

Sampling Online Social Networks

Athina Markopoulou^{1,3}



Joint work with:

Minas Gjoka³, Maciej Kurant³, Carter T. Butts^{2,3}, Patrick Thiran⁴







¹Department of Electrical Engineering and Computer Science

²Department of Sociology

³CalIT2: California Institute of Information Technologies
University of California, Irvine

⁴School of IC, EPFL, Lausanne

Online Social Networks (OSNs)

	500 million
	200 million
	130 million
	100 million
	75 million
	75 million

(November 2010)

> 1 billion users

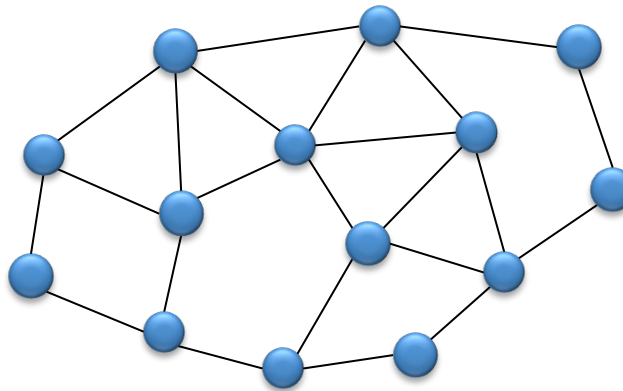
Activity: email and chat (FB), voice and video communication (e.g. skype), photos and videos (flickr, youtube), news, posting information, ...

Why study Online Social Networks?

Difference communities have different perspective

- Social Sciences
 - Fantastic source of data for studying online behavior
- Marketing
 - Influential users, recommendations/ads
- Engineering
 - OSN provider
 - Network/mobile provider
 - New apps/Third party services
- Large scale data mining
 - understand user communication patterns, community structure
 - "human sensors"
- Privacy
-

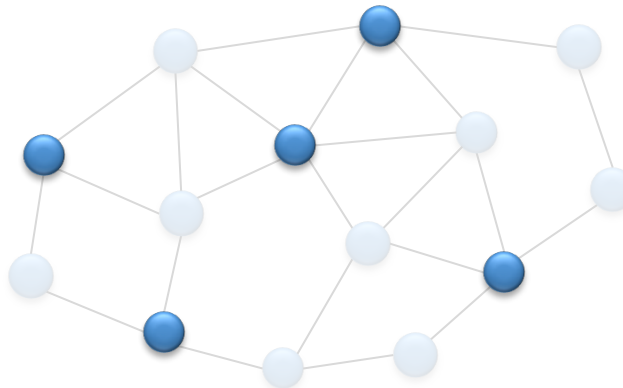
Original Graph



Interested in some property.

Graphs too large \rightarrow sampling

Sampling Nodes

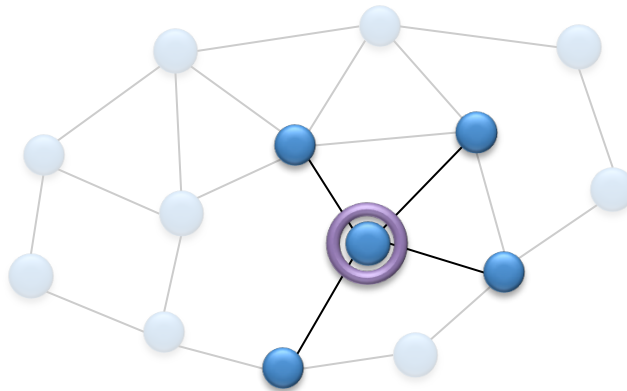


Estimate the property of interest from a sample of nodes

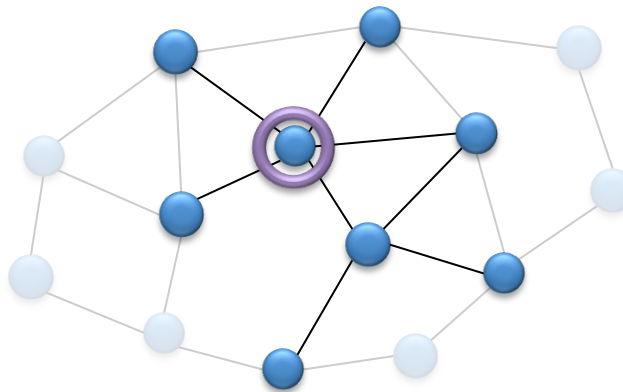
Population Sampling

- Classic problem
 - given a population of interest, draw a sample such that the probability of including any given individual is known.
- Challenge in online networks
 - often lack of a sampling frame: population cannot be enumerated
 - sampling of users: may be impossible (not supported by API, user IDs not publicly available) or inefficient (rate limited , sparse user ID space).
- Alternative: network-based sampling methods
 - Exploit social ties to draw a probability sample from hidden population
 - Use [crawling](#) (a.k.a. "link-trace sampling") [to sample nodes](#)

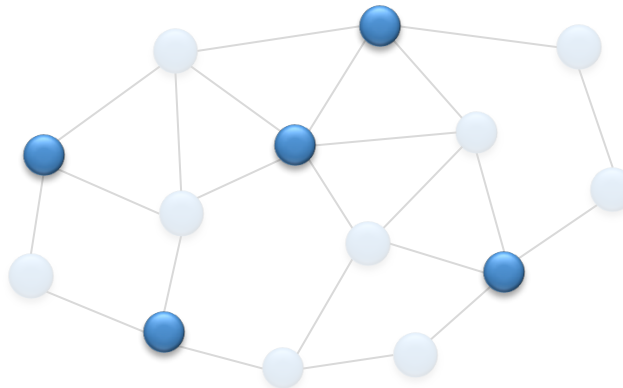
Sample Nodes by Crawling



Sample Nodes by Crawling



Sampling Nodes



Questions:

