Welcome to

DS504/CS586: Big Data Analytics

Graph Mining

Prof. Yanhua Li

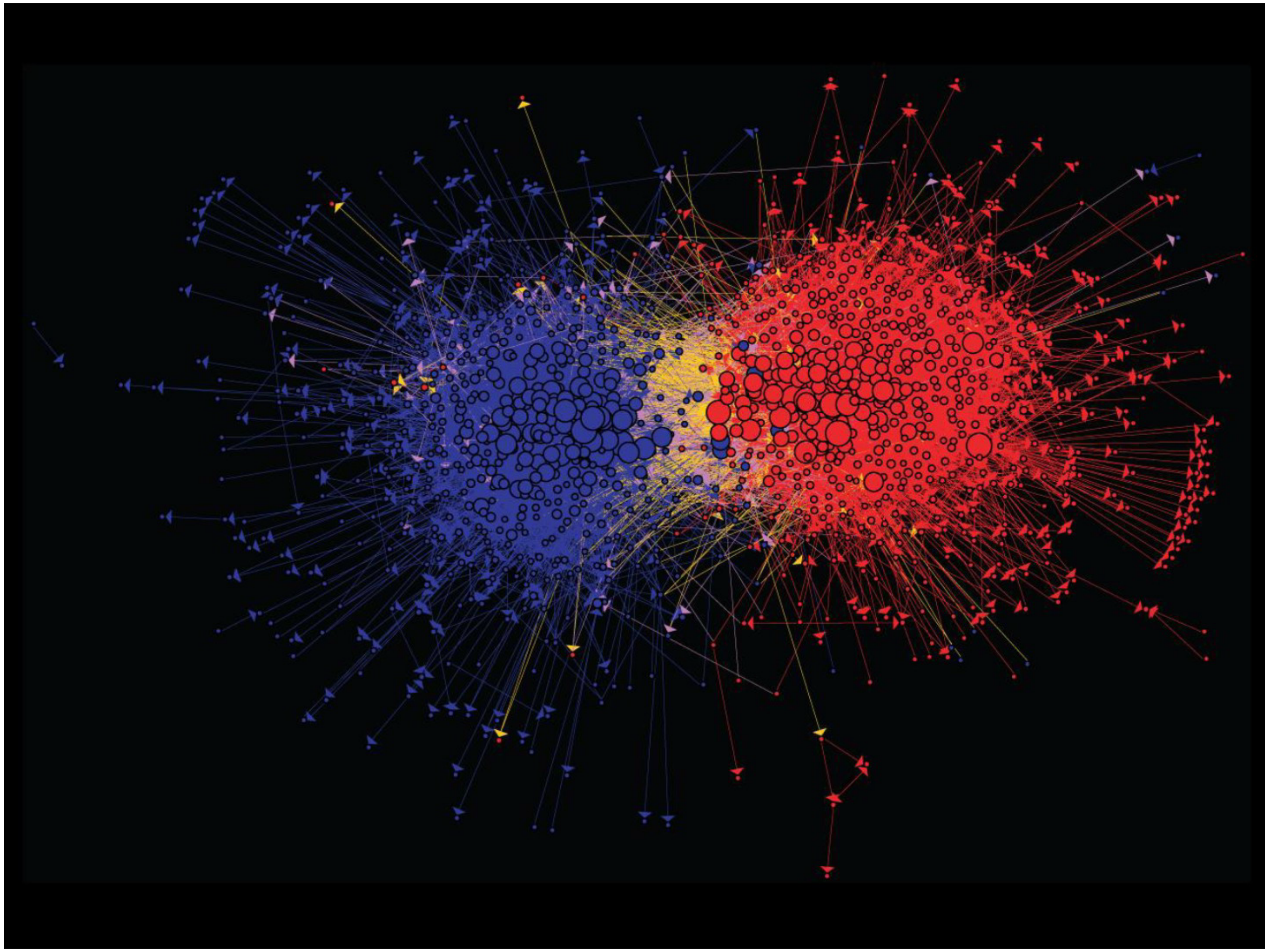
Time: 6:00pm –8:50pm R
Location: AK232
Fall 2016

