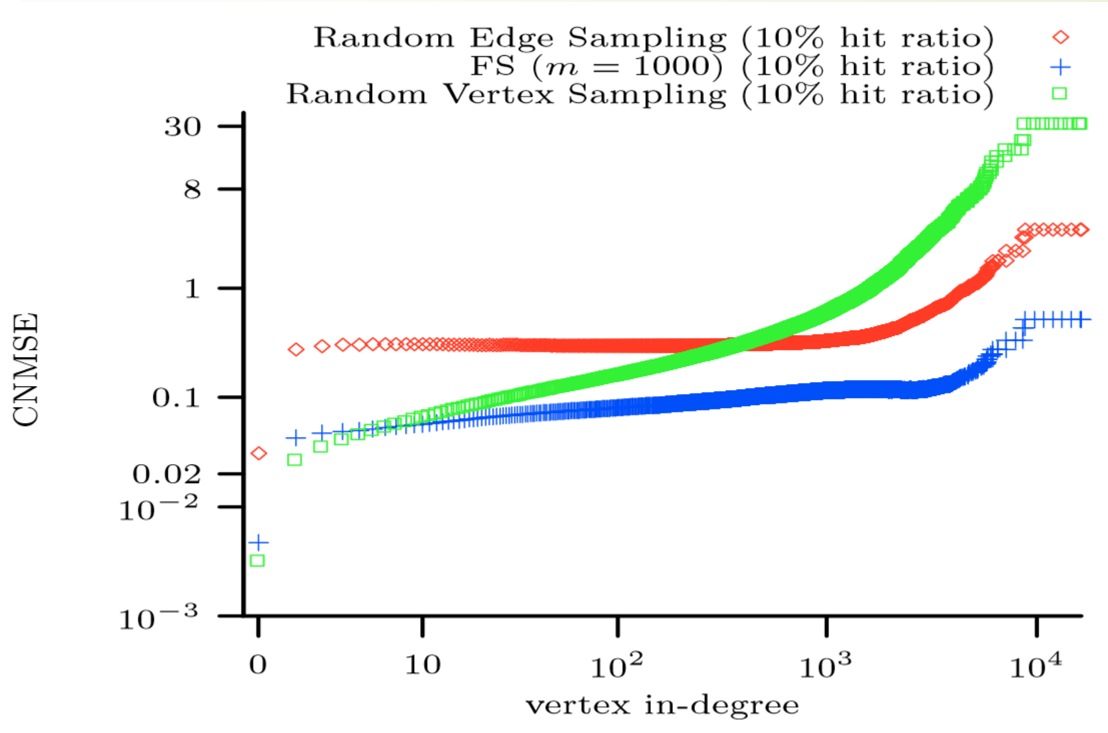
FS V.S. Stationary MultipleRW & SingleRW

- MultipleRW and SingleRW start with steady state

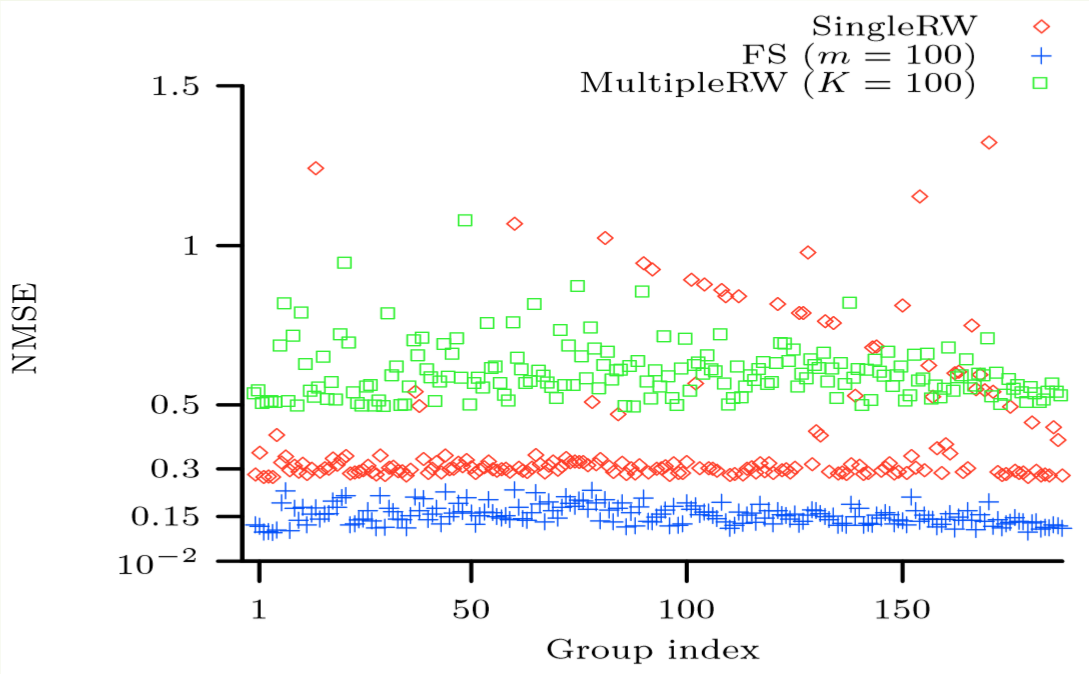


FS V.S. Random Independent Samplin





Density of Special Interest Group



Global Clustering Coefficient Estimates

- Global Clustering Coefficient

  a measure of the degree to which nodes in a graph tend to cluster together.

			$E[\hat{C}]$ (NMSE)		
Graph	B	C	FS	SingleRW	MultipleRW
Flickr	1%	0.14	0.13 (0.04)	0.12 (0.33)	0.16 (0.18)
LiveJournal	1%	0.16	0.16 (0.02)	0.16 (0.02)	0.17 (0.06)

Conclusion

- In almost all of the tests, FS is better.

Future Work

- estimating characteristics of dynamic networks
- design of new MCMC-based approximation algorithms

Thank you!