1. How do you collect a sample of nodes using crawling?
2. What can we estimate from a sample of nodes?

Related Work

- **Measurement/Characterization studies of OSNs**
 - Cyworld, Orkut, Myspace, Flickr, Youtube [...]
 - Facebook [Wilson et al. '09, Krishnamurthy et al. '08]
- **System aspects of OSNs:**
 - Design for performance, reliability [SPAR by Pujol et al, '10]
 - Design for privacy Privacy [PERSONA: Baden et al. '09]
- **Sampling techniques for WWW, P2P, recently OSNs**
 - BFS/traversal
[Mislove et al. 07, Cha 07, Ahn et al. 07, Wilson et al. 09, Ye et al. 10, Leskovec et al. 06, Viswanath 09]
 - Random walks on the web/p2p/osn
[Henzinger et al. '00, Gkantsidis 04, Leskovec et al. '06, Rasti et al. '09, Krishnamurthy'08] ...
 - Possibly time-varying graphs ...
[Stutzbach et al., Willinger et al. 09, Leskovec et al. '05]
 - Community detection ...
- **Survey Sampling**
 - Stratified Sampling [Neyman '34]
 - Adaptive cluster sampling [Thompson '90]
 -
- **MCMC literature**
 -
 - Fastest mixing Markov Chain [Boyd et al. '04]
 - Frontier-Sampling [Ribeiro et al. '10]

Outline

- Introduction
- Sampling Techniques
 - Random Walks/BFS for sampling Facebook
 - Multigraph Sampling
 - Stratified Weighted Random Walk
- What can we learn from a sample?
- Conclusion and Future Directions

Outline

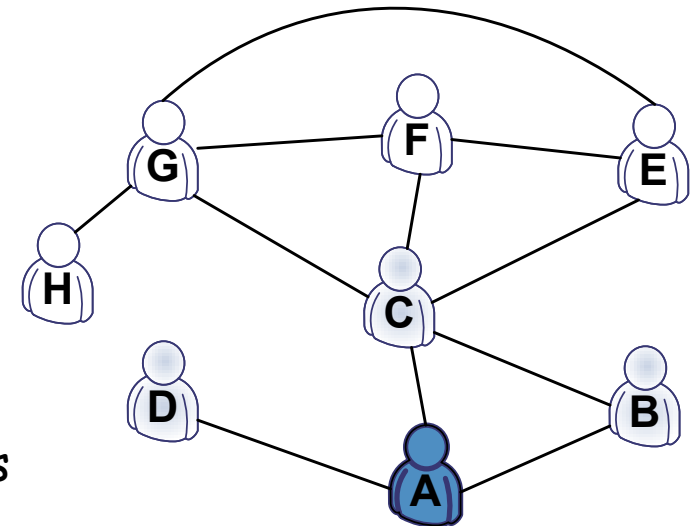
- Introduction
- Sampling Techniques
 - Random Walks/BFS for sampling Facebook
 - Multigraph Sampling
 - Stratified Weighted Random Walk
- What can we learn from a sample?
- Conclusion and Future Directions

How should we crawl Facebook?

- Before the crawl
 - Define the graph (users, relations to crawl)
 - Pick crawling method for lack of bias and efficiency
 - Decide what information to collect
 - Implement efficient crawlers, deal with access limitations
- During the crawl
 - When to stop? Online convergence diagnostics
- After the crawl
 - What samples to discard?
 - How to correct for the bias, if any?
 - How to evaluate success? ground truth?
 - What can we do with the collected sample (of nodes)?

Method 1: Breadth-First-Search (BFS)

- Starting from a seed, explores all neighbors nodes. Process continues iteratively
- Sampling 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.]



Method 2: Simple Random Walk (RW)

- Randomly choose a neighbor to visit next
- (sampling with replacement)

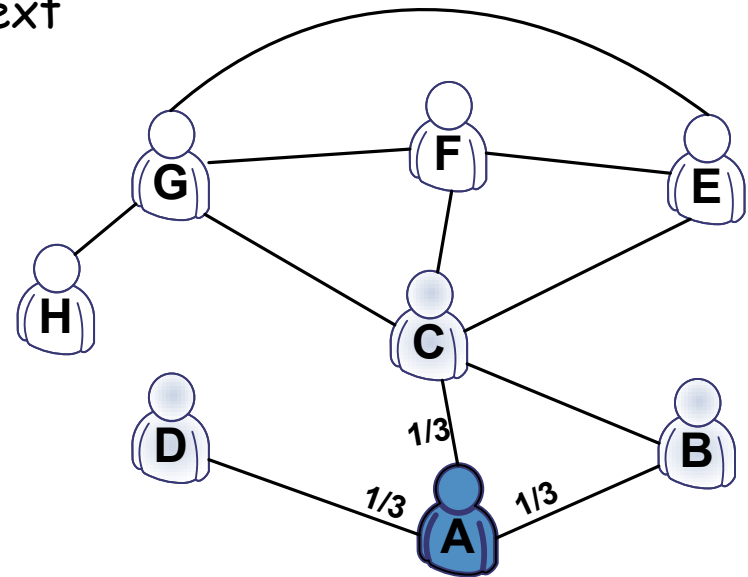
$$P_{v,w}^{RW} = \frac{1}{k_v}$$

Degree of node **u**

- leads to stationary distribution

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

- RW is biased towards high degree nodes



Next candidate

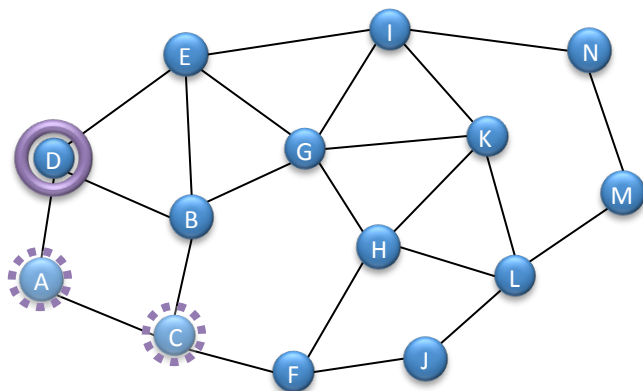


Current node

Correcting for the bias of the walk

Method 3:

Metropolis-Hastings Random Walk (MHRW):



DAAC ...



