Sampling Multivariate Heavy-tailed Distributions in Networks

Don Towsley

Umass-Amherst

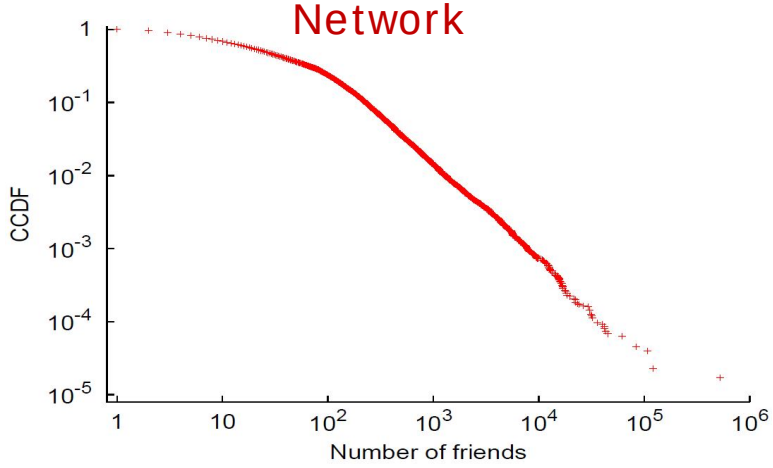
Motivation: social networks

Statistical characterization

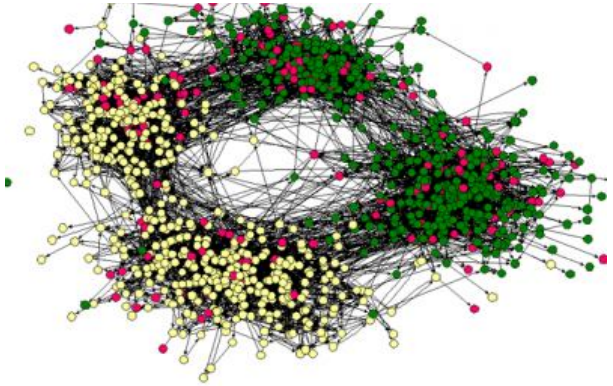
- friends?
- followers? followings
- clustering?
- motifs?
- centrality?

Often heavy-tailed

Myspace Online Social Network



High school friendship network

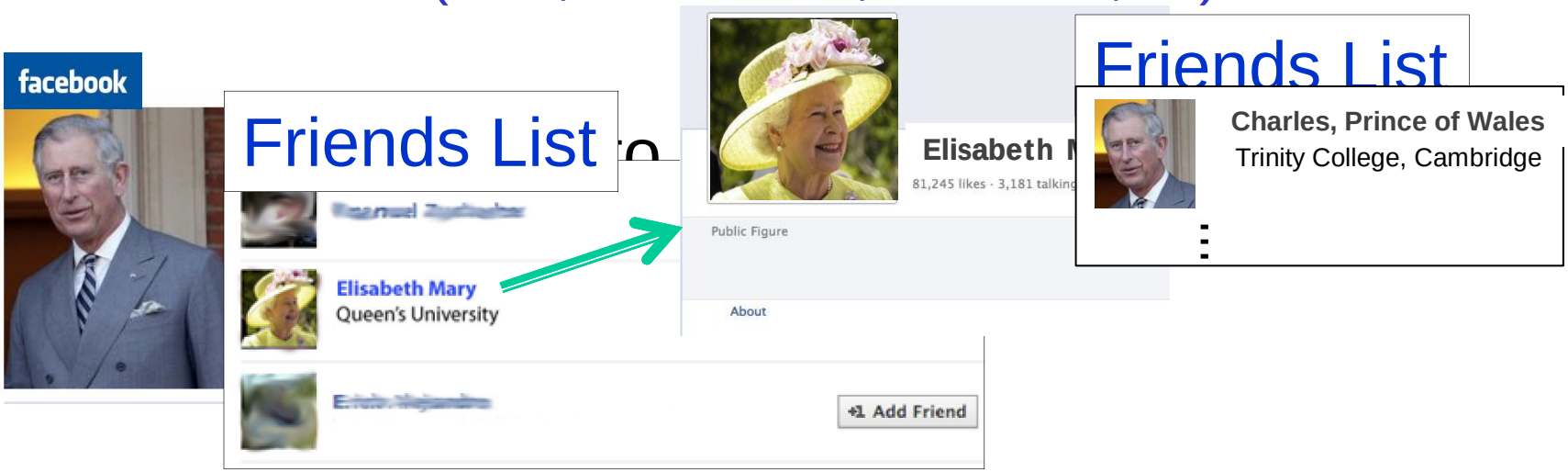


Outline

- motivation
- characterizing graphs
- sampling with random walks
  - undirected, directed graphs
  - degree distribution
- summary

On-line social networks

Can pick up node degree and neighbors at each visit (web, FaceBook, LinkedIn, …)



