Our world view

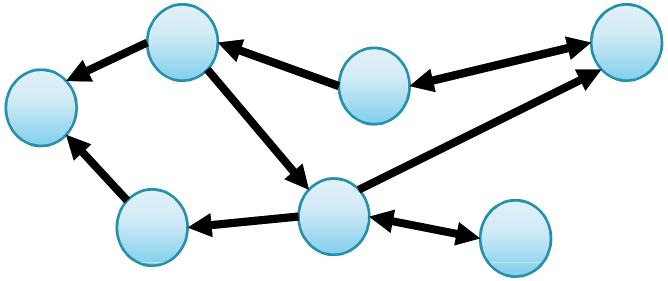
- world view is biased
- depends c
  - where yo
  - your network connections
  - the network structure

Research to enable unbiased view of the "world"

Saul Steinberg's *View of the World from 9th Avenue*

Types of networks

Directed

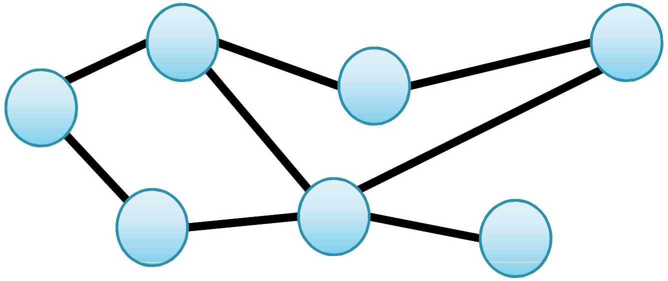


incoming and outgoing edges can be queried

E.g.:

- YouTube
- Livejournal
- Twitter
- ArXiv

Undirected



E.g.:

- OSNs (Facebook, MySpace, etc.)
- Computer networks (in general)
- Family ties (e.g. DNA mutations)

Directly observable characteristics

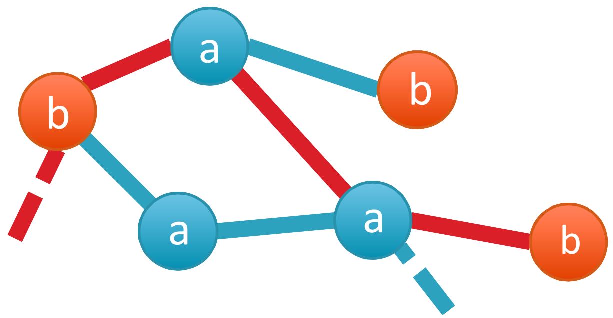
graph: G=(V,E)

Compute:

$$h(V) = \sum_{\forall v \in V} w(v)$$

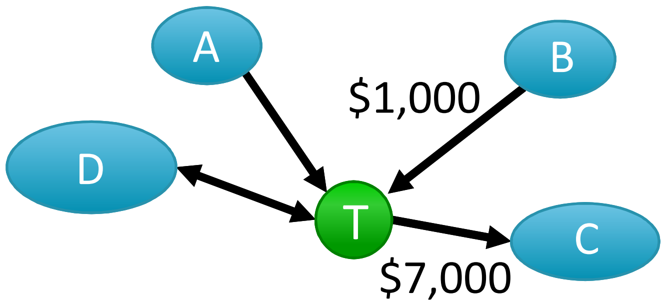
$$f(E) = \sum_{\forall (u,v) \in E} g(u,v)$$

network



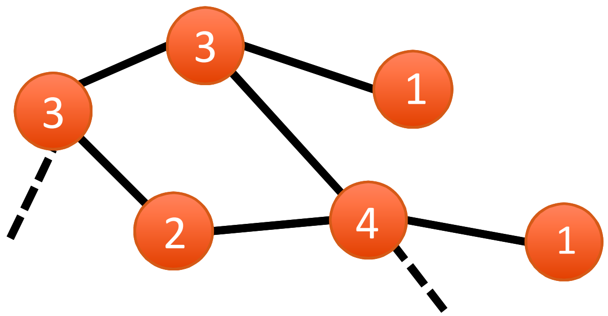
vertex/edge labels

bank transactions



money transfers

Facebook network



vertex degrees

Graph measurements

Can pick up vertex characteristics by querying

- Web, FaceBook, YouTube, …

Resource constraints: too expensive to query all vertices

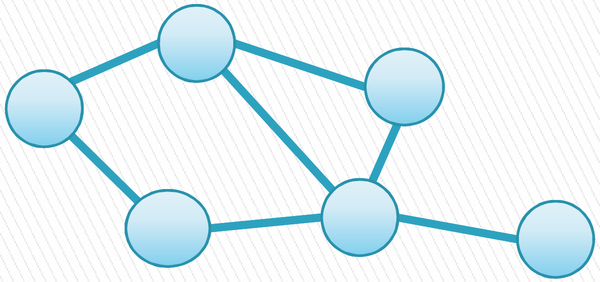
- size (100M+ vertices)
- query rate restrictions

How then? sampling/crawling

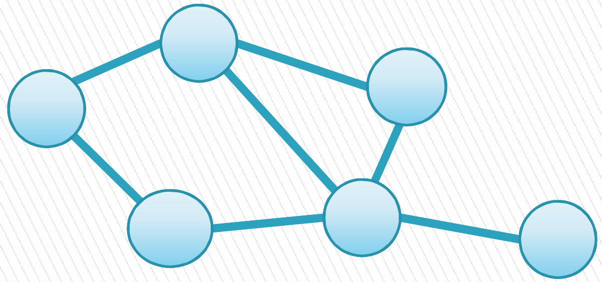
- Leslovec et al, 2006, Mislove, et al 2007, …