$\deg(A) < \deg(D) \Rightarrow$ go with probability 1



$\deg(C) > \deg(A) \Rightarrow$ go with probability $\frac{\deg(A)}{\deg(C)}$



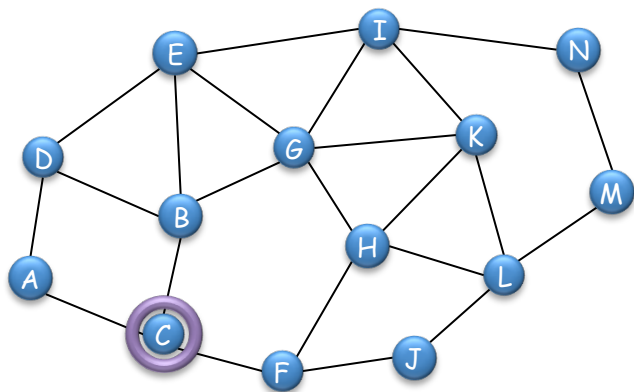
$\deg(C) > \deg(A) \Rightarrow$ go with probability $\frac{\deg(A)}{\deg(C)}$

$$\Pr(v \text{ is from Australia}) = \frac{\sum_{u \in S} 1_{\{v \text{ is from Australia}\}}}{\sum_{u \in S} 1}$$

Correcting for the bias of the walk

Method 3:

Metropolis-Hastings Random Walk (MHRW):



DAAC ...



$\deg(A) < \deg(D) \Rightarrow$ go with probability 1



$\deg(C) > \deg(A) \Rightarrow$ go with probability $\frac{\deg(A)}{\deg(C)}$

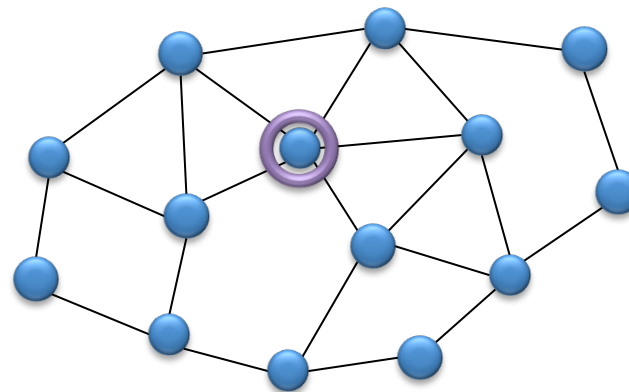


$\deg(C) > \deg(A) \Rightarrow$ go with probability $\frac{\deg(A)}{\deg(C)}$

$$\Pr(v \text{ is from Australia}) = \frac{\sum_{u \in S} 1_{\{v \text{ is from Australia}\}}}{\sum_{u \in S} 1}$$

Method 4:

Re-Weighted Random Walk (RWRW):

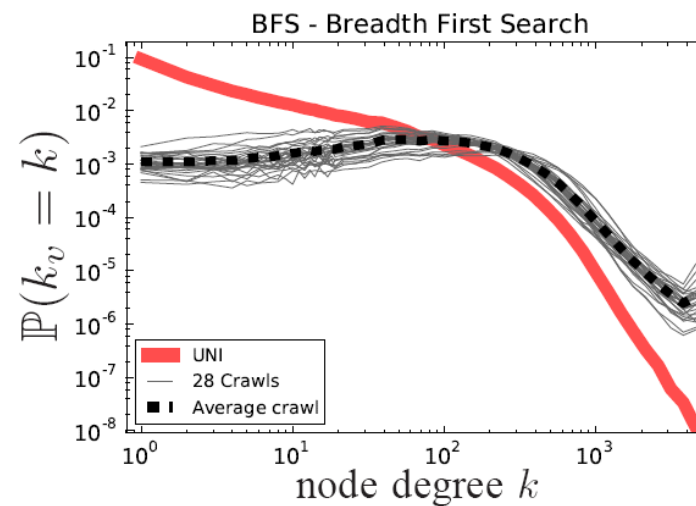
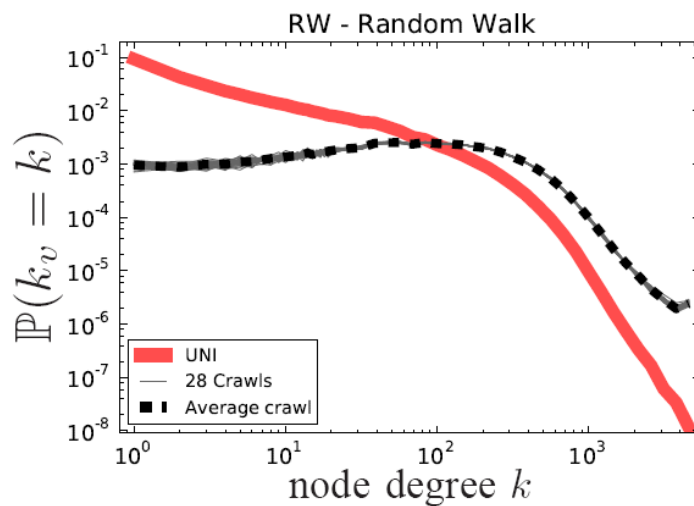
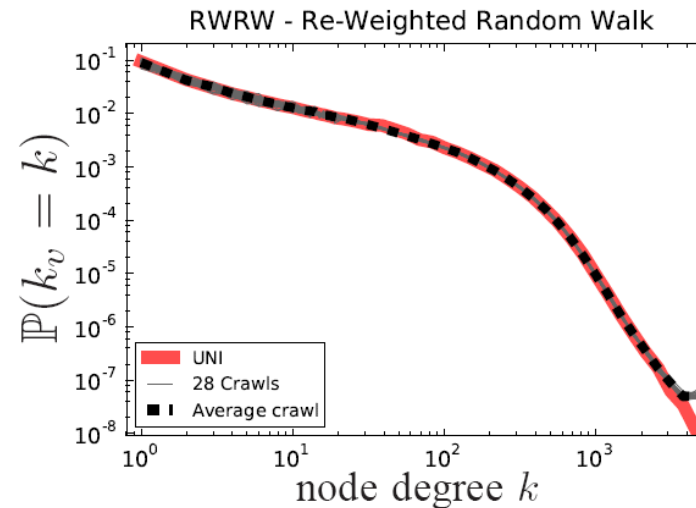
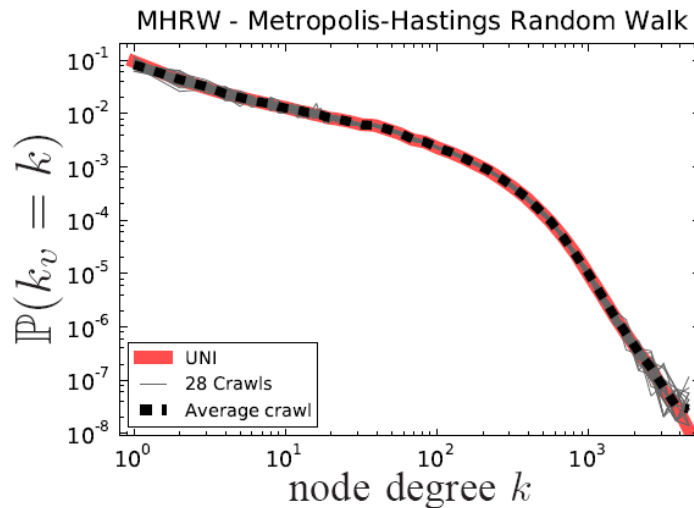