Graph Data: Social Networks

Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

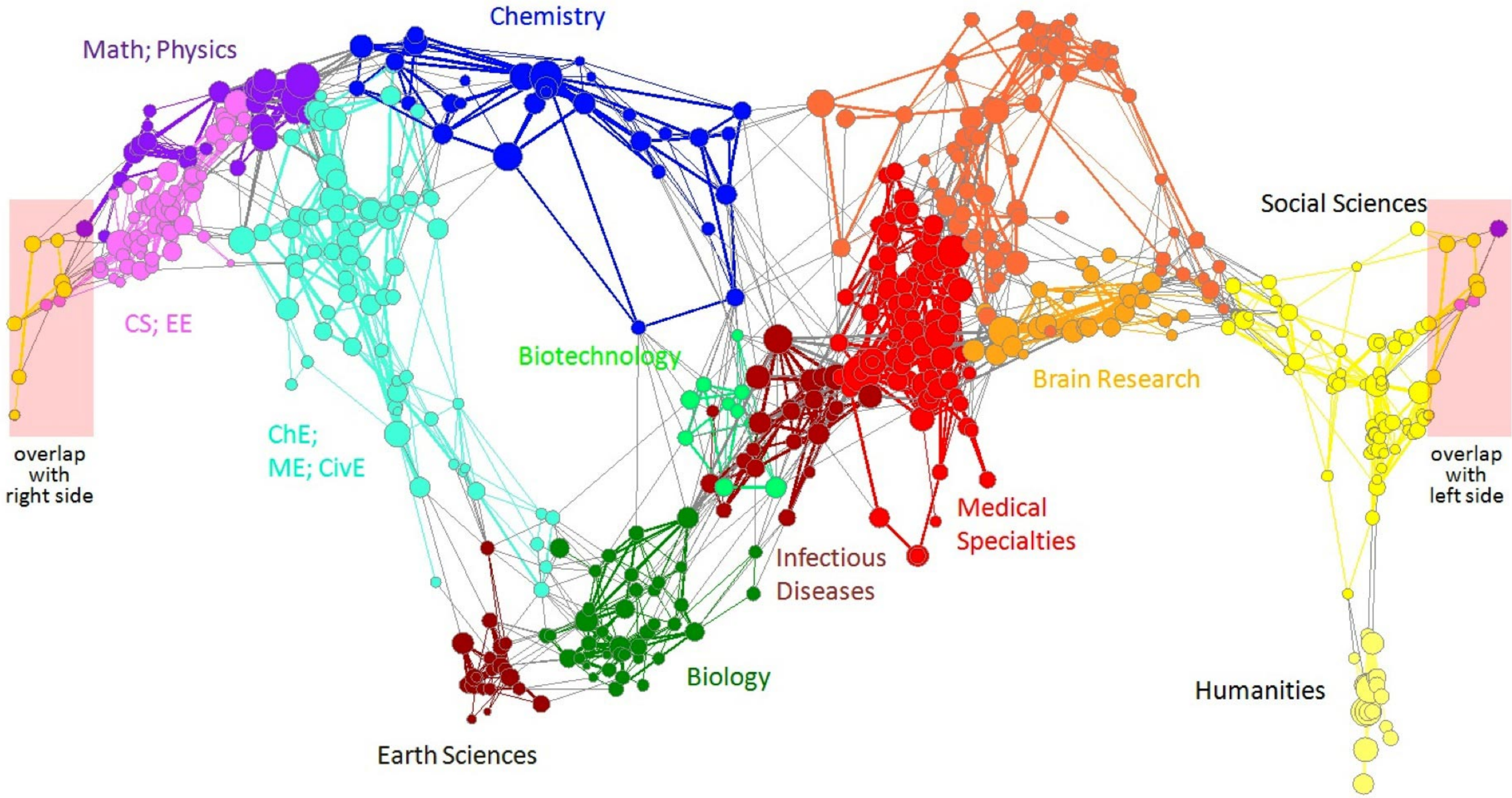
Graph Data: Media Networks



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

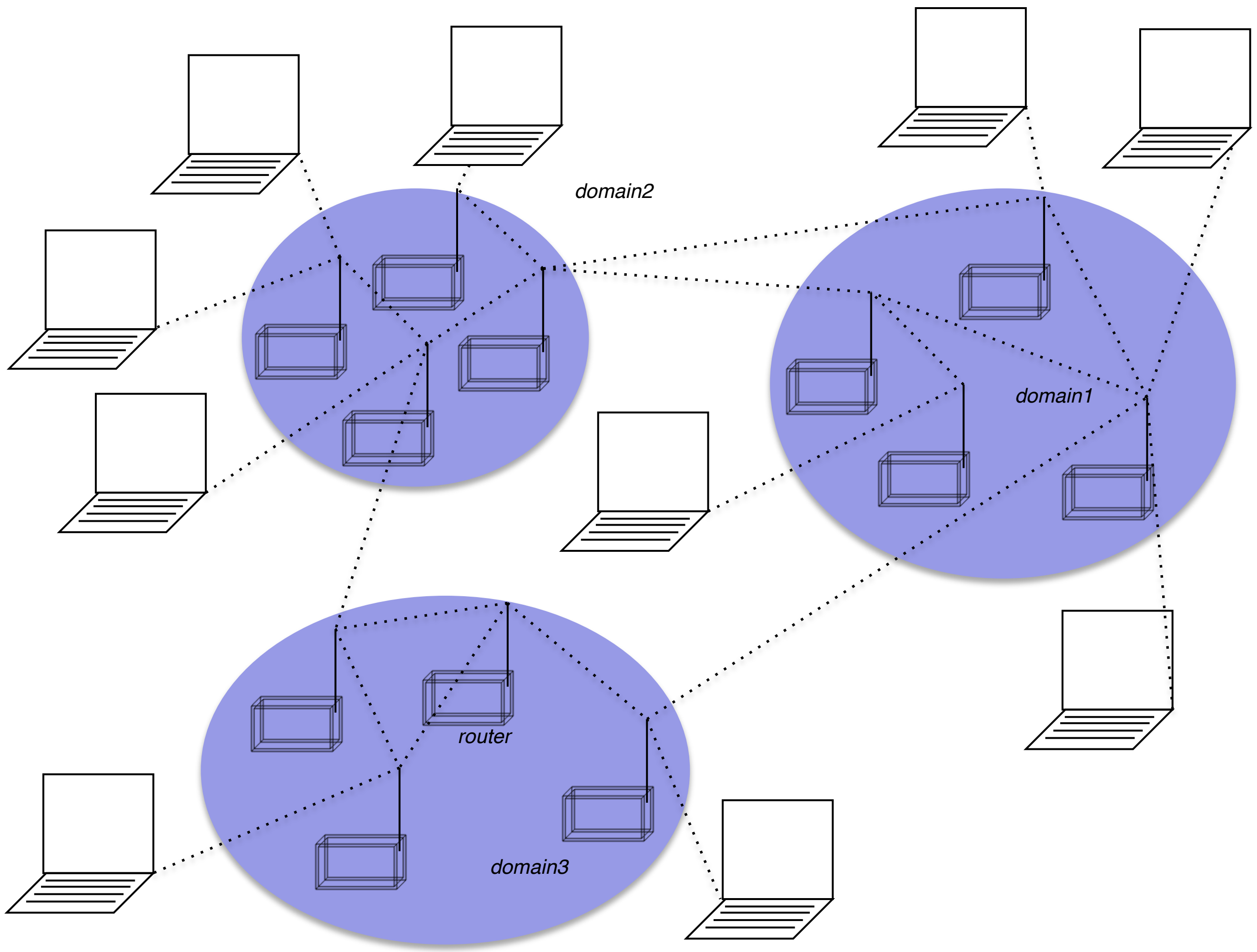
Graph Data: Information Nets



Citation networks and Maps of science

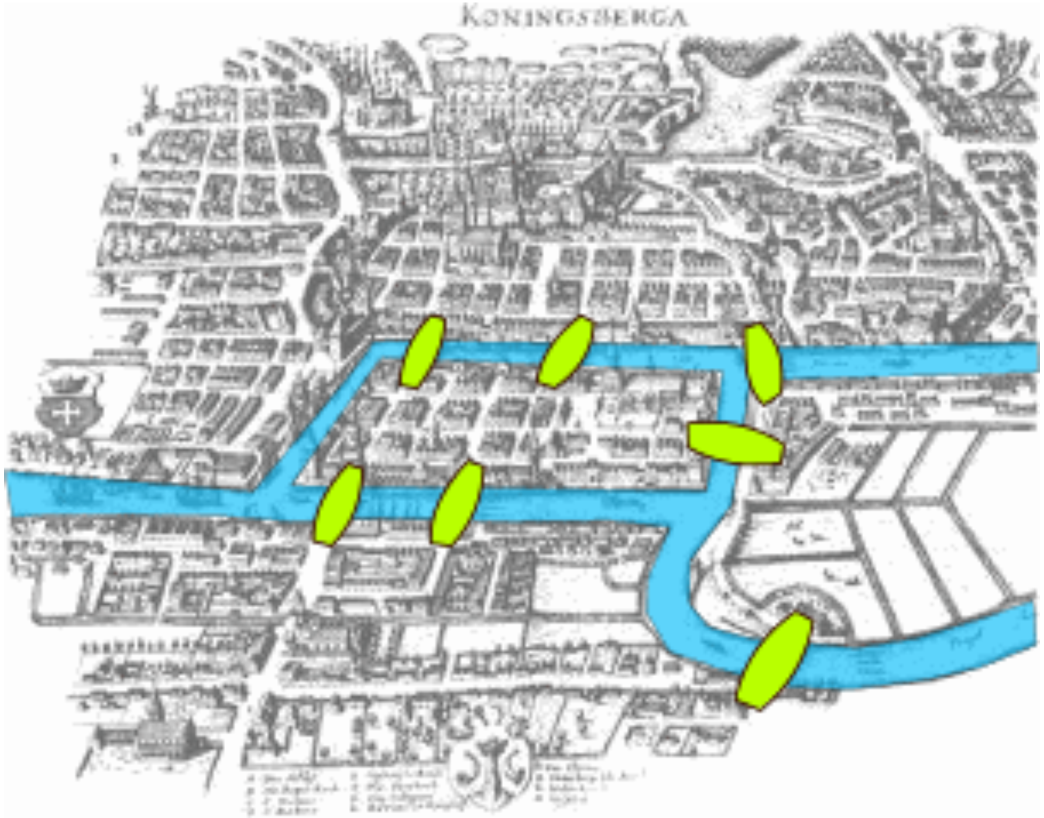
[Börner et al., 2012]

Graph Data: Communication Nets



Internet

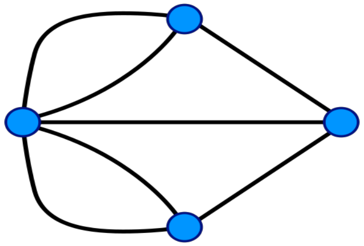
Graph Data: Topological Networks



Seven Bridges of Königsberg

[Euler, 1735]

Return to the starting point by traveling each link of the graph once and only once.