Random Sampling v.s. Crawling

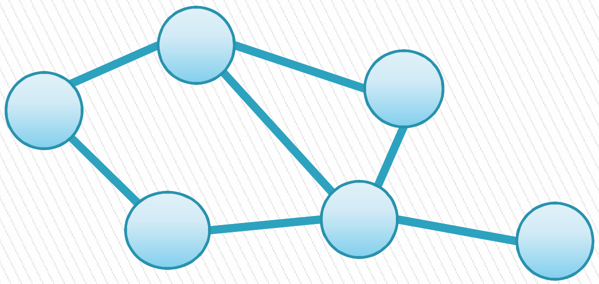
- vertex sampling



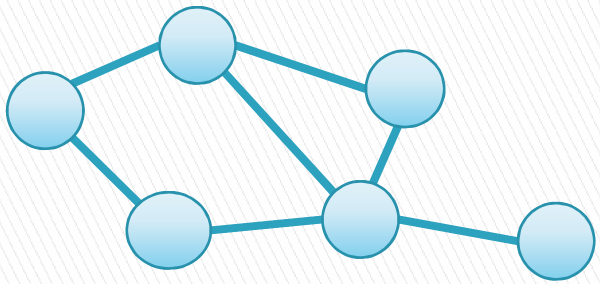
- snowball sampling



- edge sampling



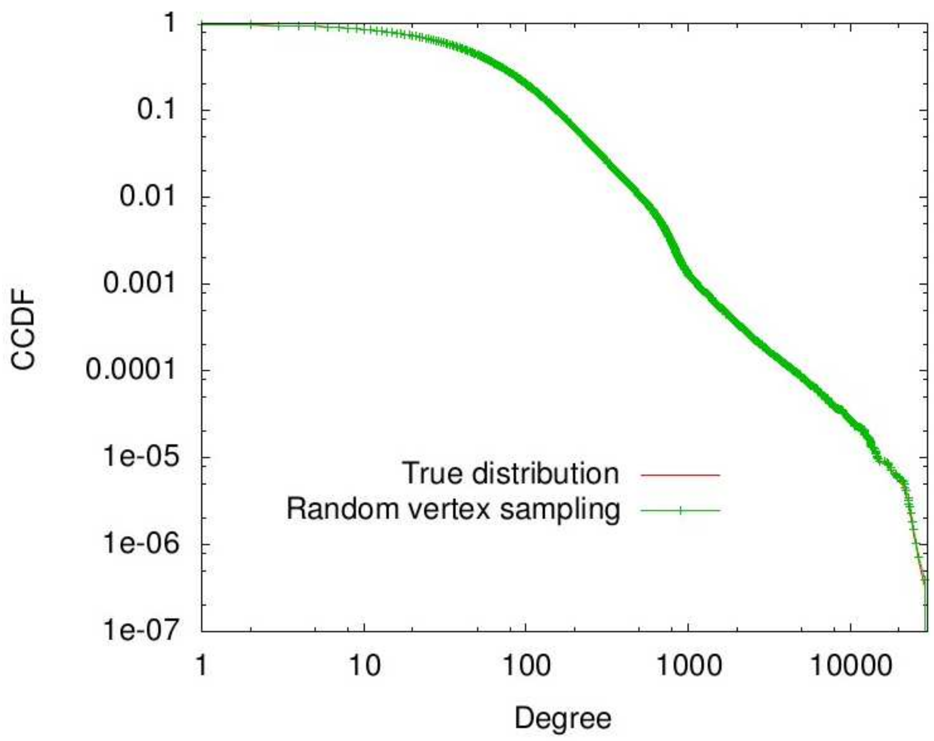
- random walk sampling

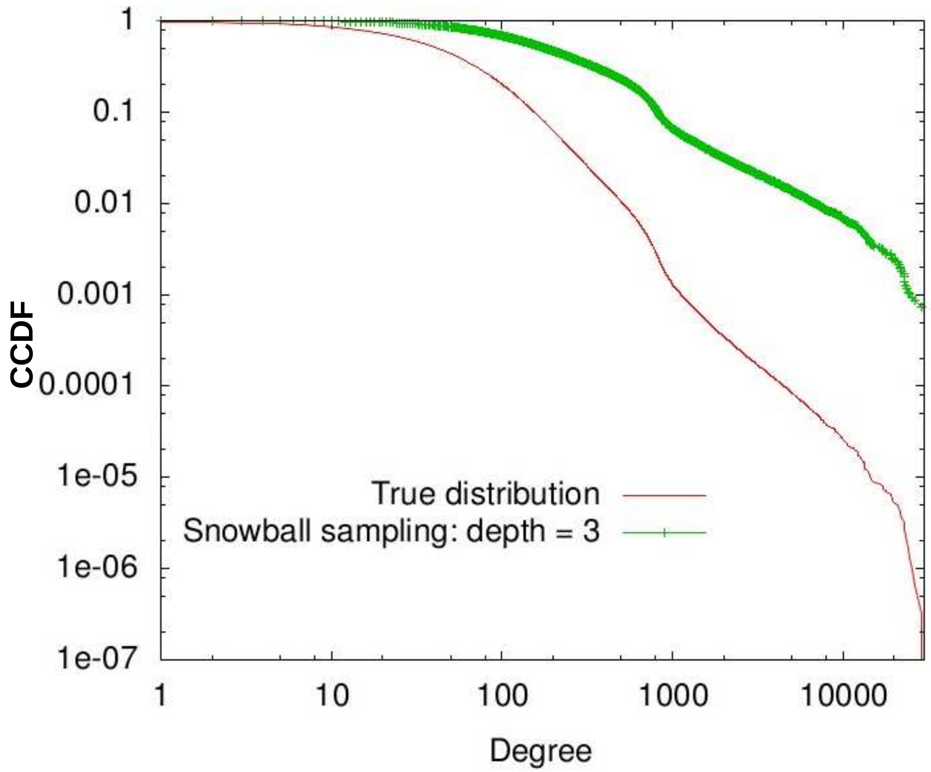


Random sampling	Crawling

Vertex sampling, snowball sampling

- Orkut data set (Mislove 2007), 3M vertices, 200M edges

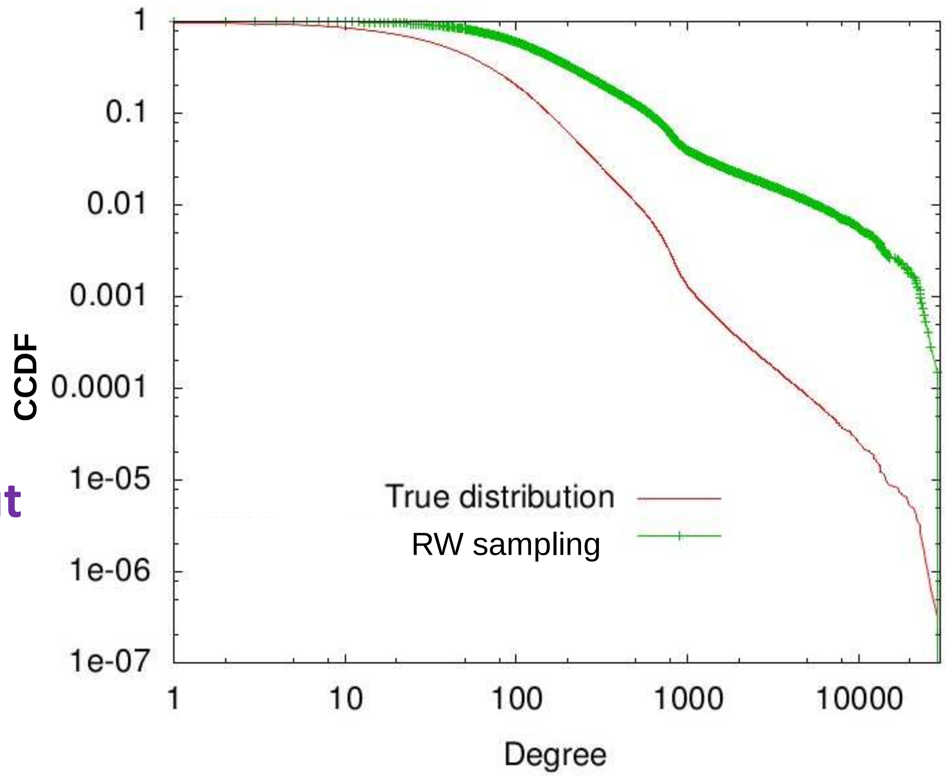




Snowball sampling highly biased

Random walks

- random walk (RW)
  - simple to implement
  - in steady state RW visits edges uniformly at random
  - RW ≡ random edge sampling **without** independence