Now apply the Hansen-Hurwitz estimator:

$$\Pr(v \text{ is from Australia}) = \frac{\sum_{u \in S} 1_{\{u \text{ is from Australia}\}} / k_u}{\sum_{u \in S} 1 / k_u}$$

Comparison in terms of bias

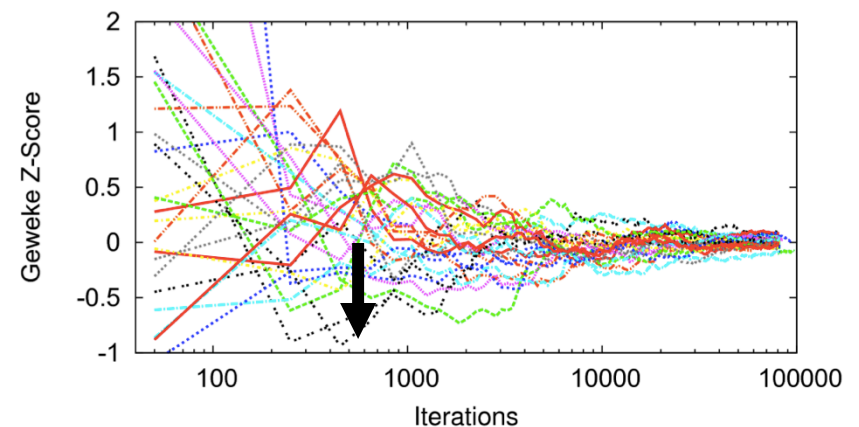
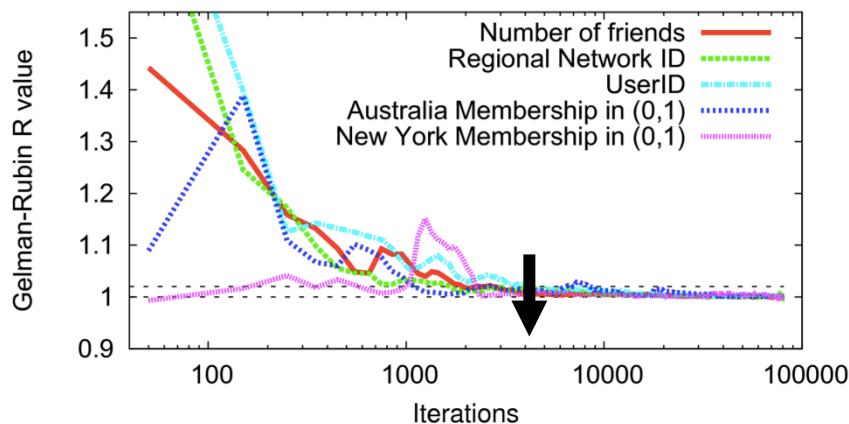
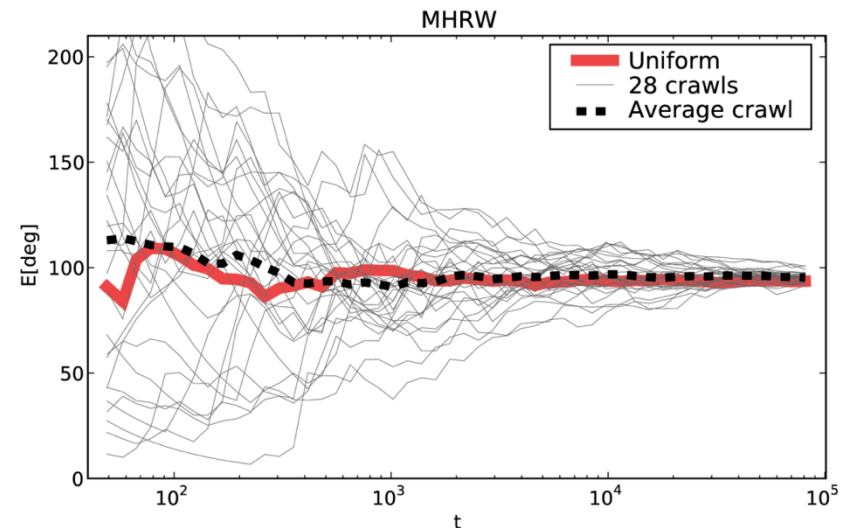
Node Degree in Facebook