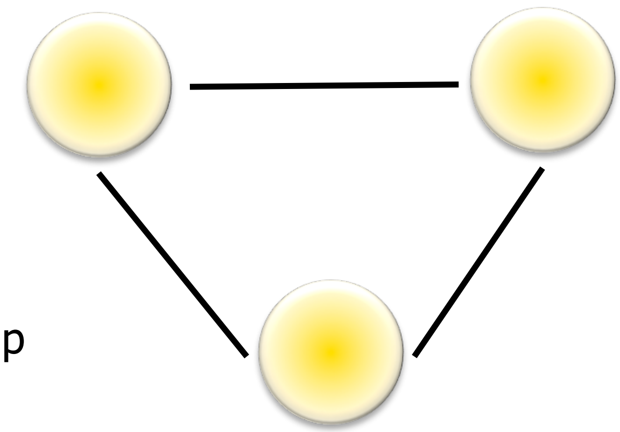


Graph representation of networks

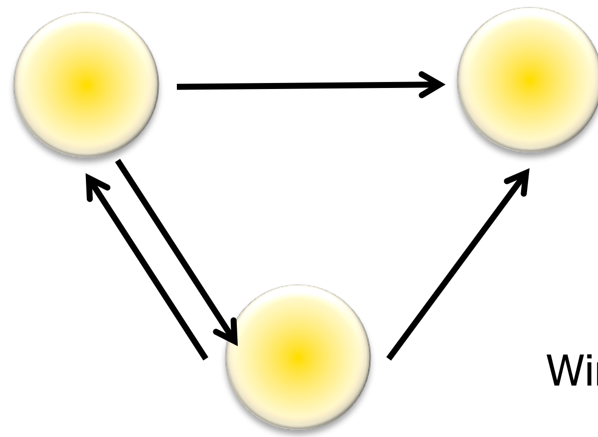
Friendship

Resistance

Co-authorship



Undirected links



Following

One-way road

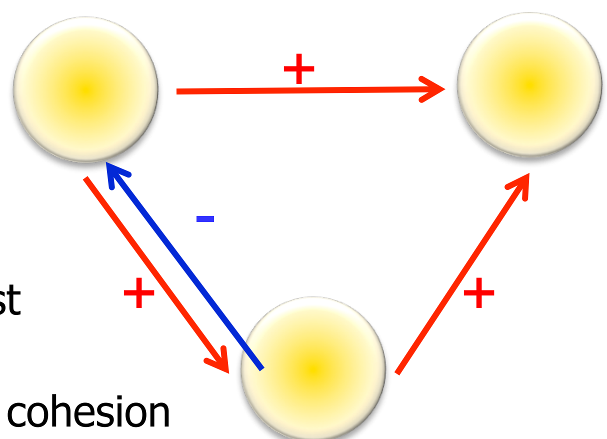
Wireless channel

Directed links

Friend & foe

Trust & distrust

Repulsion & cohesion



Signed links

Multi-relational links

Hyperlinks

… …

Mining in Big Graphs

- Network Statistic Analysis (this lecture)
  - Network Size
  - Degree distribution.

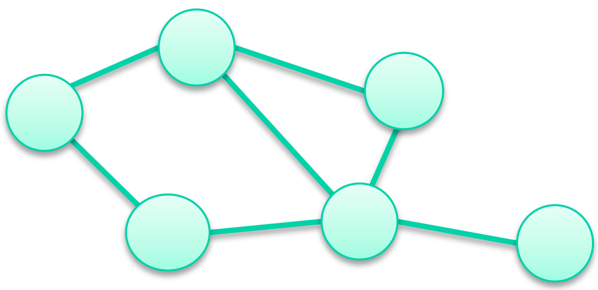
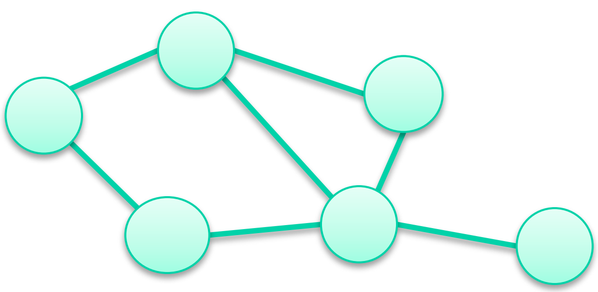
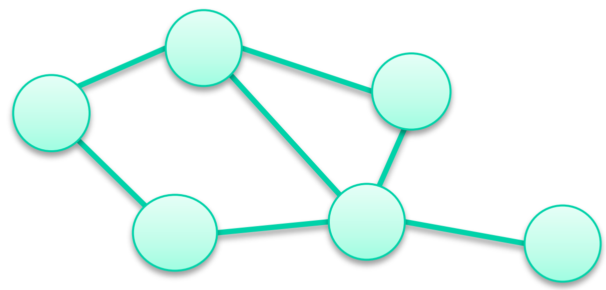
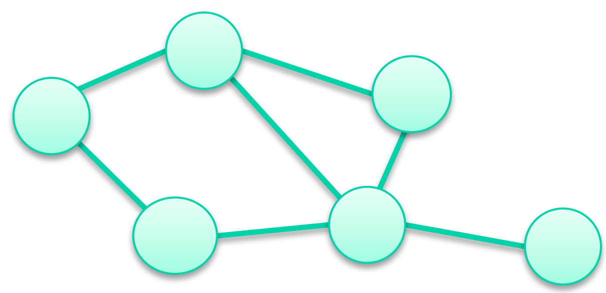
- Node Ranking (Next lecture)
  - Identifying most influential nodes
  - Viral Marketing, resource allocation

Graph Data: Social Networks

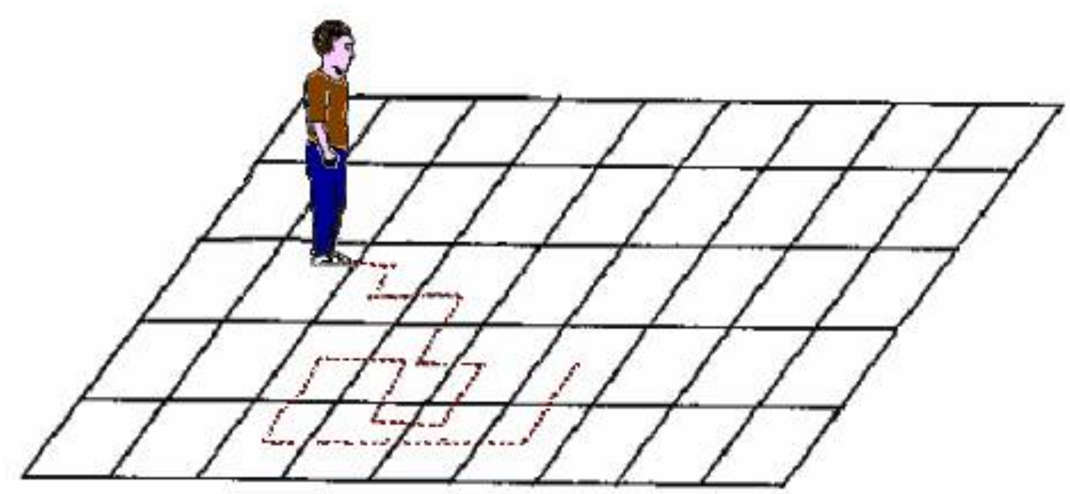
Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Sampling graphs

Random sampling (uniform & independent)	Crawling
‣ **vertex sampling**  ‣ **edge sampling** 	‣ **BFS sampling**  ‣ **random walk sampling** 

Random Walks on Graphs



Random walk sampling

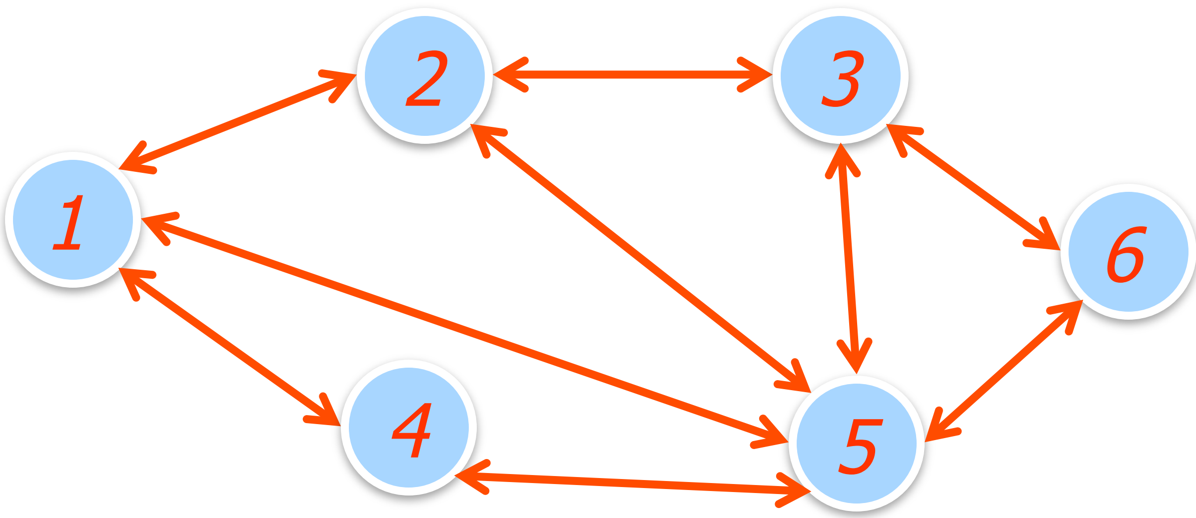
Random Walk
Routing

Molecule in liquid

Influence diffusion

Undirected Graphs

Undirected !!