- v – vertex in undirected graph G
  - $deg(v)$ – degree v
  - $|E|$ - total number of edges

P[v visited in RW] = $deg(v)/|E|$

- $\theta_i = \pi_i \times \text{avg. degree}/i$

- obtains unbiased estimates of

$$h(V) = \sum_{\forall v \in V} w(v)$$

$$f(E) = \sum_{\forall (u,v) \in E} g(u,v)$$

Estimation from sampling

- random vertex sampling (uniform + independent)
  - unbiased
  - not always possible
  - high overhead
    - MySpace – 10% of ID space populated
    - Orkut – 7% of ID space populated
- snowball sampling
  - biased (but under certain conditions bias can be removed)
- random walk sampling
  - *Markov Chain Monte Carlo estimation*
  - estimator asymptotically unbiased
    e.g. RDS (Heckathorn 1997)

Sampling error – independent degrees

degree distribution θ_i ; B samples

- error metric: Normalized root Mean Squared Error

$$\text{NMSE}(i) = \frac{\sqrt{E[(\hat{\theta}_i - \theta_i)^2]}}{\theta_i}$$

- random vertex sampling

θ **head**: **GOOD**
θ **tail**: **BAD**

$$\text{NMSE}(i) = \sqrt{(1/\theta_i - 1)/B}$$

- random w… …ing

Power-law tails more accurate with RW

θ **head**: **BAD**
θ **tail**: **GOOD**

$$\text{NMSE}(i) = \sqrt{(1/\pi_i - 1)/B}\,, \quad i > 0\,,$$

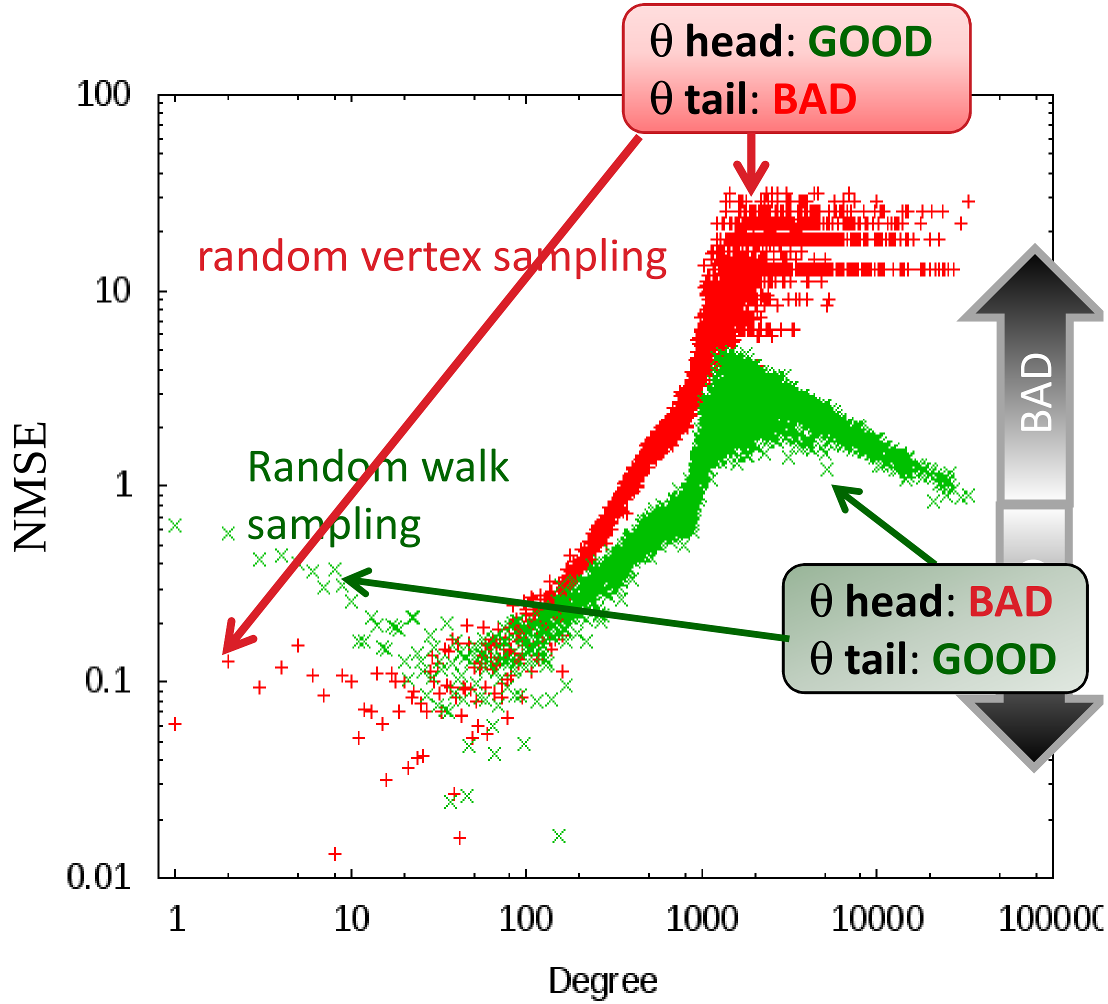
$$\theta_i = \pi_i \times \text{avg. degree}/i$$

What happens when i > avg.degree ?