Online Convergence Diagnostics

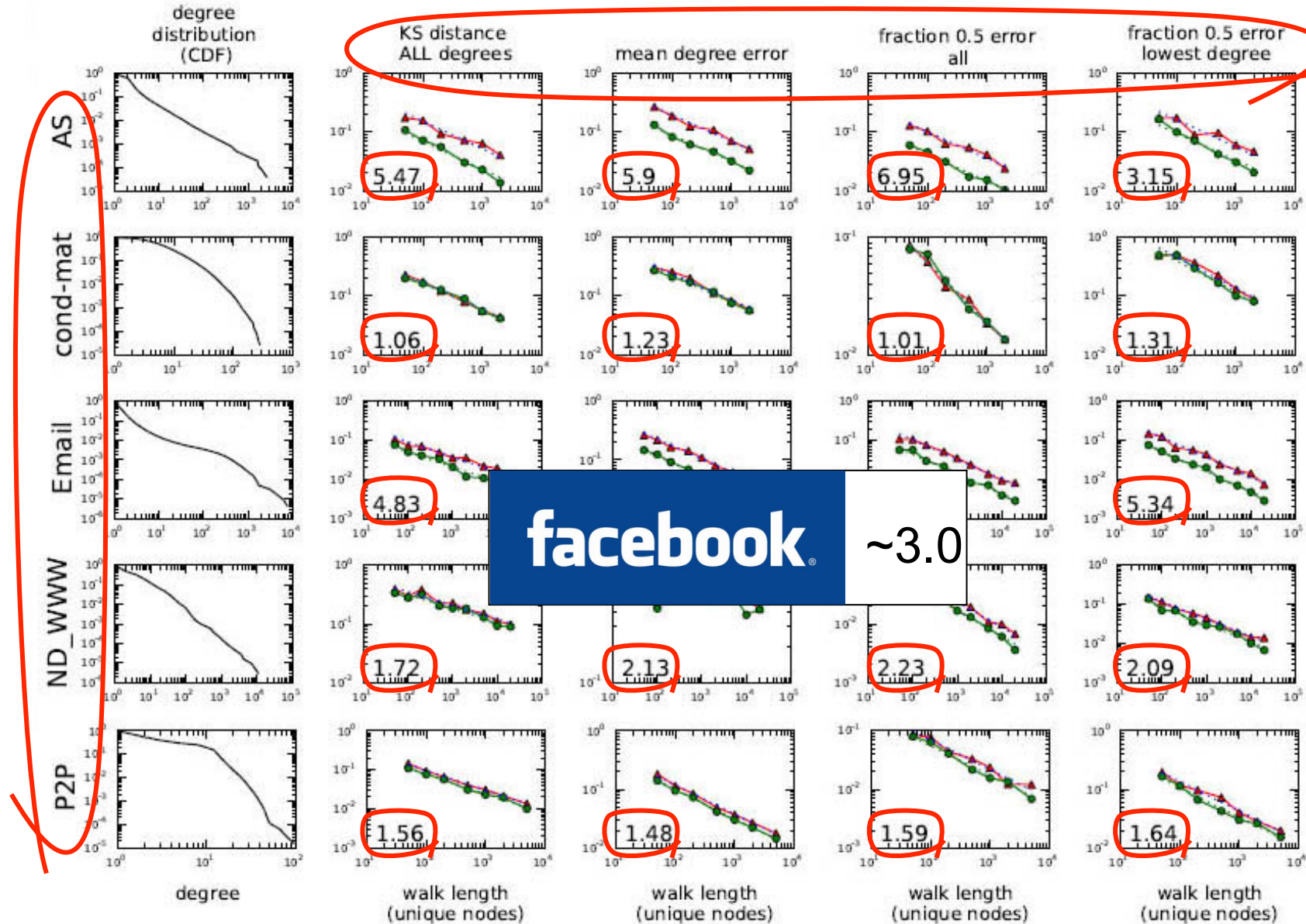
- Inferences assume that samples are drawn from stationary distribution
- No ground truth available in practice
- MCMC literature, online diagnostics

Acceptable convergence between 500 and 3000 iterations (depending on property of interest)



Comparison in Terms of Efficiency

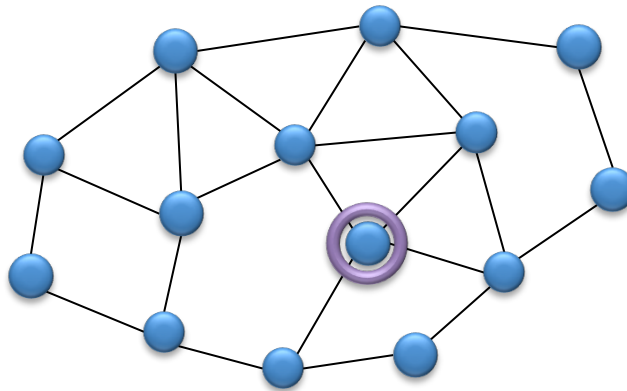
MHRW vs. RWRW



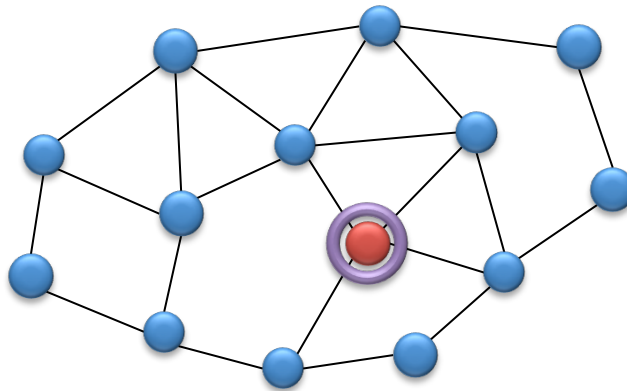
MHRW vs. RWRW

- Both do the job: they yield an unbiased sample
- RWRW converges faster than MHRW
 - for all practical purposes (1.5-8 times faster)
 - pathological counter-examples exist.
- MHRW easy/ready to use - does not require reweighting
- In the rest of our work, we consider only (RW)RW.
- How about BFS?