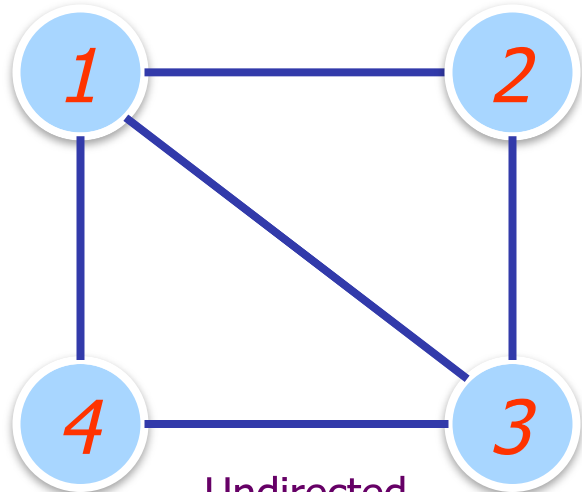


Random Walk

- Adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad D = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

Symmetric



Undirected

- Transition Probability Matrix

$$P_{ij} = \frac{1}{k_i}$$

$$P = A \bullet D^{-1} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}$$

- |E|: *number of links*
- *Stationary Distribution*

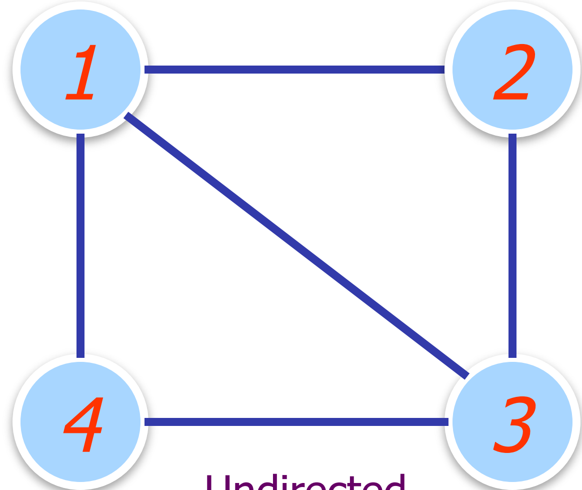
$$\pi_i = \frac{d_i}{2|E|}$$

Metropolis-Hastings Random Walk

- Adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad D = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

Symmetric



Undirected

- Transition Probability Matrix

$$P^{MH}_{\upsilon,w} = \begin{cases} \frac{1}{k_\upsilon}\min(1, \frac{k_\upsilon}{k_w}) & \text{if } w \text{ neighbor of } \upsilon \\ 1 - \sum_{y \neq \upsilon} P^{MH}_{\upsilon,y} & \text{if } w = \upsilon \end{cases}$$

$$P = A \bullet D^{-1} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \end{pmatrix}$$

- |E|: *number of links*
- *Stationary Distribution*

$$\pi_\upsilon = \frac{1}{|V|}$$

Walking in Facebook: A Case Study of Unbiased Sampling of OSNs

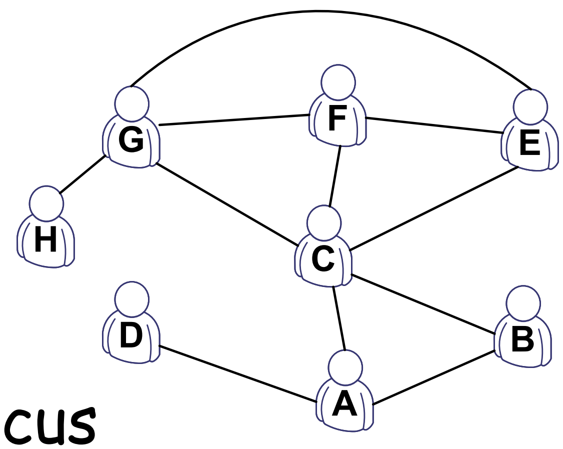
Minas Gjoka, Maciej Kurant ‡,
Carter Butts, Athina Markopoulou
UC Irvine, EPFL ‡

Outline

- Motivation and Problem Statement
- Sampling Methodology
- Data Analysis
- Conclusion

Online Social Networks (OSNs)

- A network of declared friendships between users
- Allows users to maintain relationships
- Many popular OSNs with different focus
  - Facebook, LinkedIn, Flickr, …



Social Graph

Why Sample OSNs?

- Representative samples desirable
  - study properties
  - test algorithms

- Obtaining complete dataset difficult
  - companies usually unwilling to share data
  - tremendous overhead to measure all (~100TB for Facebook)

Problem statement

- Obtain a **representative sample of users** in a given OSN by exploration of the social graph.
  - in this work we sample Facebook (FB)
  - explore graph using various crawling techniques

Related Work

- Graph traversal (BFS)
  - A. Mislove et al, IMC 2007
  - Y. Ahn et al, WWW 2007
  - C. Wilson, Eurosys 2009
- Random walks (MHRW, RDS)
  - M. Henzinger et al, WWW 2000
  - D. Stutbach et al, IMC 2006
  - A. Rasti et al, Mini Infocom 2009

Outline

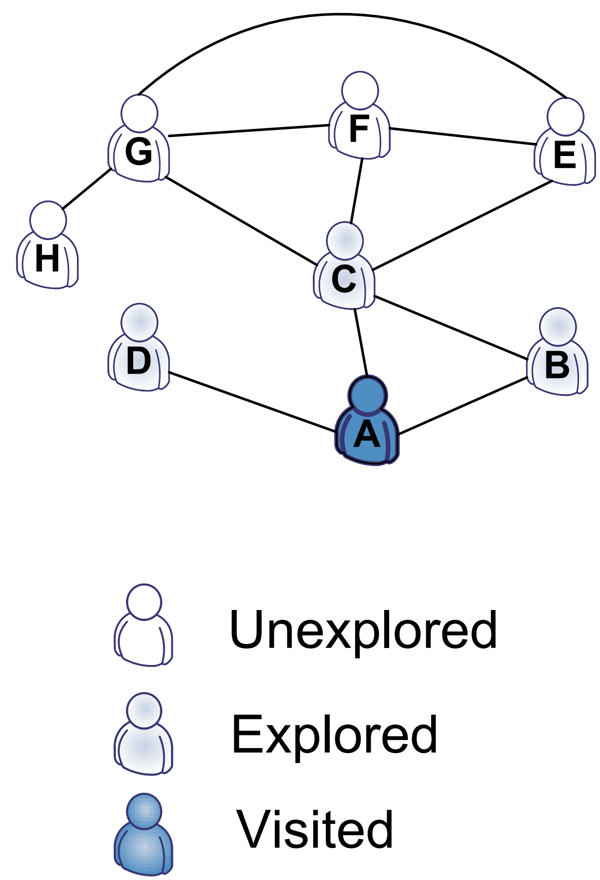
- Motivation and Problem Statement
- Sampling Methodology
  - crawling methods
  - data collection
  - convergence evaluation
  - method comparisons
- Data Analysis
- Conclusion

(1) Breadth-First-Search (BFS)

- Starting from a seed, explores all neighbor nodes. Process continues iteratively without replacement.

- BFS leads to bias towards high degree nodes
  Lee et al, "Statistical properties of Sampled Networks", Phys Review E, 2006

- Early measurement studies of OSNs use BFS as primary sampling technique
  i.e [Mislove et al], [Ahn et al], [Wilson et al.]