Simulation 1, Orkut
Random Walk vs. Random vertex sampling

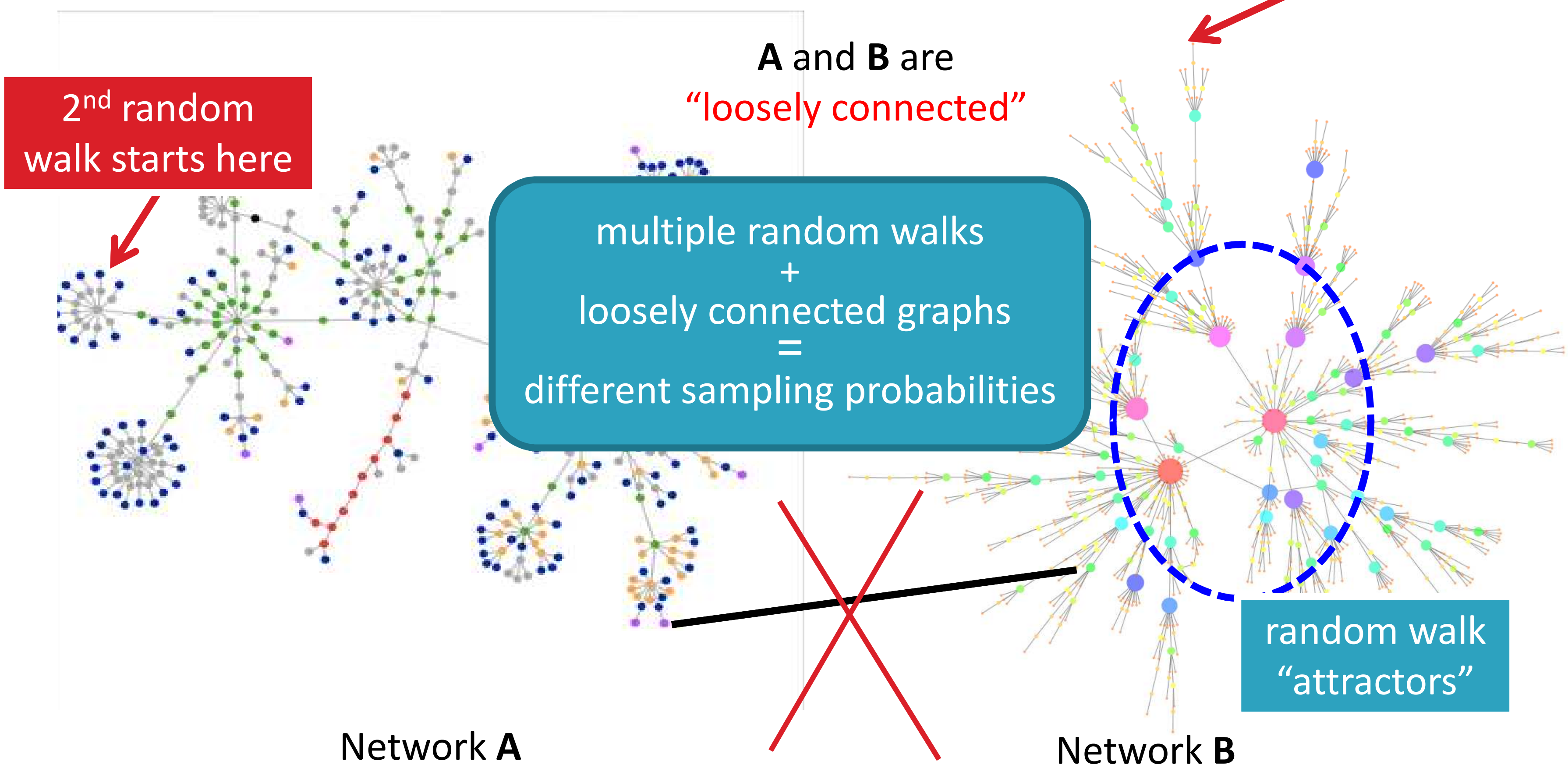
0.3% vertices sampled

- random vertex sampling
- random walk sampling



Random Walk drawback

Consider the following (extreme) thought experiment:



Multiple dependent random walks : Frontier Sampling (FS)

B – sampling budget

Let $S = \{v_1, v_2, \ldots, v_m\}$ be a set of m vertices

(1) start from $v_r \in S$ w.p. $\propto$ $\deg(v_r)$

(2) walk one step from v_r

(3) add **walked** edge to E' and **update** v_r

(4) return to (1) (until $m + |E'| = B$)

FS facts

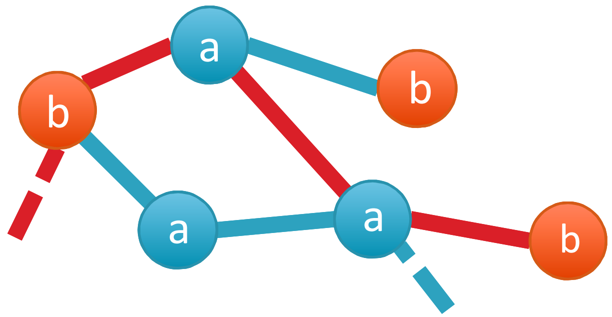
- centrally coordinated
- when stationary
  - edges sampled uniformly
  - vertices sampled ∝ vertex degree

- like a RW, FS estimates:

$$h(V) = \sum_{\forall v \in V} w(v)$$

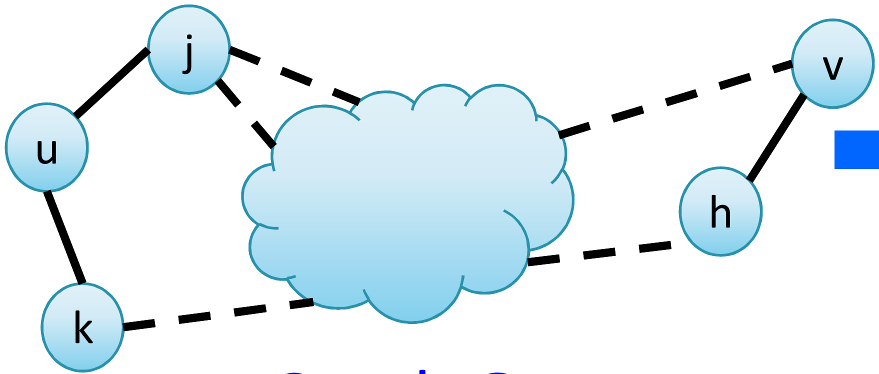
$$f(E) = \sum_{\forall (u,v) \in E} g(u,v)$$

network



vertex labels

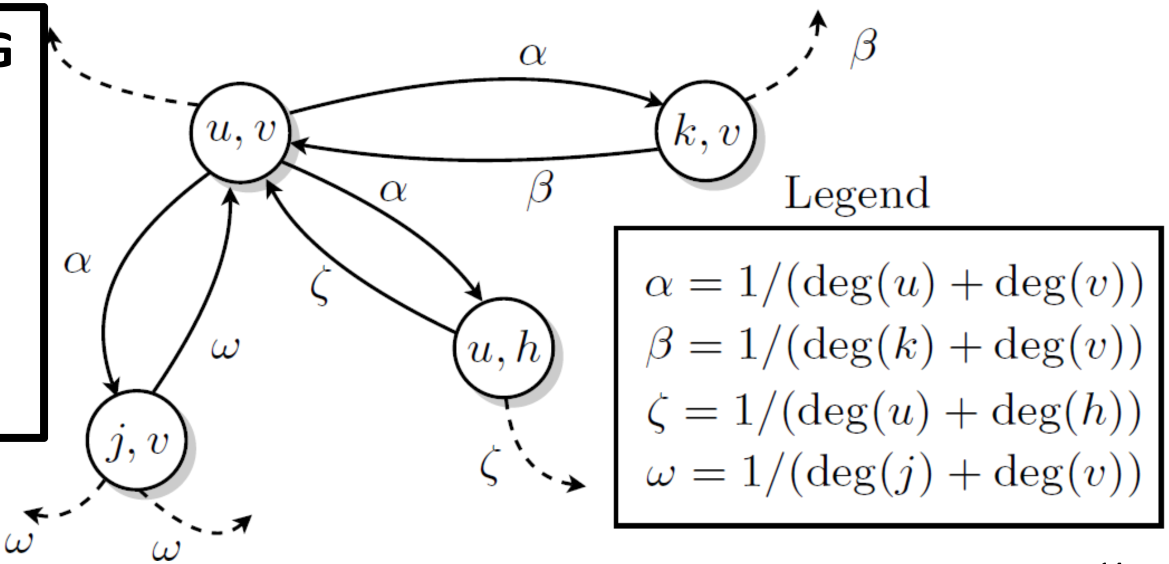
FS: An m–dimensional random walk



Graph **G**

Frontier Sampling
Discrete-time Markov Chain

- $\mathbf{G}^m$ = m-th Cartesian power of **G**
- Frontier sampling
  $\equiv$
  single random walk over $\mathbf{G}^m$