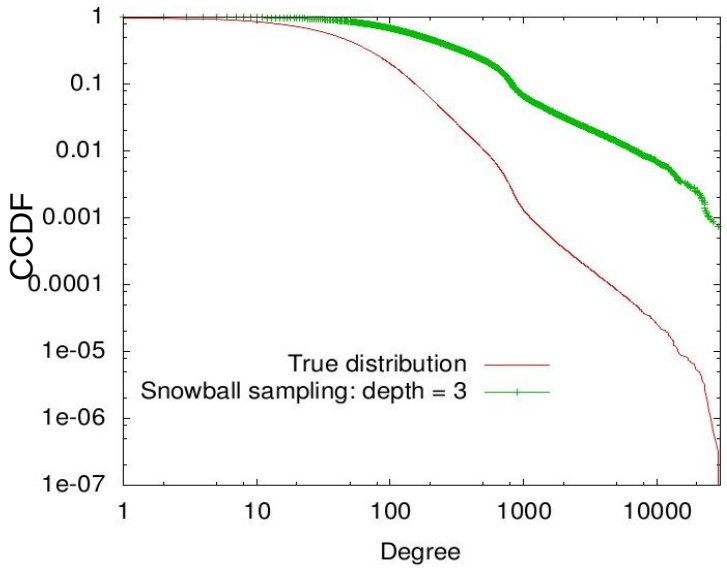
How then? sampling/crawling

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

Sampling vs. crawling

- sampling
  - random node sampling
    - unbiased estimates
    - expensive
- crawling
  - snowball sampling, breadth first search
    - biased estimates
  - random walk (RW)
    - select next node uniformly from neighbors

Orkut OSN, >2M nodes



RW sampling: undirected graph

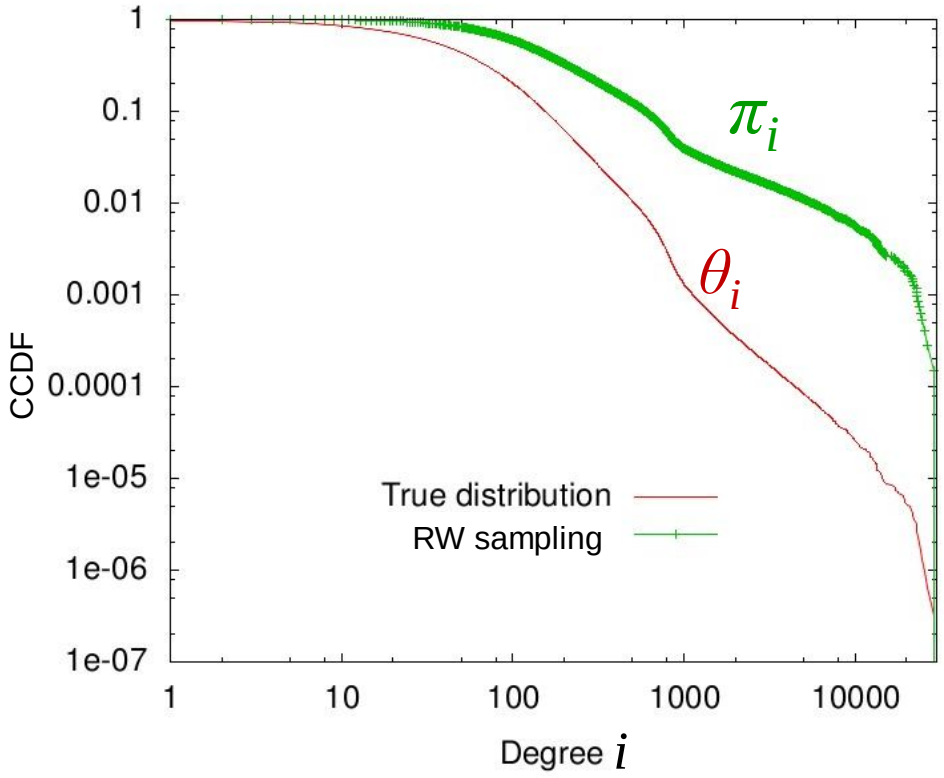
Bias removal?

- Markov model
- at steady state visits edges uniformly at random

Model:

θ_i - P[vertex degree = i]

π_i - P[visited degree = i]



RW sampling: undirected graph

Bias removal?

- Markov model
- at steady state visits edges uniformly at random

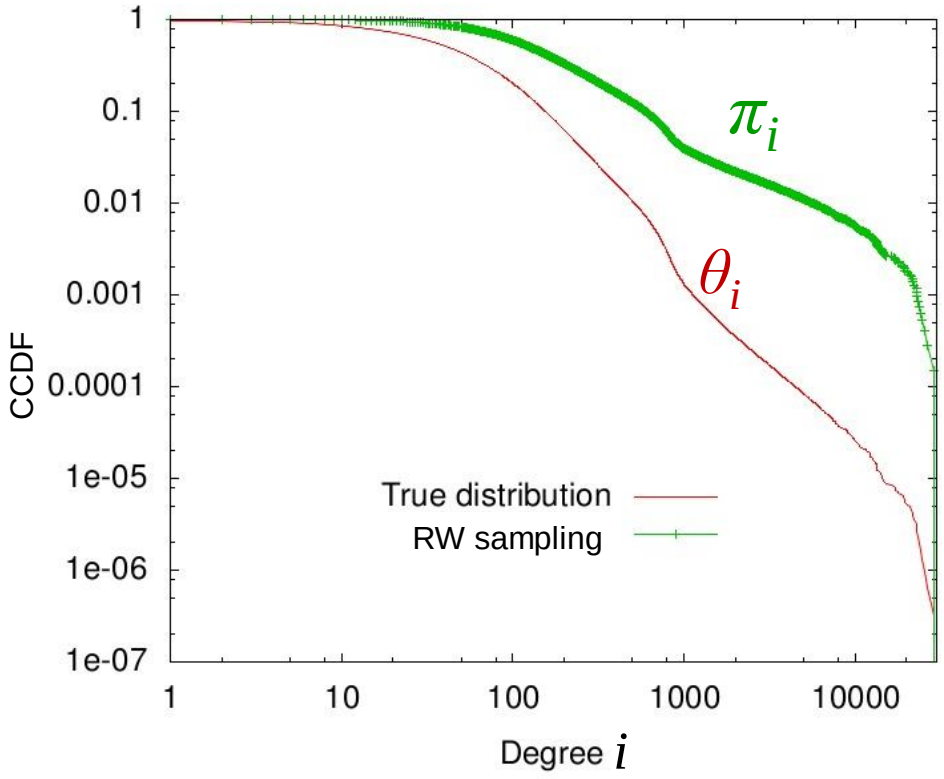
Model:

θ_i - P[vertex degree = i]

π_i - P[visited degree = i]

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

or $\theta_i = \text{Norm} \times \pi_i / i$



produces asymptotic unbiased estimates

Sampling errors

- estimate θ_i (avg. degree $\bar{d}$) with B samples
- error metric

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

- random node sampling

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

smaller if $i < \bar{d}$

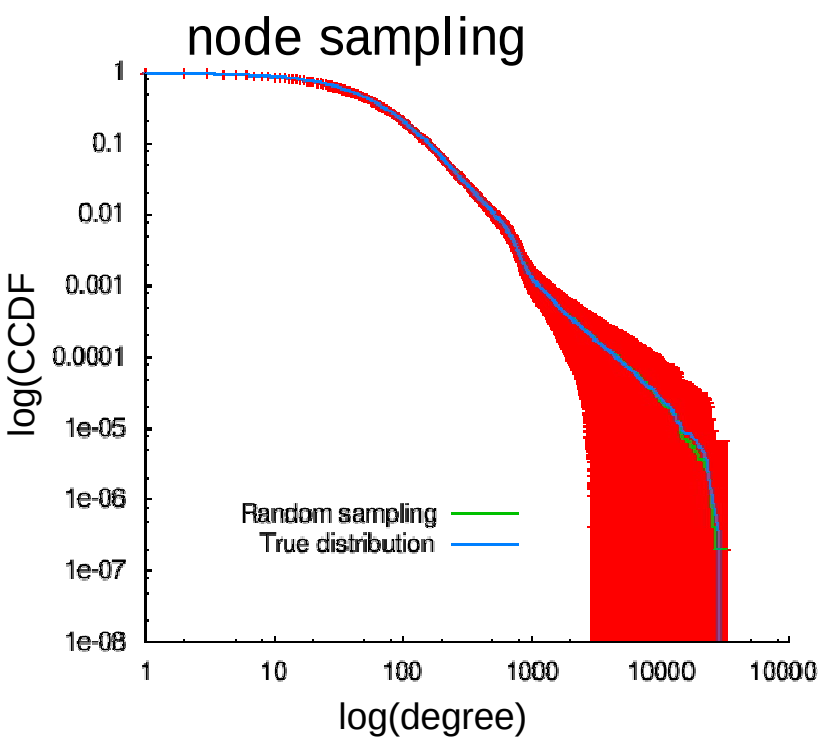
- random walk sampling (≈ random edge)

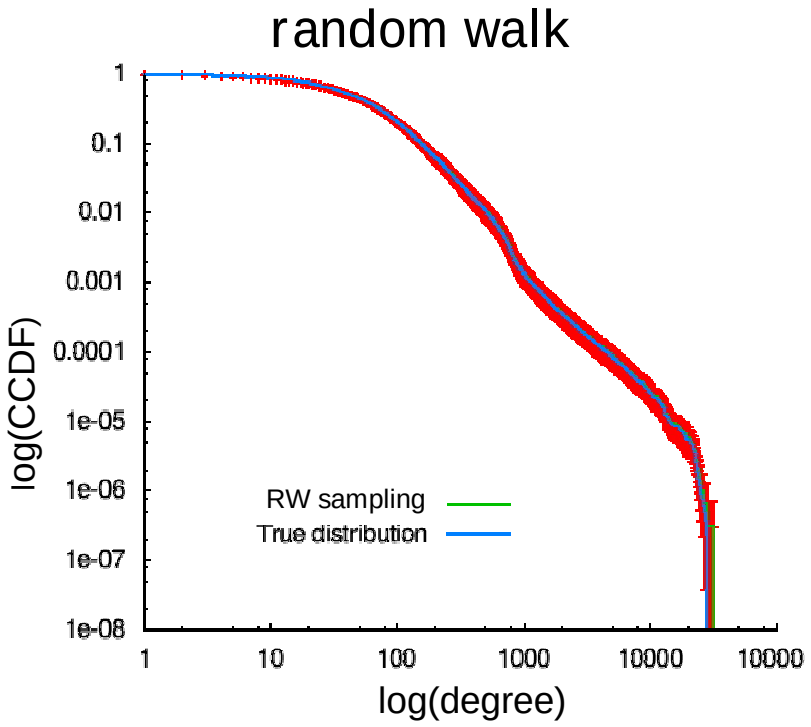
$$NRMSE(i) = \sqrt{((\bar{d}/i)(1/\theta_i) - 1/B}$$