(2) Random Walk (RW)

- Explores graph one node at a time with replacement

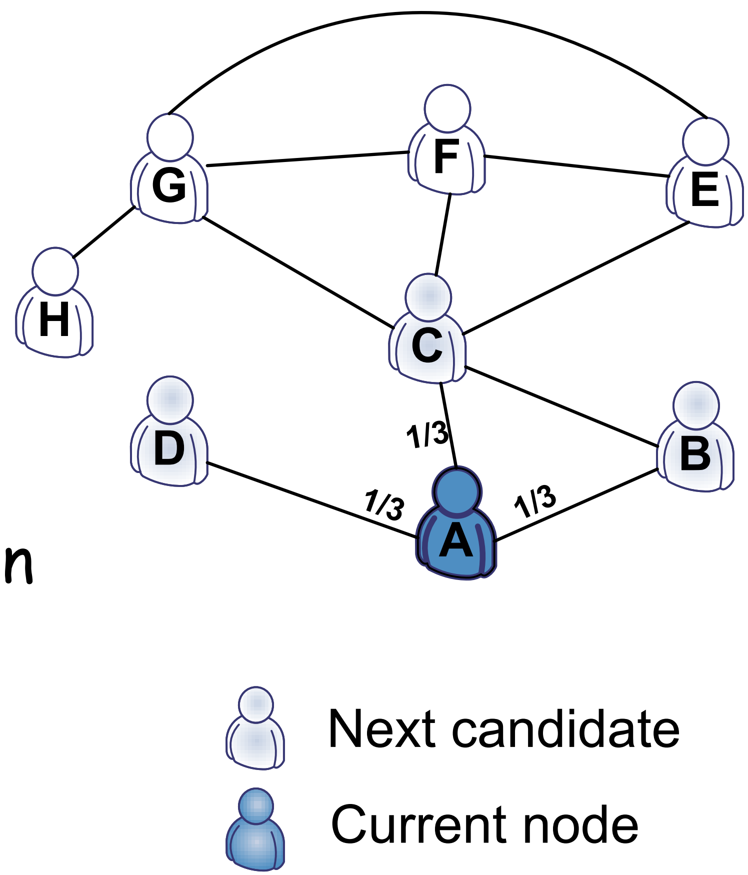
$$P^{RW}_{v,w} = \frac{1}{k_v}$$

Degree of node **u**

- In the stationary distribution

$$\pi_v = \frac{k_v}{2 \cdot |E|}$$

Number of edges



Minas Gjoka, UC Irvine

(3) Re-Weighted Random Walk (RWRW)

Hansen-Hurwitz estimator

- Corrects for degree bias at the end of collection
- Without re-weighting, the probability distribution for node property A is:

$$p(A_i) = \frac{\sum_{u \in A_i} 1}{\sum_{u \in V} 1} = \frac{|A_i|}{|V|}$$

Subset of sampled nodes with value i

All sampled nodes

- Re-Weighted probability distribution :

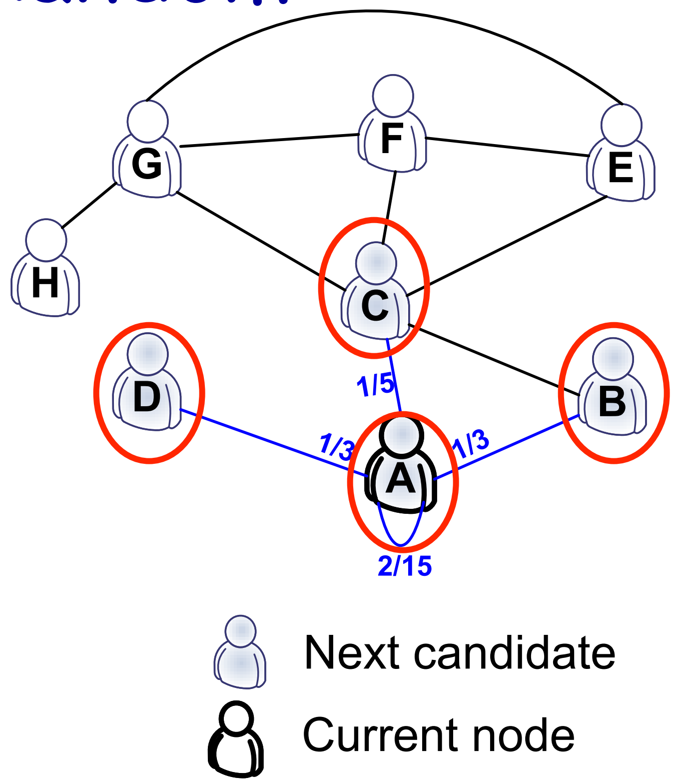
$$p(A_i) = \frac{\sum_{u \in A_i} 1/k_u}{\sum_{u \in V} 1/k_u}$$

Degree of node u

(4) Metropolis-Hastings Random Walk (MHRW)

- Explore graph one node at a time with replacement

$$P^{MH}_{\upsilon,w} = \begin{cases} \frac{1}{k_\upsilon}\min(1, \frac{k_\upsilon}{k_w}) & \text{if } w \text{ neighbor of } \upsilon \\ 1 - \sum_{y \neq \upsilon} P^{MH}_{\upsilon,y} & \text{if } w = \upsilon \end{cases}$$



- In the stationary distribution

$$\pi_\upsilon = \frac{1}{|V|}$$

$$P^{MH}_{AA} = 1 - (\frac{1}{3} + \frac{1}{3} + \frac{1}{5}) = \frac{2}{15}$$

$$P^{MH}_{AC} = \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{5}$$

Uniform userID Sampling (UNI)

- As a ***basis for comparison***, we collect a uniform sample of Facebook userIDs (UNI)
  - rejection sampling on the 32-bit userID space

- UNI not a general solution for sampling OSNs
  - userID space must not be sparse
  - names instead of numbers

Summary of Datasets

Sampling method	MHRW	RW	BFS	UNI
#Valid Users	28x81K	28x81K	28x81K	984K
# Unique Users	957K	2.19M	2.20M	984K

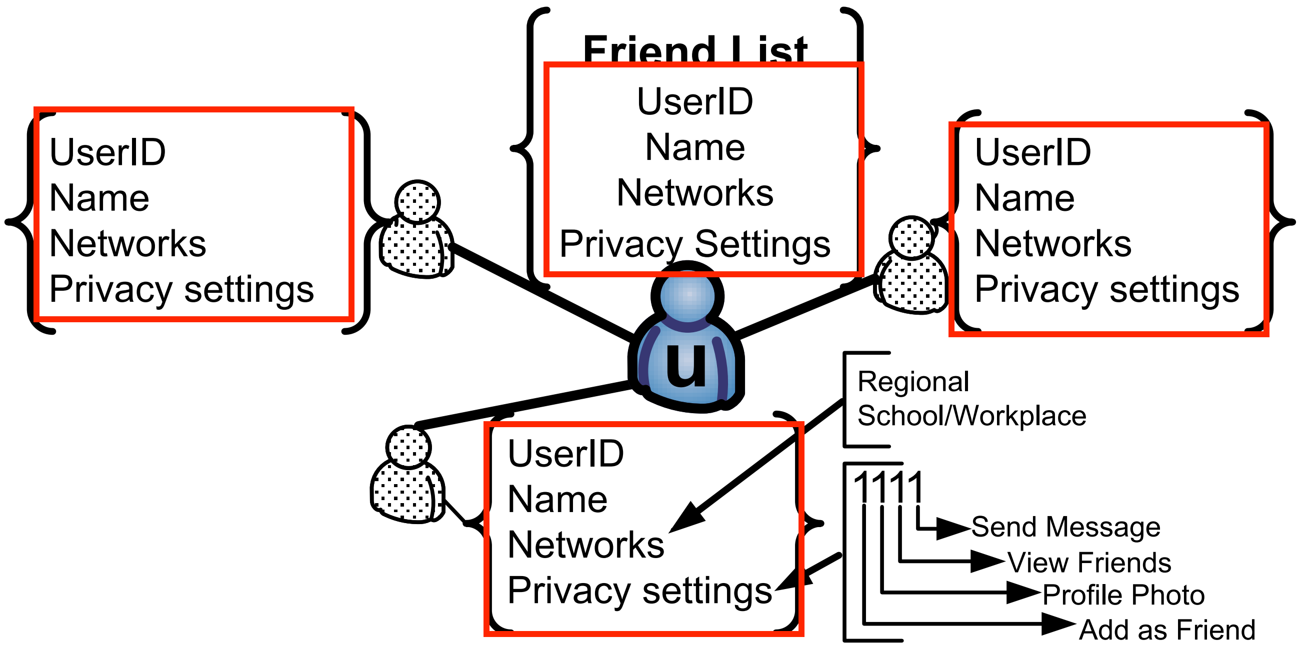
- Egonets for a subsample of MHRW
  - local properties of nodes