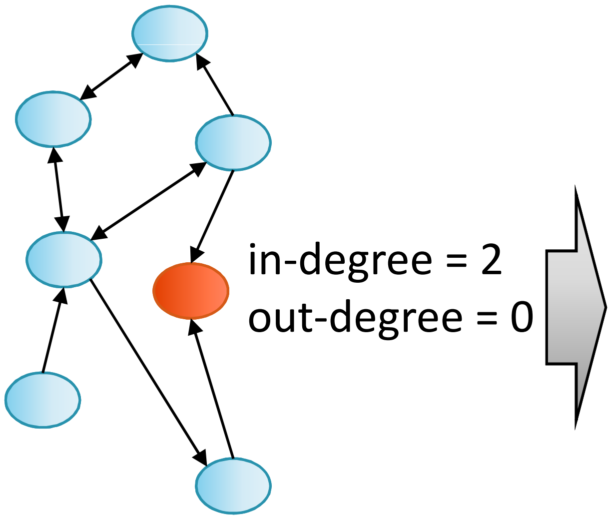


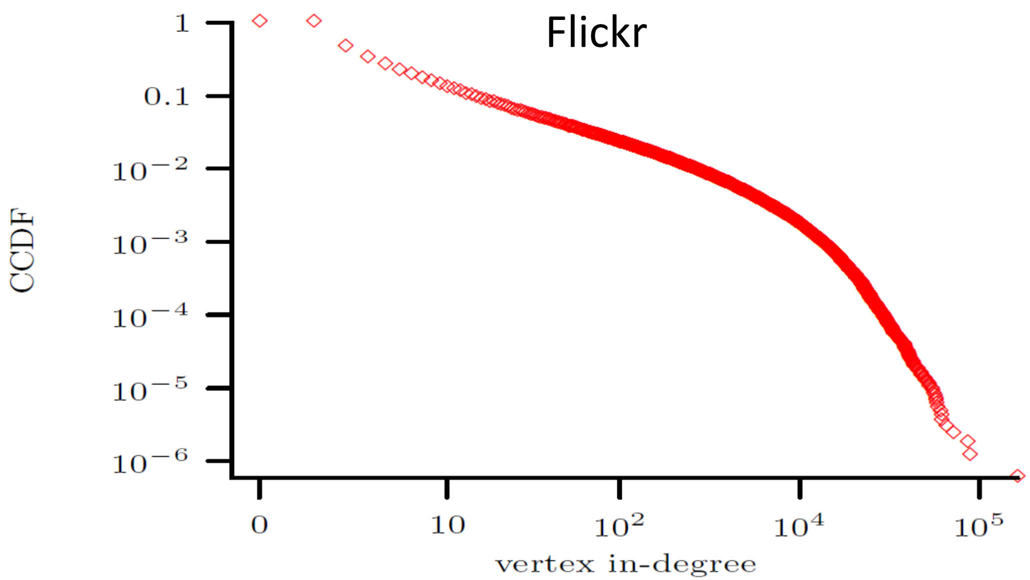
Simulation scenarios

- Flickr graph (Mislove 2007), 1.7M vertices, 5M edges.
  Largest connected component = 1.6M vertices

- LiveJournal graph, 5M vertices, 77M edges

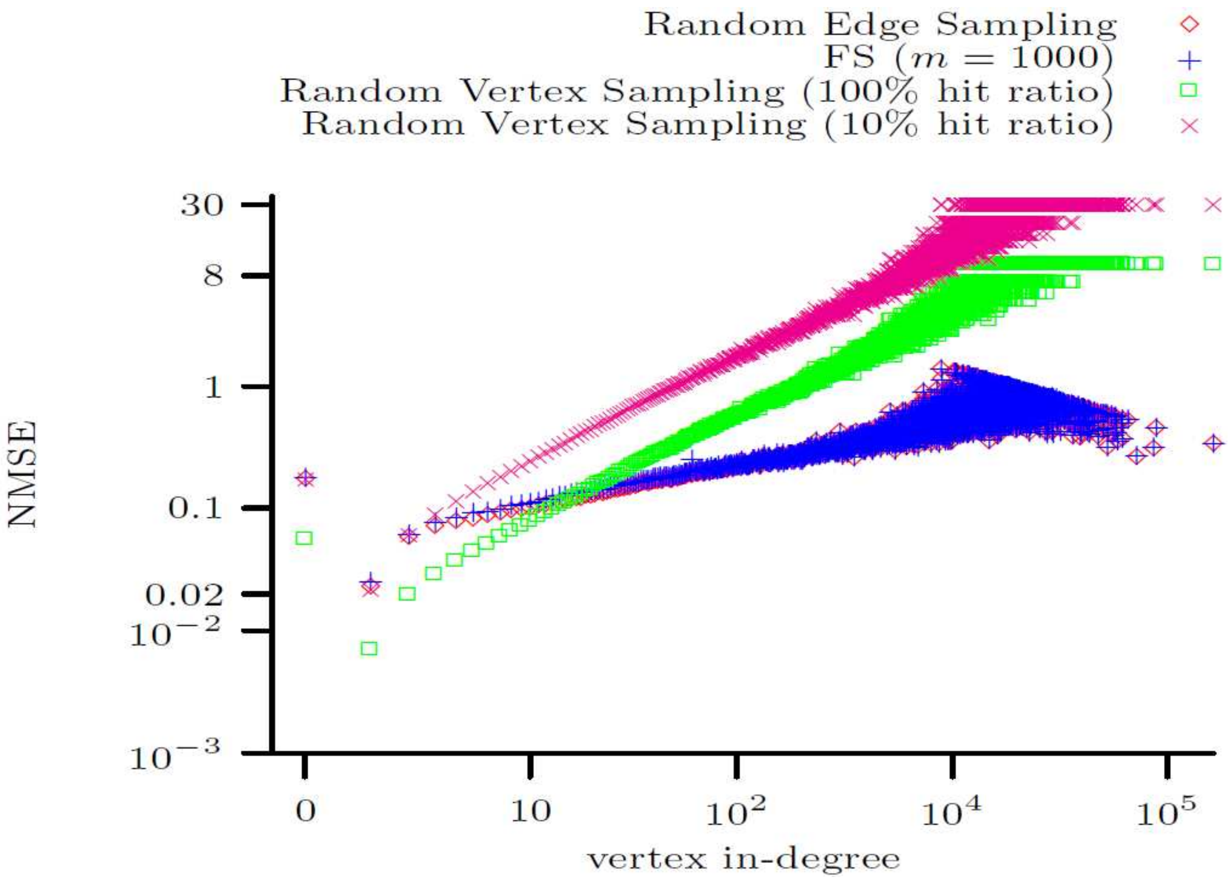
- **Objective:** Estimate the fraction of vertices with in-degree i