smaller if $i > \bar{d}$

Heavy tails more accurate with RW sampling

Node sampling vs. RW: Orkut





RW sampling effective for estimating heavy tails

Directed graphs: hidden in-edges

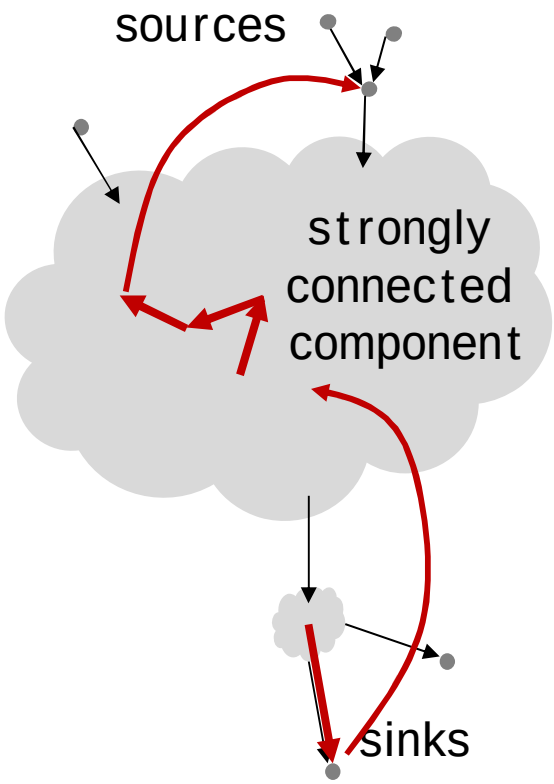
Challenges: sources, sinks

- walk outgoing edges
- add random jumps

Problem: RW steady state distribution not computable

Solution (DURW):

- during walk, construct undirected graph consistent with walk
- walk undirected graph on revisits
- use undirected RW estimation



Estimating joint in/out degree distribution

Impossibility result

- when indegree heavy-tailed
  - samples contain “almost no” statistical information unless > ½ of edges sampled
  - Fisher information → zero as max degree → ∞ exponentially fast

- different result when tail is light
  - estimation much easier

Directed graphs: visible edges

- transform digraph to undirected graph
- collect samples using RW

$$s_1, s_2, \dots, s_n, \qquad s_k = (i_k, o_k)$$

- estimate

$$\hat{\varphi}_{i,j} = \frac{1}{n} \sum_k \frac{h_{ij}(s_k)}{\hat{\pi}(s_k)} h_{ij}(s_k), \qquad i, j = 0,1, \dots$$

$$h_{ij}(s_k) = \begin{cases} 1, & i_k = i, o_k = j \\ 0, & \text{otherwise} \end{cases}$$

$$\hat{\pi}(s_k) = C \times \deg(s_k)$$

- $\deg(s_k)$ is degree of new undirected graph, C chosen to make $\hat{\pi}(\quad)$ a distribution

RW-based degree distribution estimation

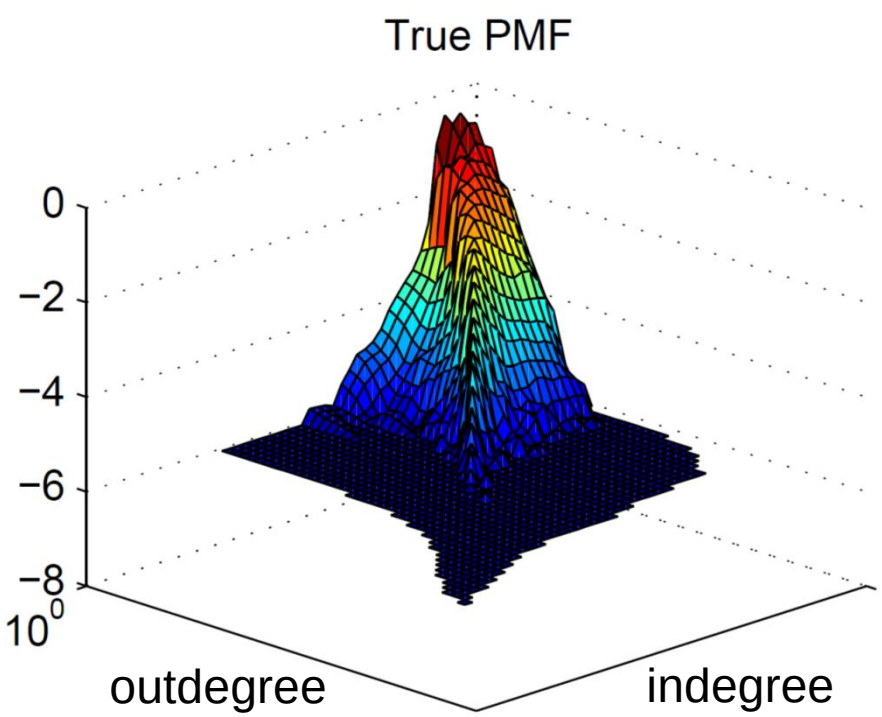
- performance on real datasets
  - vs. uniform vertex sampling
  - vs. DURW for marginal outdegree distribution