- Datasets available at:
  http://odysseas.calit2.uci.edu/research/osn.html

Data Collection

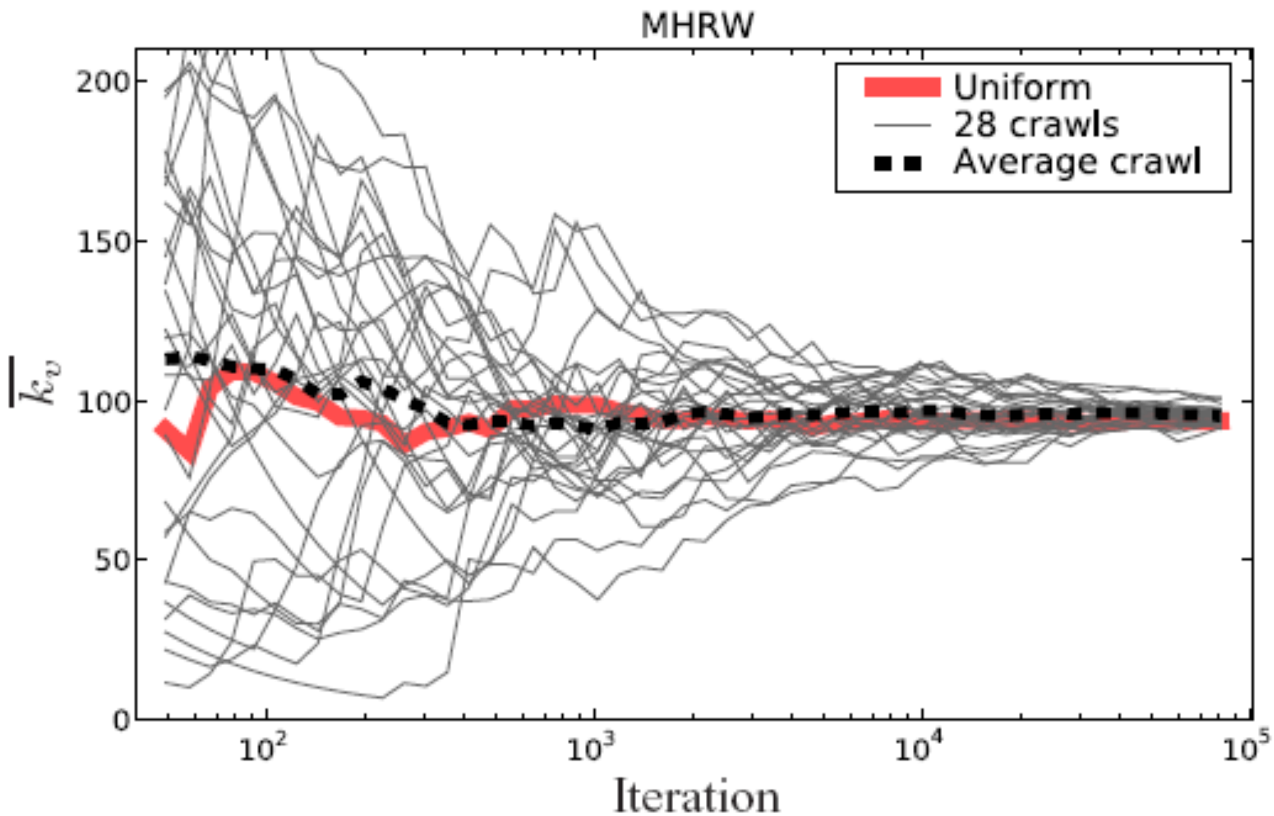
Basic Node Information

- What information do we collect for each sampled node **u**?



Detecting Convergence

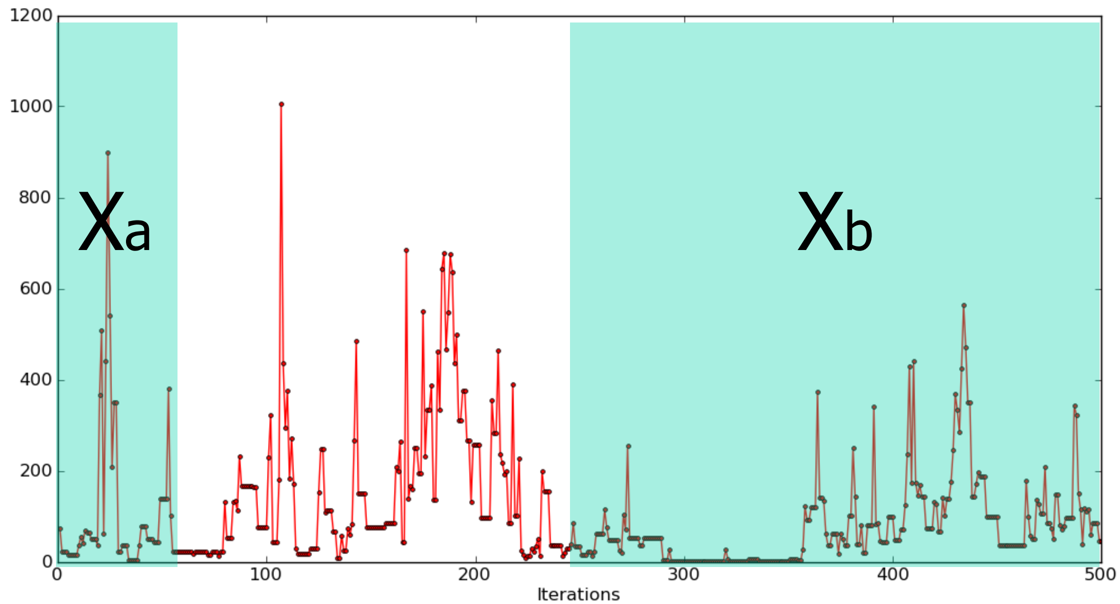
- Number of samples (iterations) to loose dependence from starting points?



Online Convergence Diagnostics

Geweke

- Detects convergence for a single walk. Let X be a sequence of samples for metric of interest.



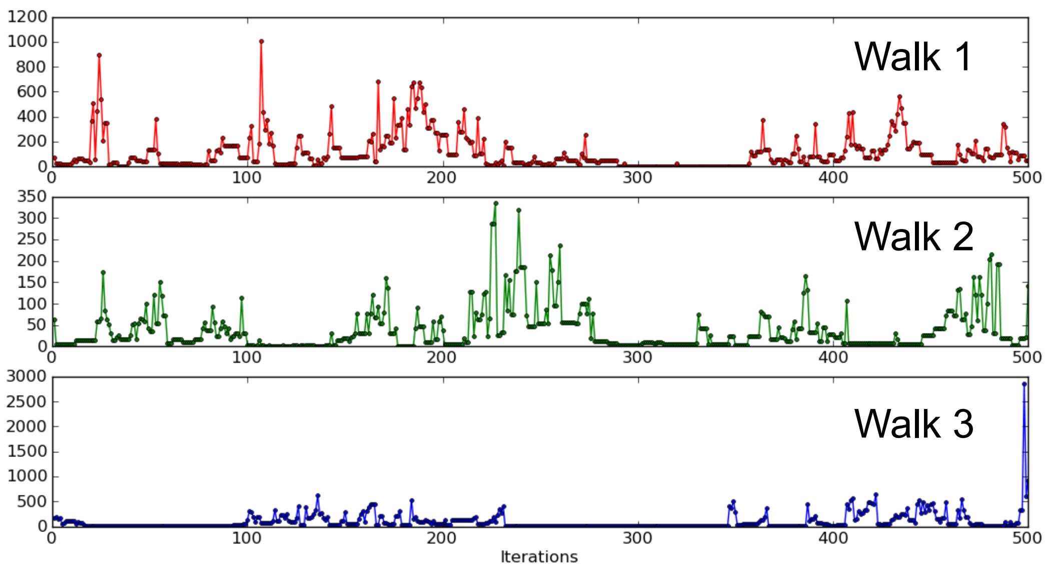
$$z = \frac{E(X_a) - E(X_b)}{\sqrt{Var(X_a) - Var(X_b)}}$$