FS v.s. Independent sampling

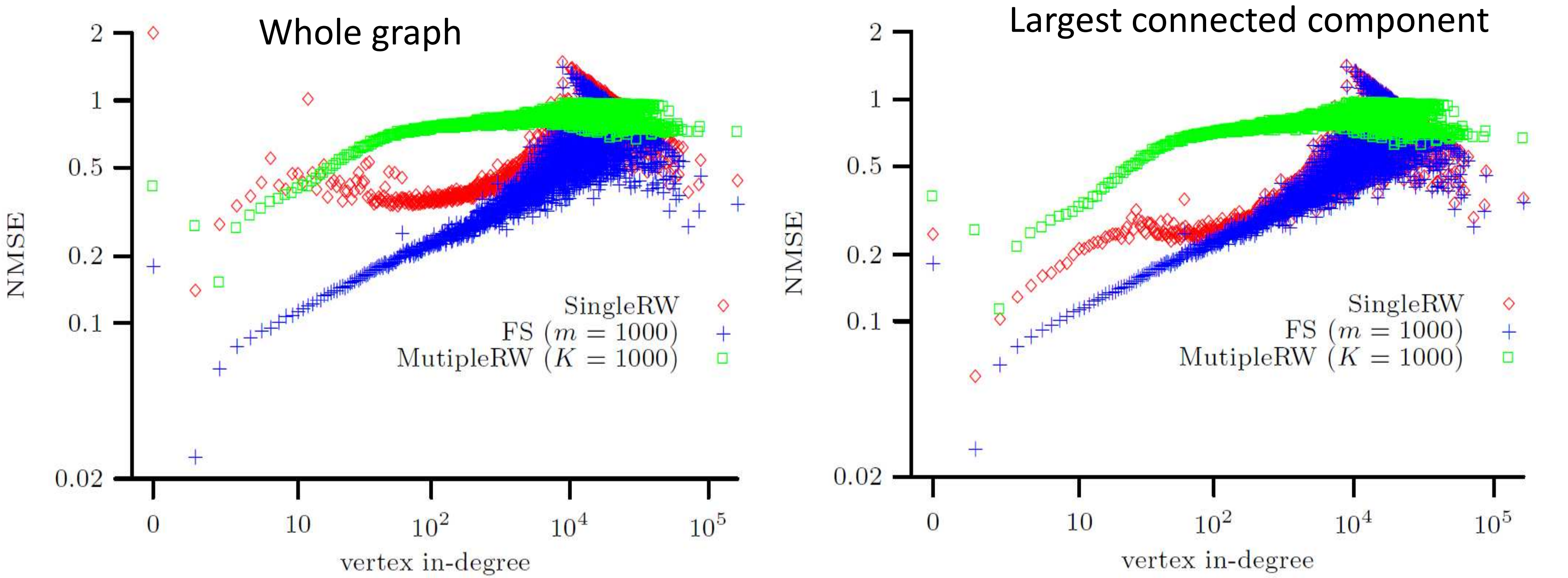
- LiveJournal graph
- Budget = 1% vertices



FS almost as good as independent edge sampling!

FS v.s. RW

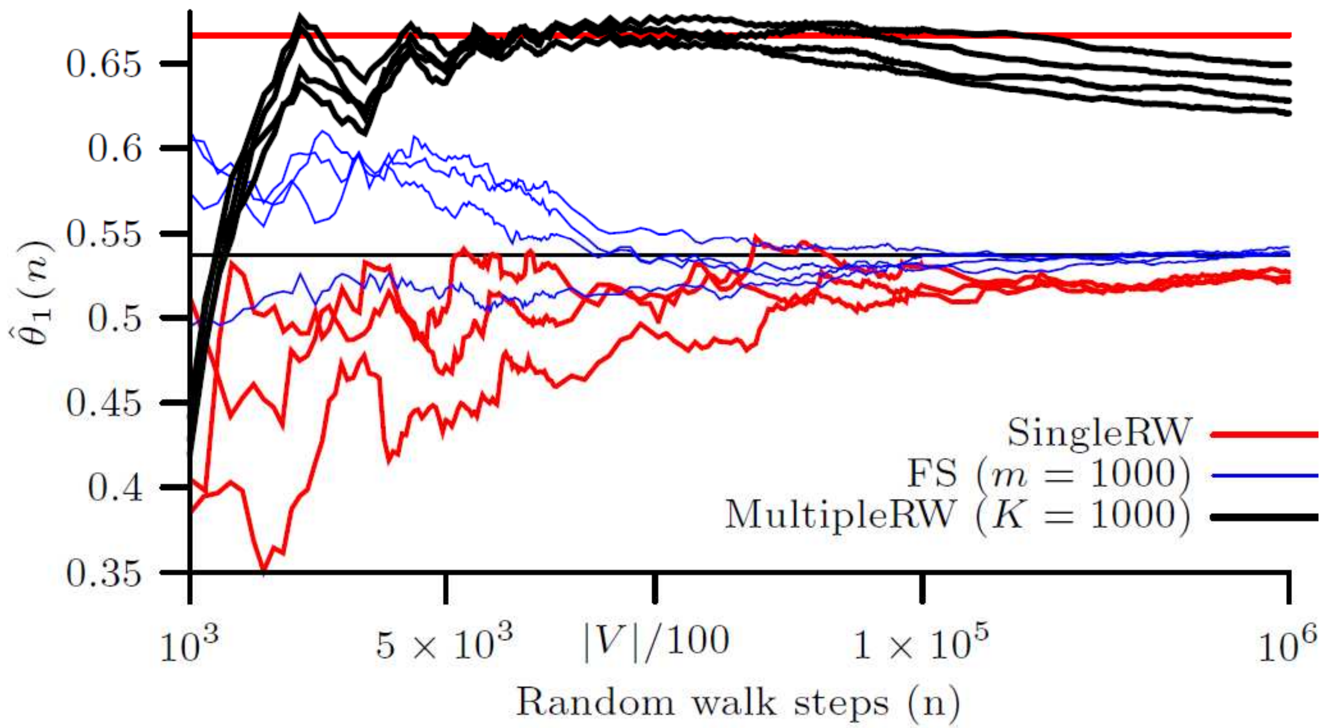
- Flickr graph
- Budget = 1% vertices



FS more accurate than random walks

Sample paths (whole graph)