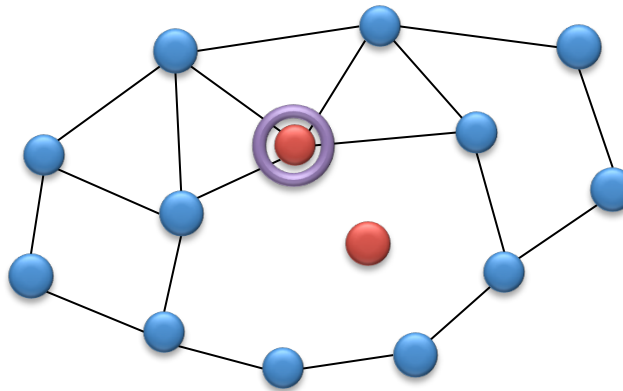
Sampling without replacement



Sampling without replacement



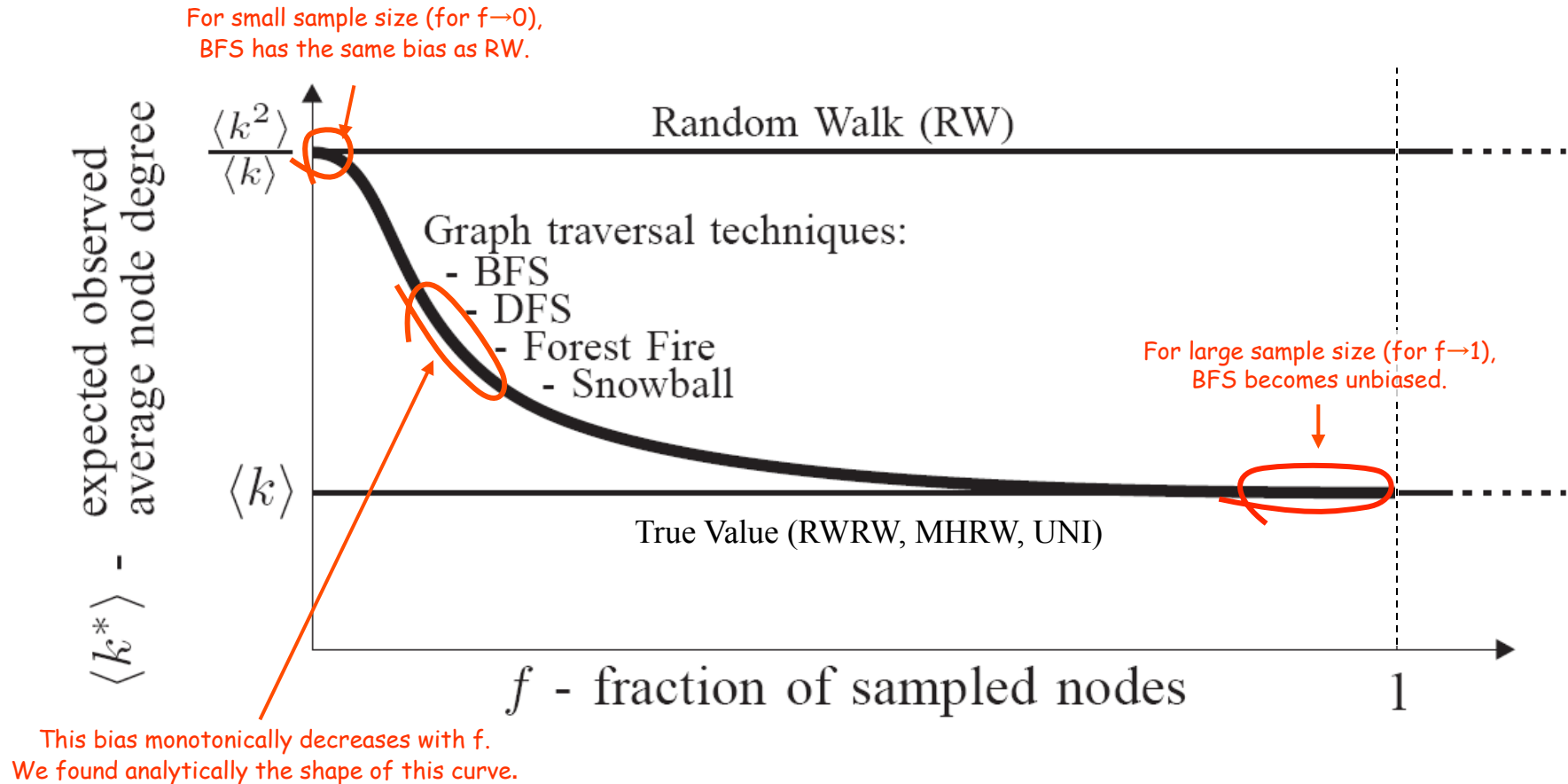
Sampling without replacement



Examples:

- BFS (Breadth-First Search)
- DFS (Depth-First Search)
- Forest Fire....
- RDS (Respondent-Driven Sampling)
- Snowball sampling

BFS degree bias



true: $p_k = \Pr\{\text{degree}=k\}$

biased: $q_k^{\text{BFS}}(f) = \frac{p_k(1 - (1-t(f))^k)}{\sum_l p_l(1 - (1-t(f))^l)}$

corrected: $\hat{p}_k^{\text{BFS}} = \frac{\hat{q}_k}{1 - (1-t(f))^k} \cdot \left(\sum_l \frac{\hat{q}_l}{1 - (1-t(f))^l} \right)^{-1}$