J. Geweke, "Evaluating the accuracy of sampling based approaches to calculate posterior moments" in Bayesian Statistics 4, 1992

Online Convergence Diagnostics

Gelman-Rubin

- Detects convergence for m>1 walks



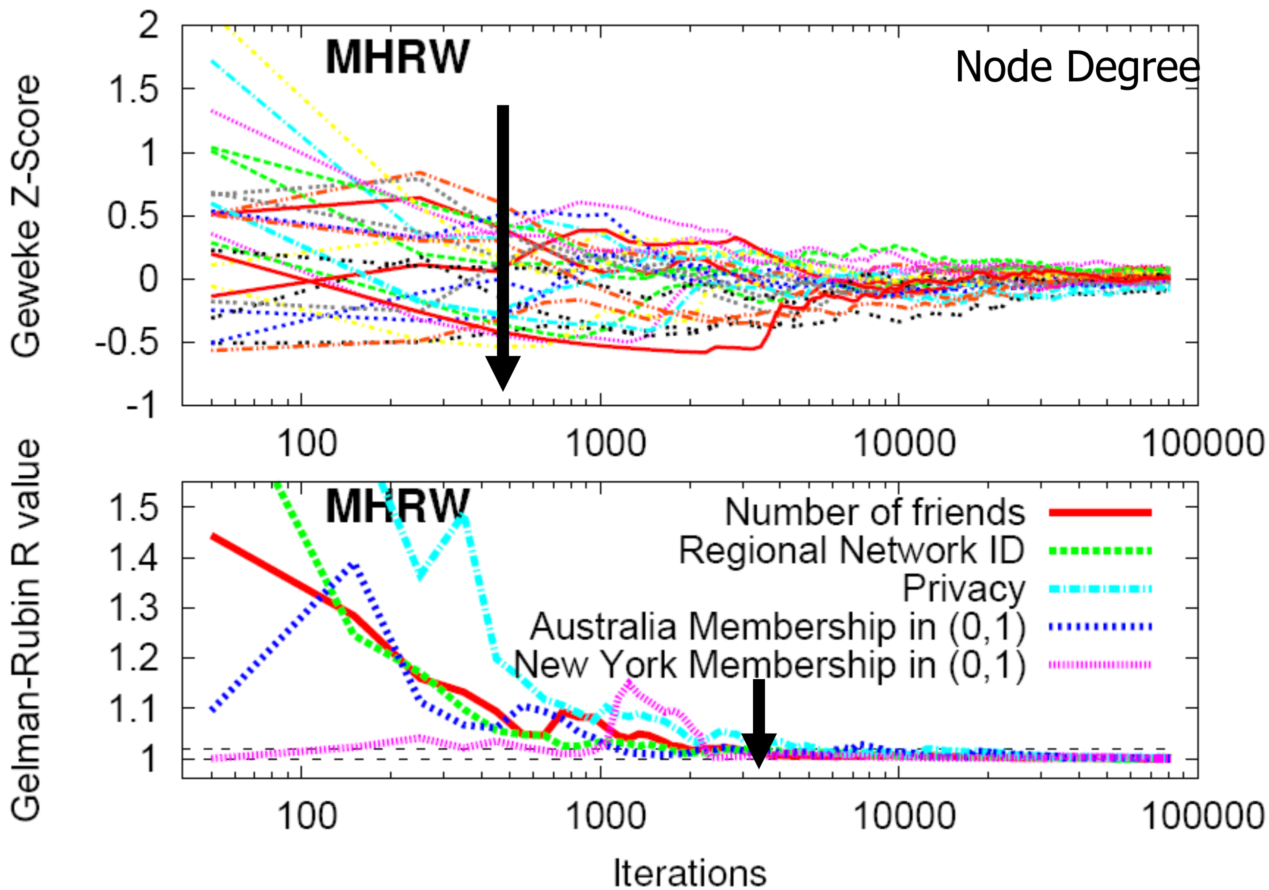
Between walks variance

$$\sqrt{R} = \sqrt{\left(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}\right)}$$

Within walks variance

A. Gelman, D. Rubin, "Inference from iterative simulation using multiple sequences" in Statistical Science Volume 7, 1992

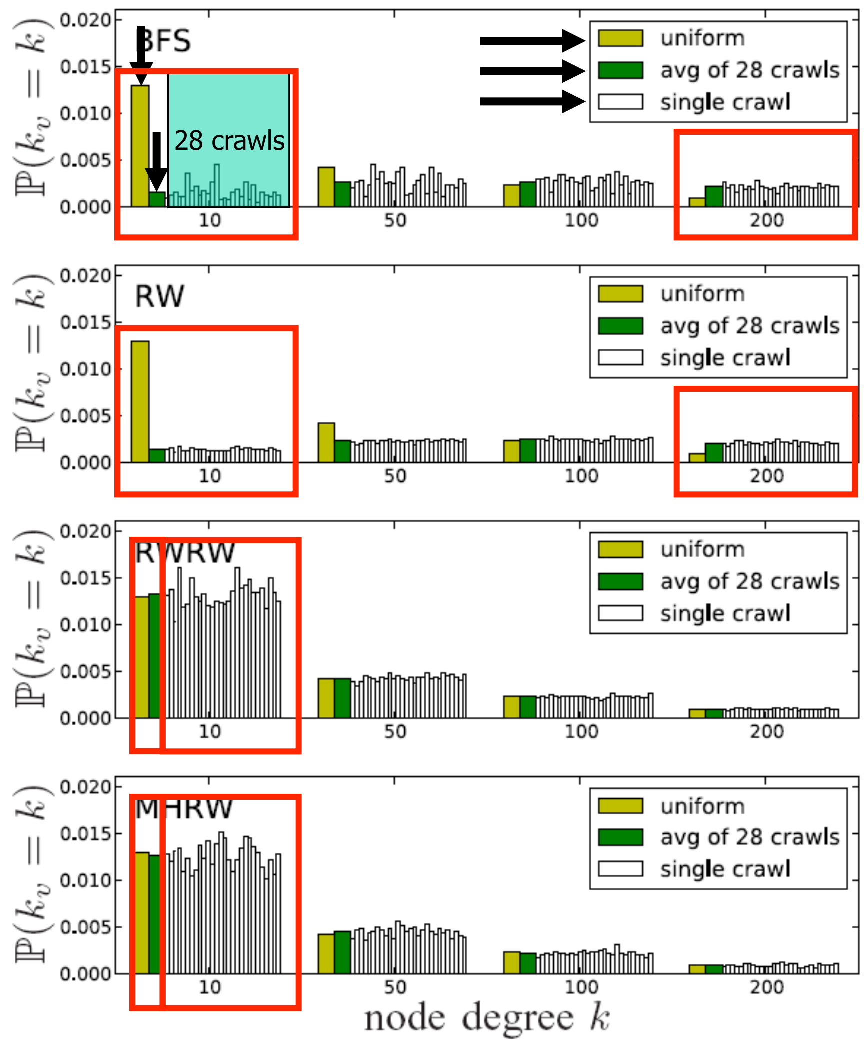
When do we reach equilibrium?