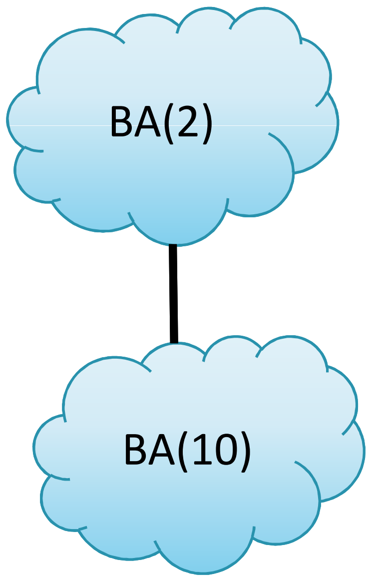
- Plot evolution $\hat{\theta}_1$(n) , where n = number of steps
- 4 sample paths = 4 curves

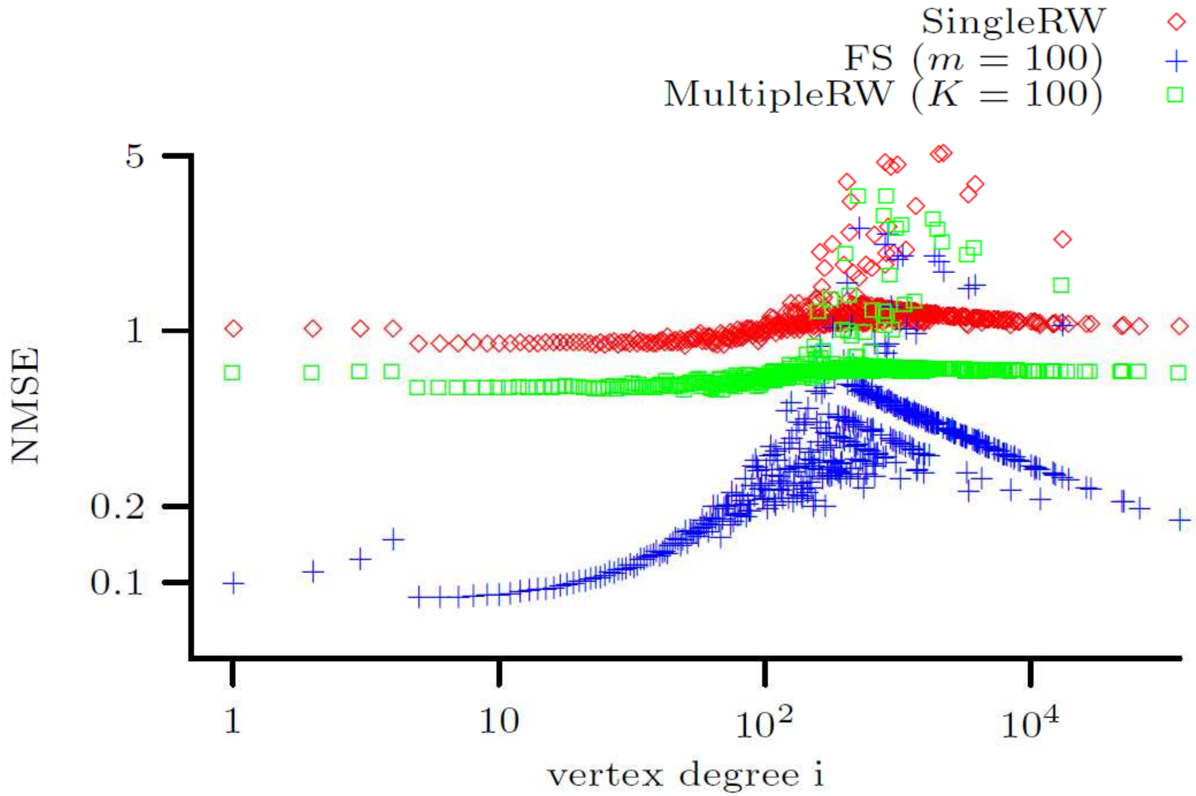


Controlled experiment

- BA(k) – Barabási-Albert graph with average degree k
- Budget = 10% vertices

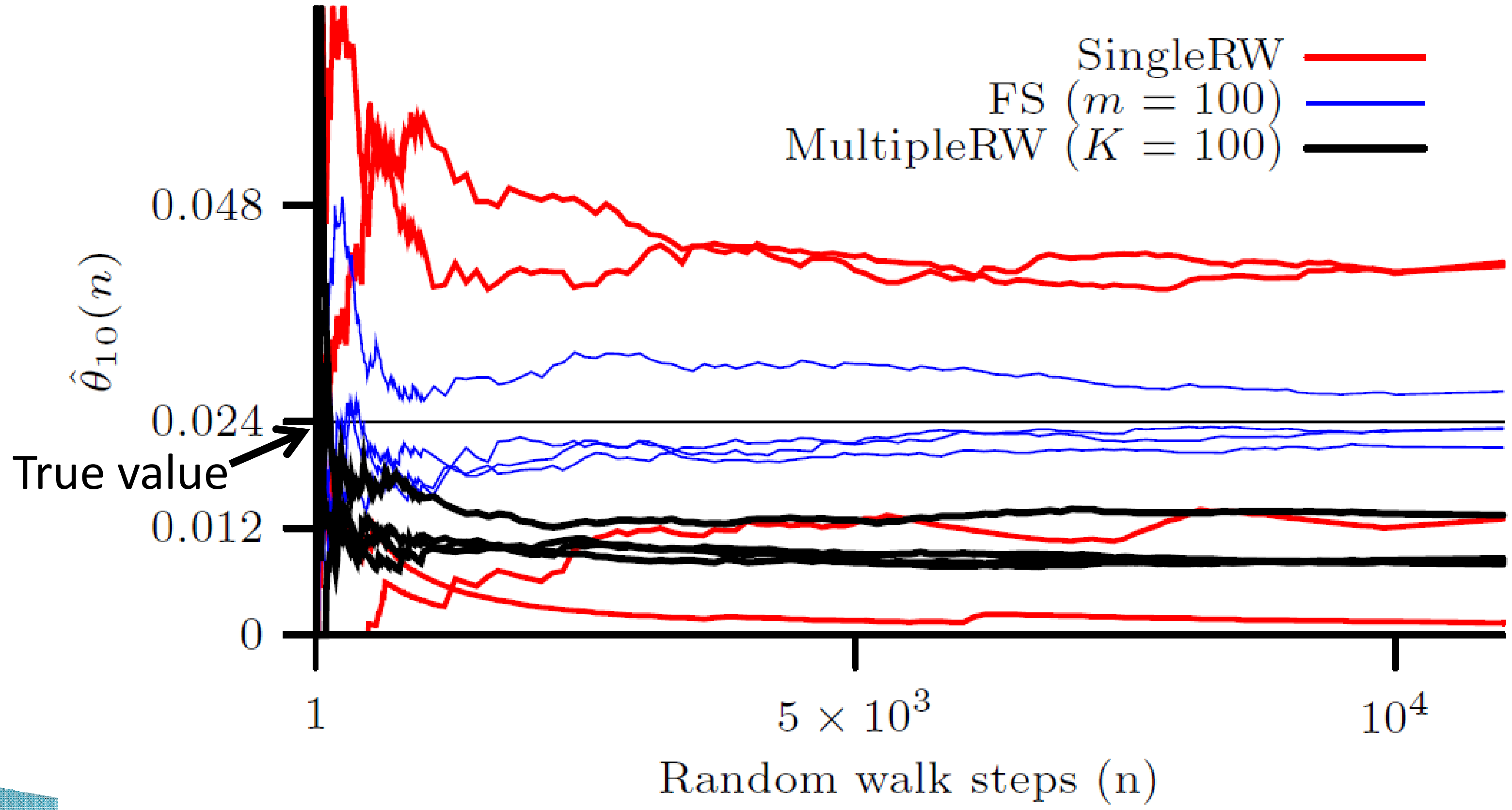
Graph





Controlled experiment (cont)