Correction exact for RG(pk)
Approximate for general graphs

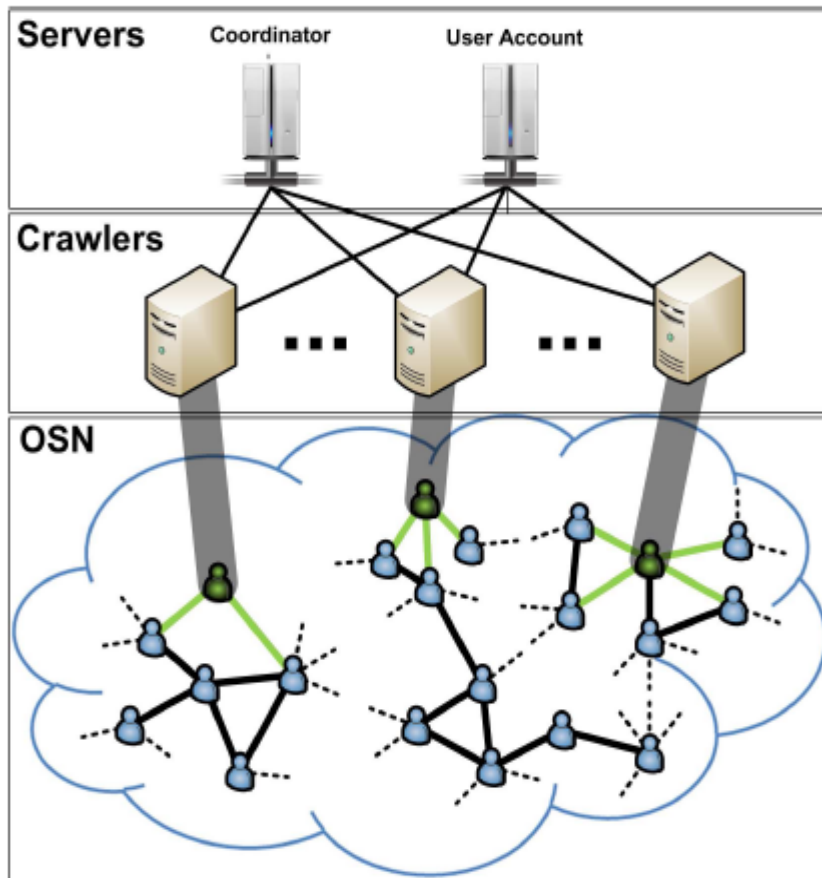
On the bias of BFS

- We computed analytically the bias of BFS in $RG(p_k)$
 - Same bias for all sampling w/o replacement, for $RG(p_k)$
 - Can correct for the bias of node attribute frequency
 - Given sample of nodes; $(v, x(v), \deg(v))$; BFS fraction f
 - Exact for $RG(p_k)$
 - Well enough (on avg, not in variance) in real-life topologies
 - In general, a difficult problem
-
- M. Kurant, A. Markopoulou, P. Thiran "Towards Unbiasing BFS Sampling", in *Proc. of ITC'22* and to appear *IEEE JSAC on Internet Topologies*
 - Python code available at: <http://mkurant.com/maciej/publications>

Data Collection Challenges

- Facebook is not easy to crawl
 - rich client side Javascript
 - interface changes often
 - stronger than usual privacy settings
 - limited data access when using API. Used HTML scraping.
 - unofficial rate limits that result in account bans
 - large scale
 - growing daily
- Designed and implemented efficient OSN crawlers.

Speeding Up Crawling



- Distributed implementation decreased time to crawl ~1million users **from ~2weeks to <2 days.**

Distributed data fetching

- cluster of 50 machines
- coordinated crawling

Parallelization

- Multiple machines
- Multiple processes per machine (crawlers)
- Multiple threads per process (parallel walks)