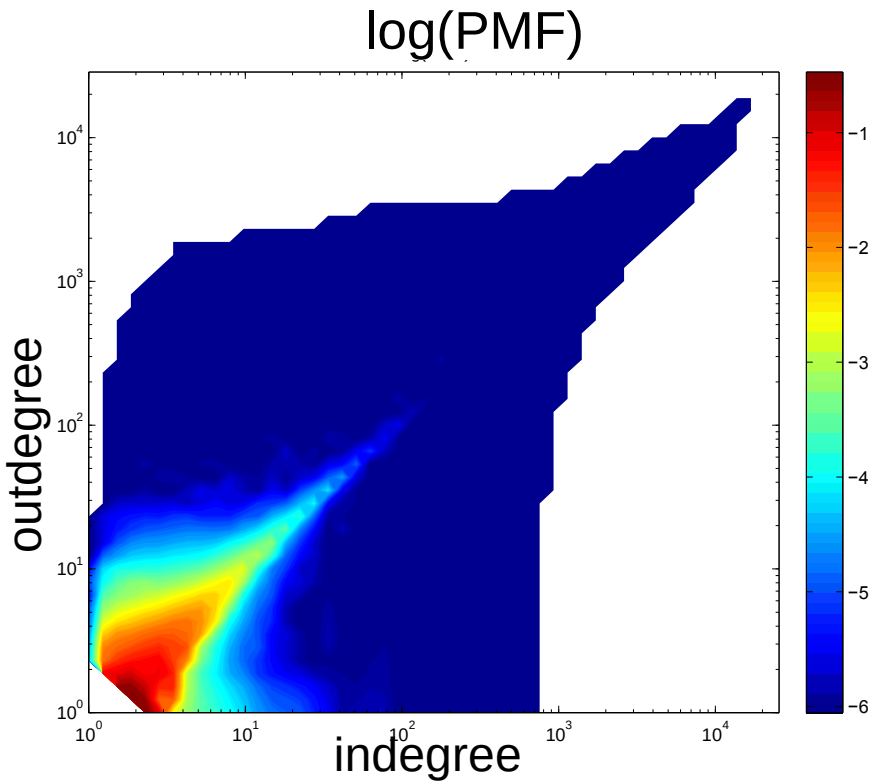
Graph	# nodes	# edges	E[out-deg]	Type
Flickr [10]	1,715,255	22,613,981	18.1	OSN
YouTube [10]	1,138,499	4,945,382.	5.3	OSN
LiveJournal [10]	5,204,176	77,402,652	18.7	OSN
Wiki-Talk [2]	2,394,385	5,021,410	3.9	usr talk
Web-Google [1]	875,713	5,105,039	9.87	Web

- in/out degree distribution heavy tailed

Behavior of RW: YouTube

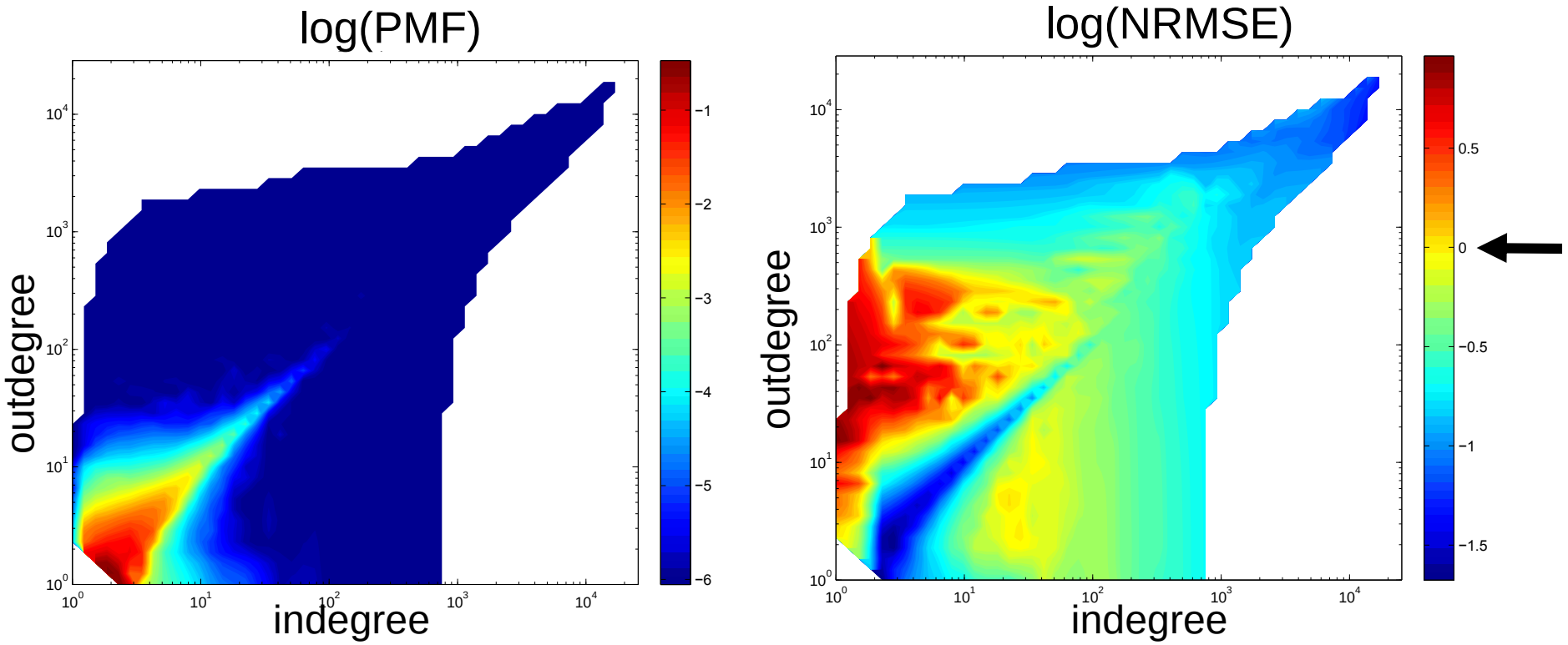
- empirical joint degree distribution
- Neyman-Pearson correlation: 0.95
- reciprocity: 0.79
  - fraction incoming edges paired with outgoing edges