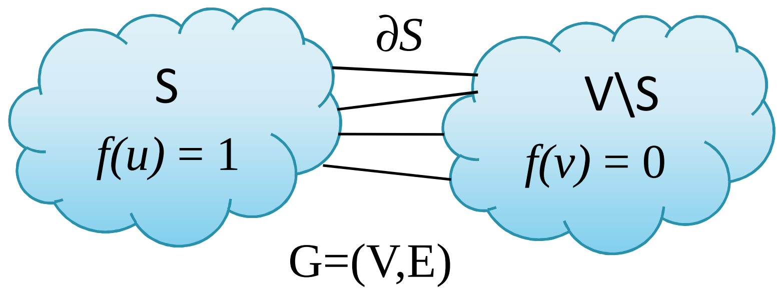
- Plot evolution $\hat{\theta}_{10}$, where n = number of steps
- 4 sample paths = 4 curves



Q: could we estimate clusters? (tentative)

- the graph conductance (normalized cut)

$$h_G = \inf_S \frac{|\partial S|}{\min\{\mathrm{vol}(S), \mathrm{vol}(V \backslash S)\}}$$



- can be estimated from the Dirichlet quotient

$$R_S(f) = \frac{\sum_{\forall (u,v) \in E} (f(u) - f(v))^2}{\sum_{\forall u \in S} f(u)^2 \deg(u)}$$

Dirichlet (experiment):
- *f(u)* = 0 if *id(u)*= odd
- *f(u)* = 1 if *id(u)*= even

Graph: Flickr (LCC)
|V|/10 steps
true R_S(f) = 0.00103

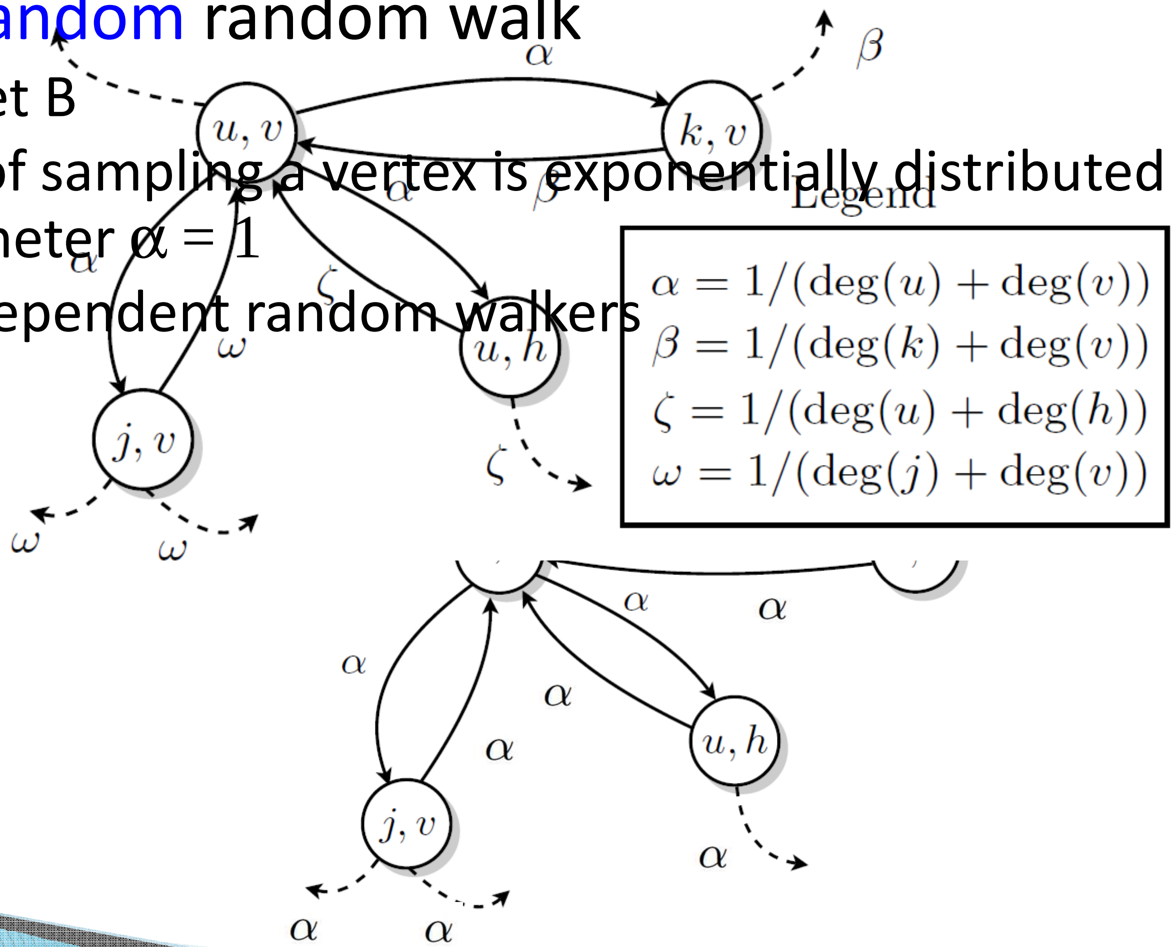
estimated: R_S(f) = 0.00103
FS: NMSE = 0.31
RndEdge: NMSE= 0.18

Frontier sampling (FS)

- must be centrally coordinated?

Discrete-time Markov Chain

- FS: A random random walk
  - Budget B
  - Cost of sampling a vertex is exponentially distributed with parameter $\alpha = 1$
  - m independent random walkers



Conclusions

- Random walks are promising approach
- Real world graphs demand new random walk strategies
- Multiple independent random walks not enough
- Dependent random walks are a powerful and unexplored

A lesson from the past

the Portuguese
"World Map" in 1459

- proved incomplete (Columbus et al. 1492)
- wrong proportions

Lesson:

understanding our "world" requires principled measurement methods

The Fra Mauro world map (1459)

source: Wikipedia