Burn-in determined to be 3K

Methods Comparison
Node Degree

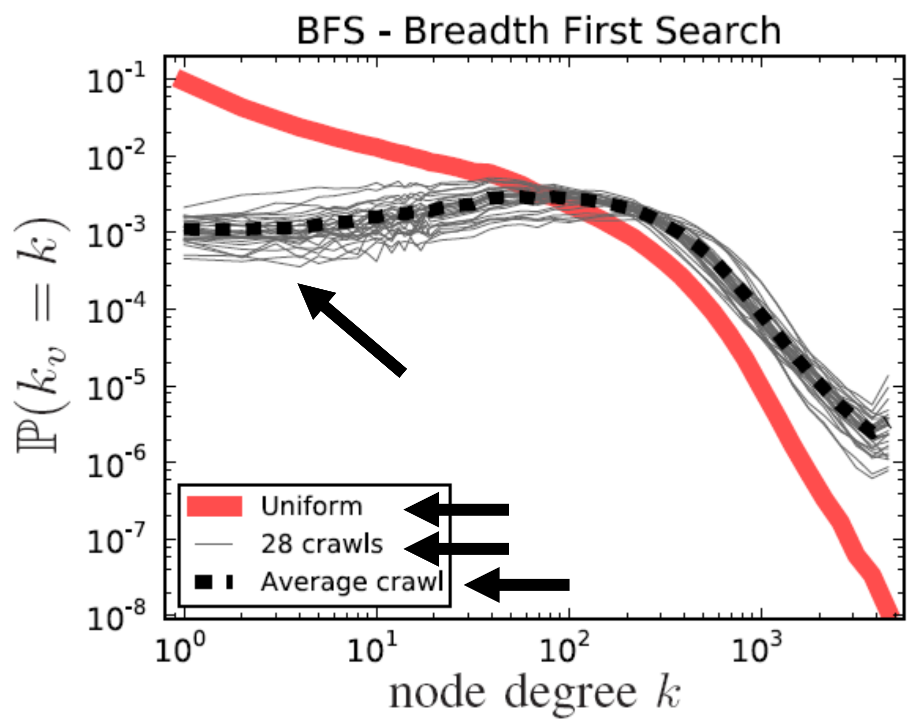
- Poor performance for BFS, RW
- MHRW, RWRW produce good estimates
  - per chain
  - overall



Sampling Bias
BFS

- Low degree nodes under-represented by two orders of magnitude

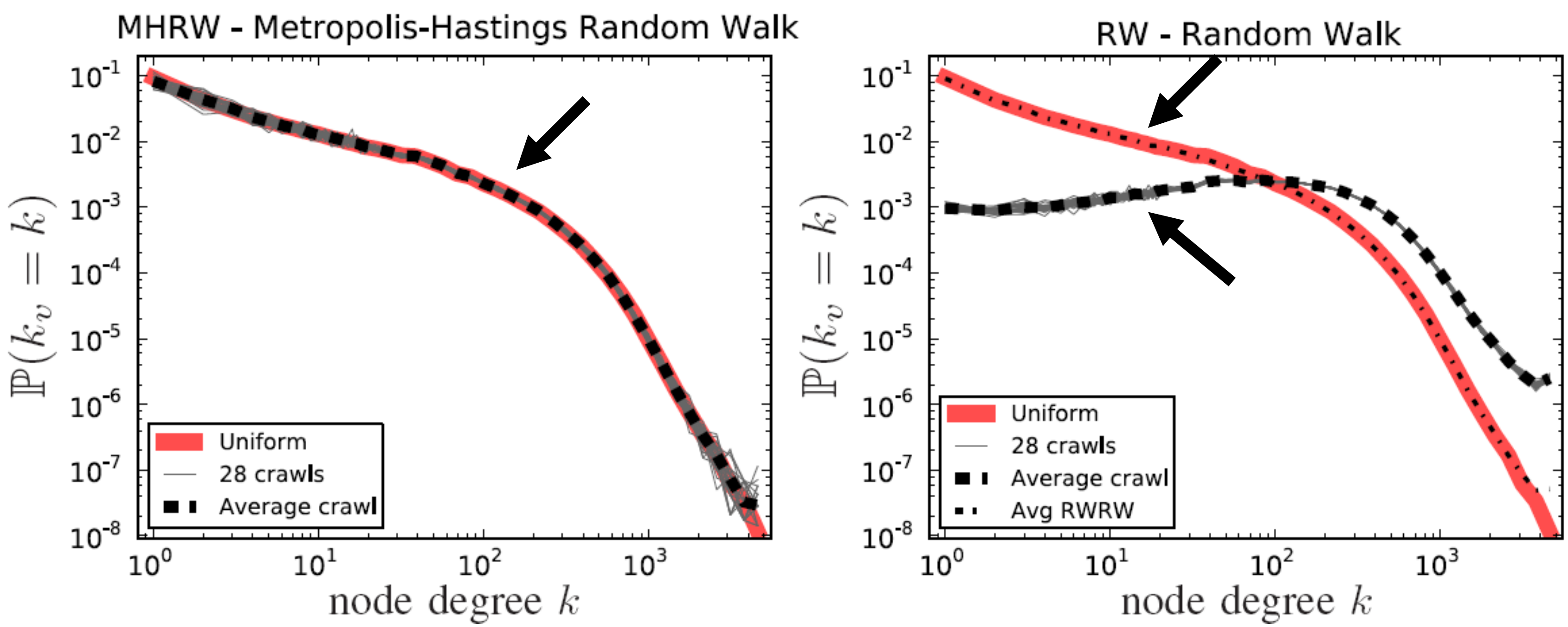
- BFS is biased



Minas Gjoka, UC Irvine

Sampling Bias

MHRW, RW, RWRW



- Degree distribution identical to UNI (MHRW,RWRW)
- RW as biased as BFS but with smaller variance in each walk

Practical Recommendations for Sampling Methods

- Use MHRW or RWRW. Do not use BFS, RW.

- Use formal convergence diagnostics
  - assess convergence online
  - use multiple parallel walks

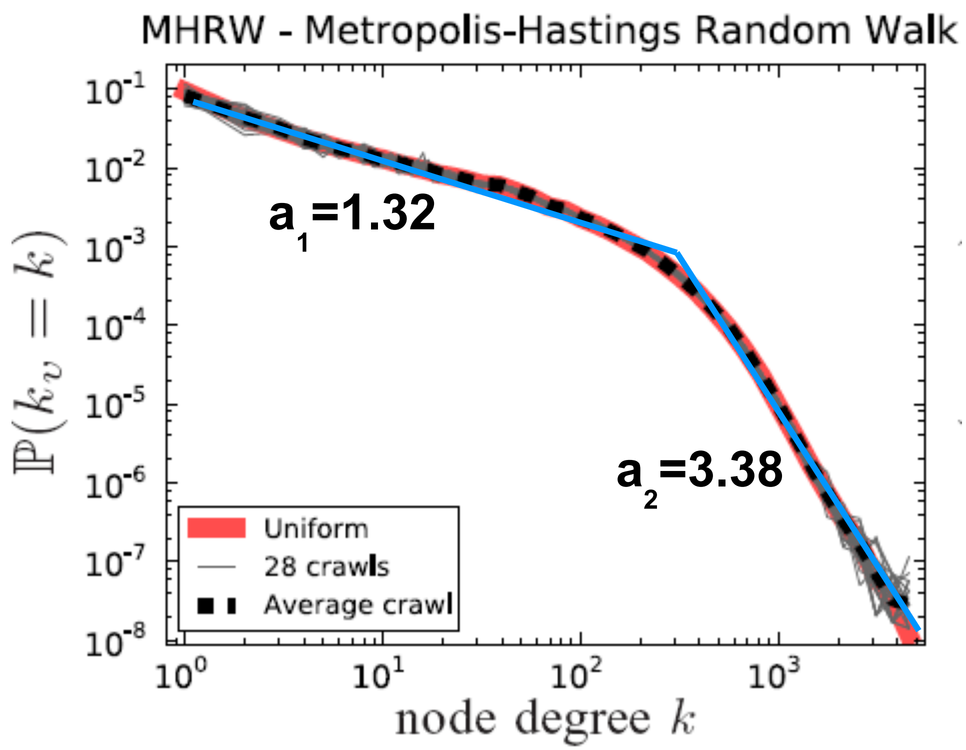
- MHRW vs RWRW
  - RWRW slightly better performance
  - MHRW provides a "ready-to-use" sample

Outline

- Motivation and Problem Statement
- Sampling Methodology
- Data Analysis
- Conclusion

FB Social Graph

Degree Distribution,
Power Law Distribution $f(k)=b*k^{-a}$



- Degree distribution not a power law

Any Comments & Critiques?

Next Class: Graph Mining (II)

- Do assigned readings before class
- Submit reviews/critiques
- Attend in-class discussions