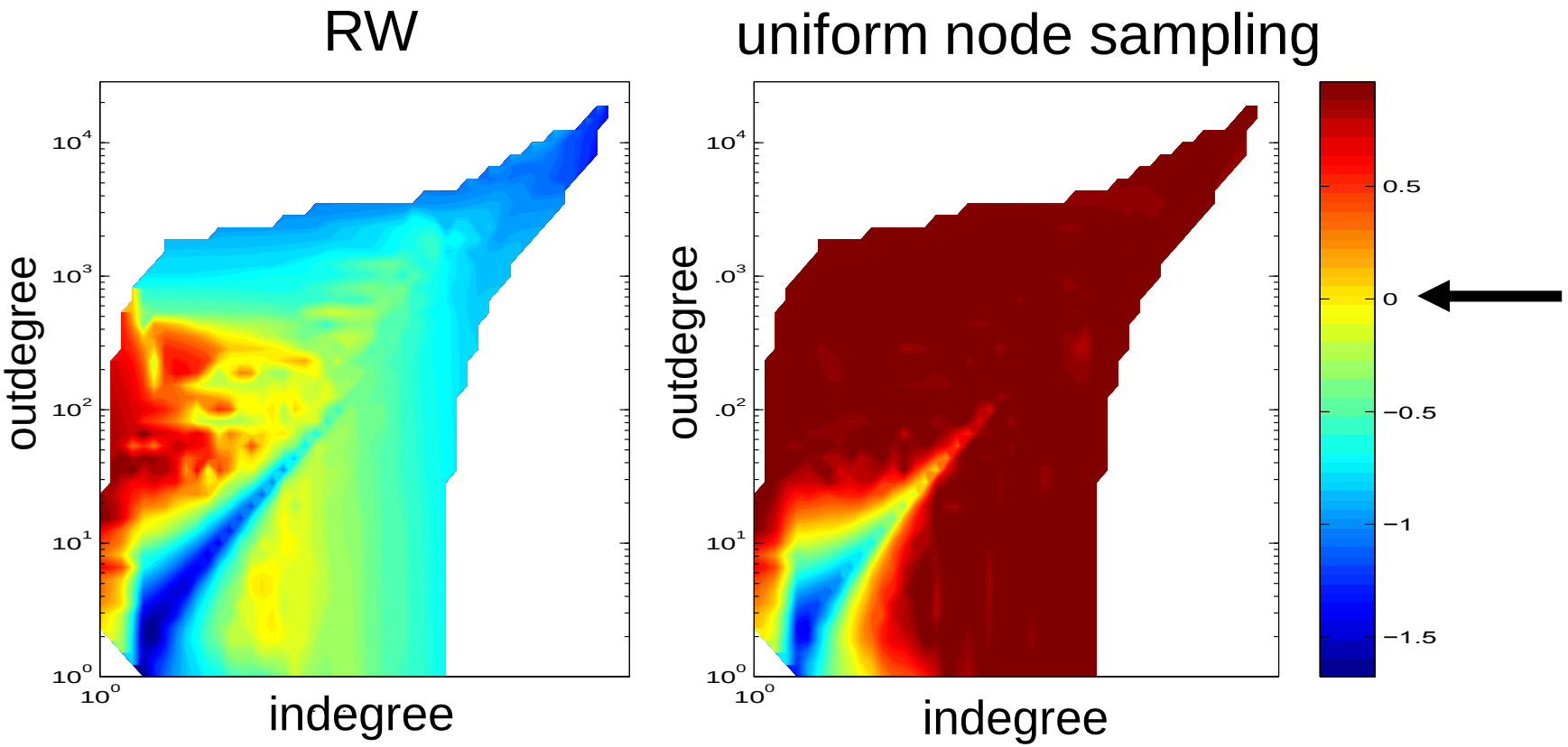
Behavior of RW on real datasets

- YouTube
- NMRSE $\mathrm{NRMSE}(\hat{\phi}_j) = \frac{\sqrt{E[(\hat{\phi}_j - \phi_j)^2]}}{\phi_j}, j = 1, 2, \ldots,$
- sampling budget 10% of graph size (B = 0.1)



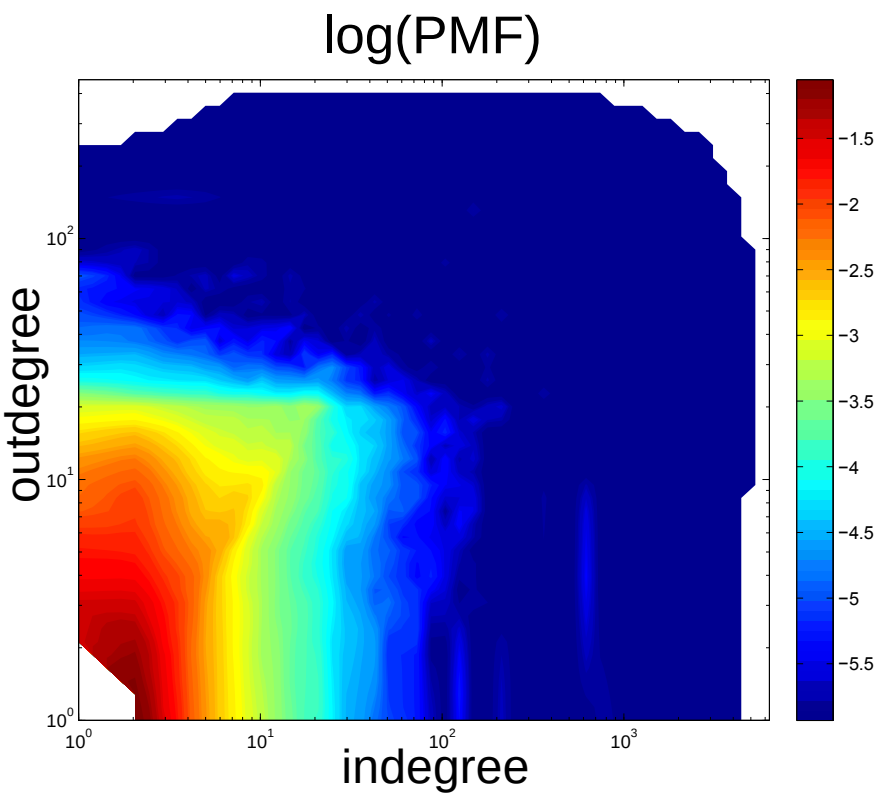
RW vs. uniform node sampling

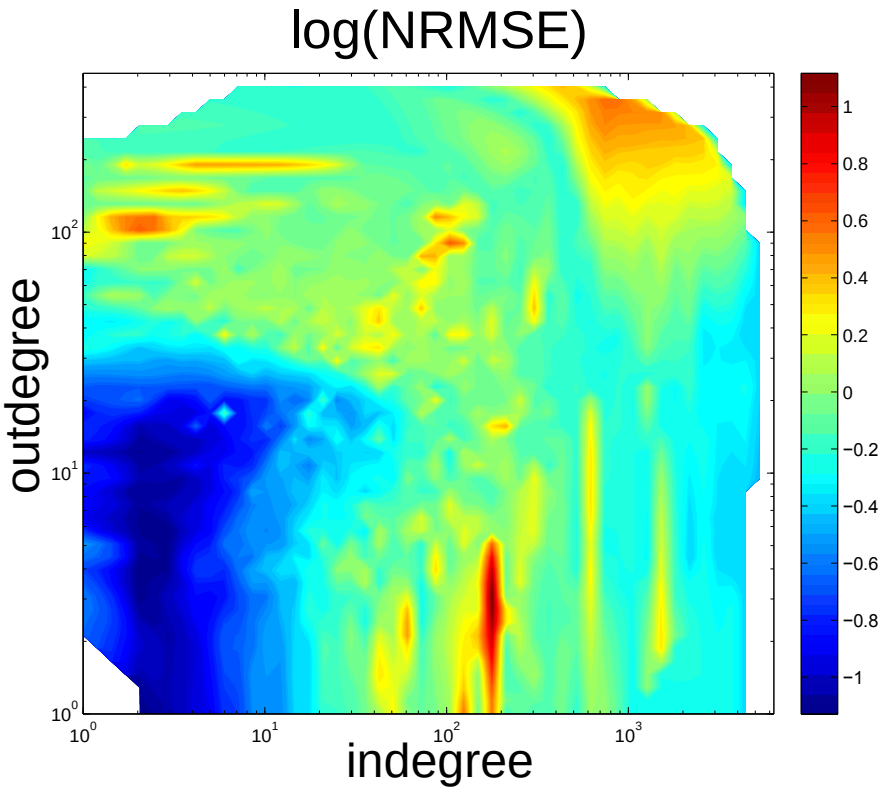
- YouTube, log(NRMSE) with B = 0.1



Web-Google

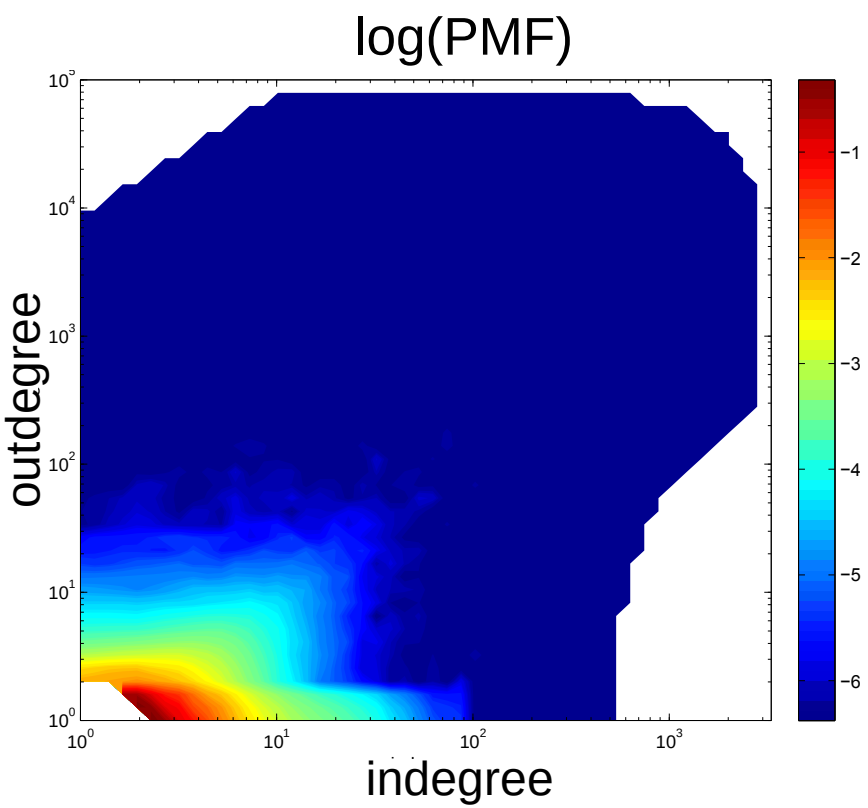
- B = 0.1

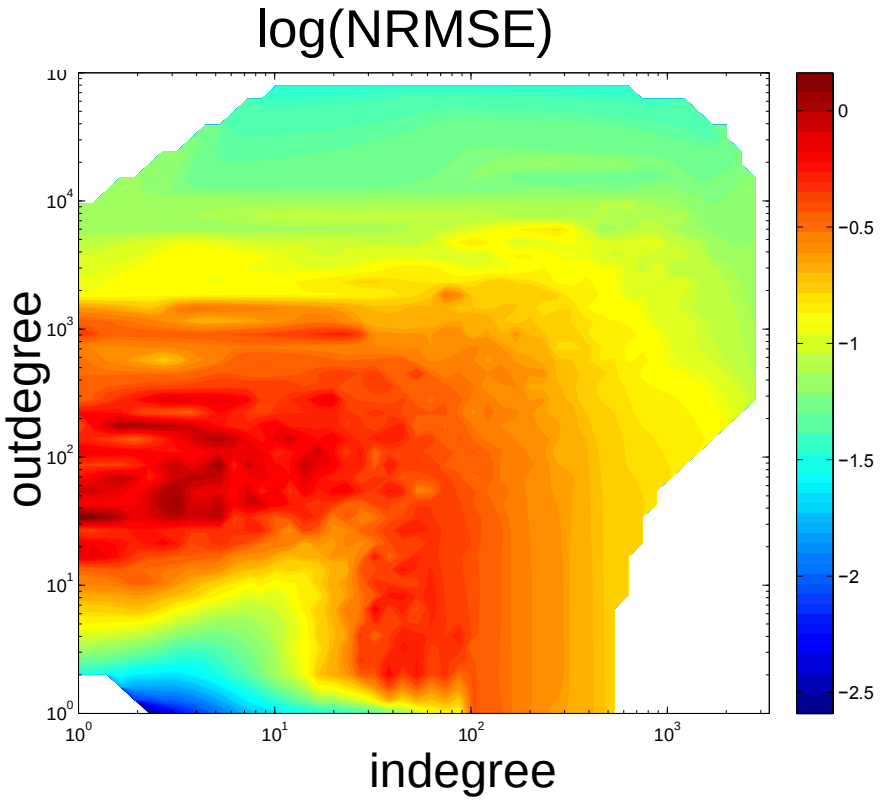




Wiki-Talk

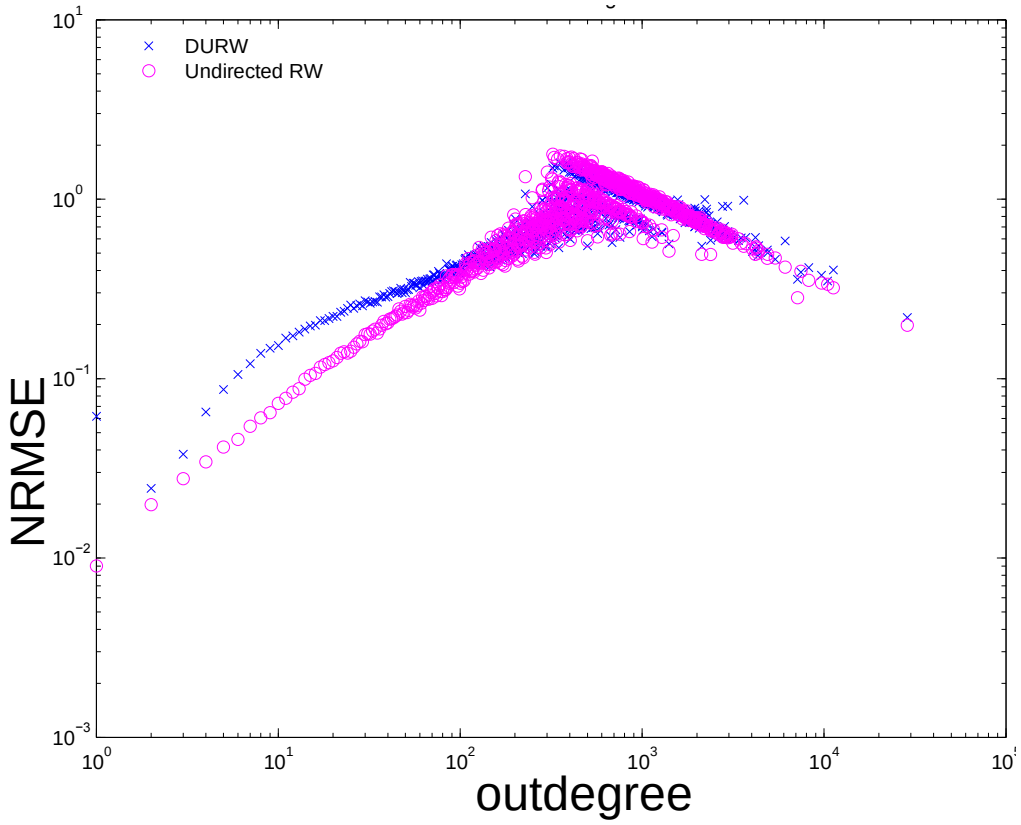
- B = 0.1





Out-degree distribution estimation: DURW vs. RW

- RW based on all edges provides (slightly) lower errors
- efficient use of indegree information?



Issues

- real world networks exhibit wide range of reciprocity, statistical dependence
- reciprocity has little effect on estimation quality
- effect of tail dependence?
- other statistics?
  - clustering coefficient
  - centrality

Summary

- random walk sampling → asymptotically unbiased estimates
- more effective than other techniques for characterizing heavy tails
  - in/out degree distribution
  - variables positively correlated with degree

Questions

- dealing with transients
  - frontier sampling – coupled RWs
- negative correlations