RW, MHRW, BFS

Datasets

1. Facebook users, April-May 2009

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

2. Last.FM multigraph, July 2010

3. Facebook social graph, October 2010

- ~2 days, 25 independent walks, 1M unique users, RW and Stratified RW

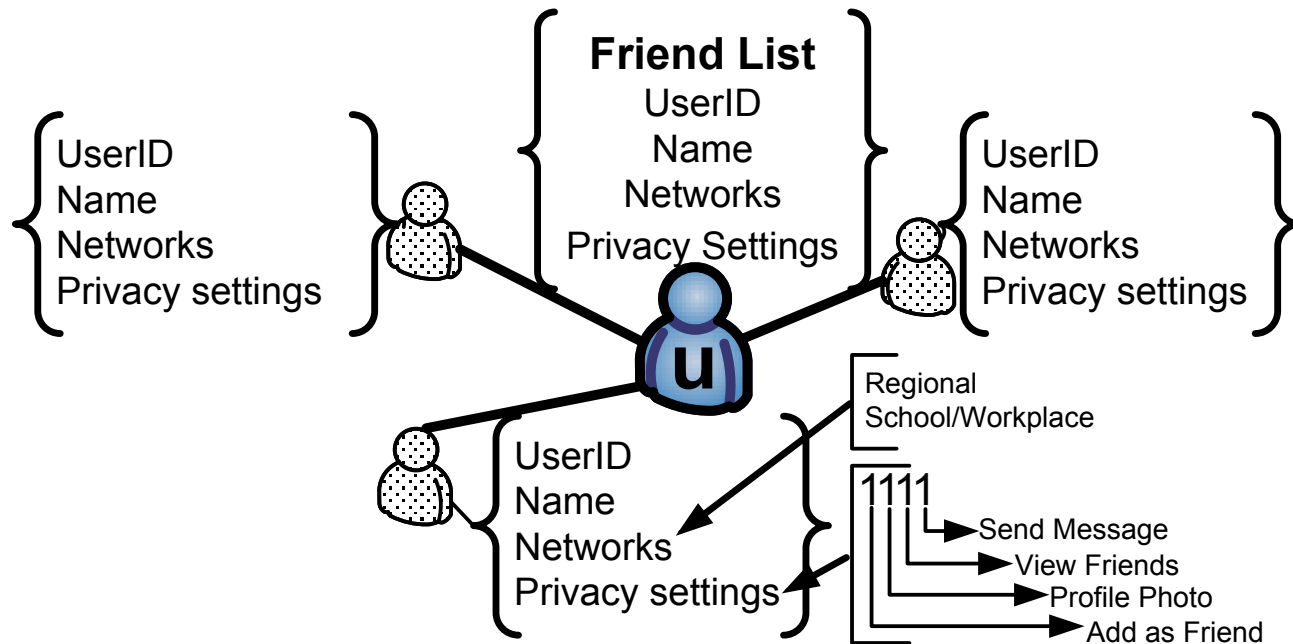
4. Category-to-category Facebook graphs

Publicly available at: <http://odysseas.calit2.uci.edu/research/osn.html>

Requested ~1000 times since April 2010

Information Collected

At each sampled node



- Also collected extended egonets for a subsample of MHRW
 - 37k egonets with ~6 million neighbors

Crawling Facebook

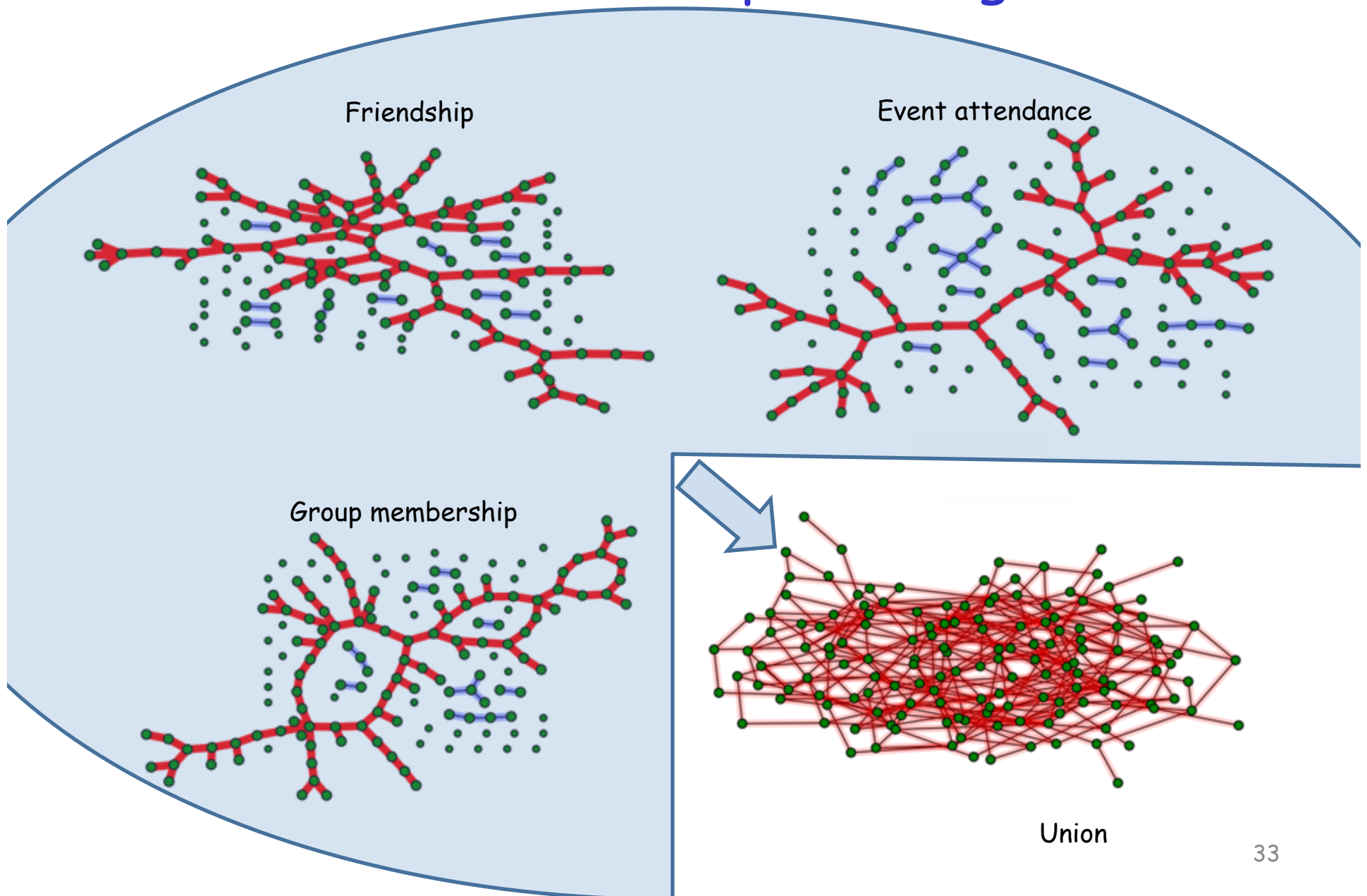
Summary

- Compared different methods
 - MHRW, RWRW performed remarkably well
 - BFS, RW lead to substantial bias - which we can correct for
 - RWRW (more efficient) vs. MHRW (ready to use)
 - Practical recommendations
 - use of online convergence diagnostics
 - proper use of multiple chains
 - implementation matters
 - Obtained and made publicly available uniform sample of Facebook:
 - <http://odysseas.calit2.uci.edu/research/osn.html>
-
- M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs", in *Proc. of IEEE INFOCOM '10*
 - M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, "Practical Recommendations for Sampling OSN Users by Crawling the Social Graph", to appear in *IEEE JSAC on Measurements of Internet Topologies 2011*.
 - M. Kurant, A. Markopoulou, P. Thiran, "Towards Unbiased BFS Sampling", *ITC'10, JSAC'11*

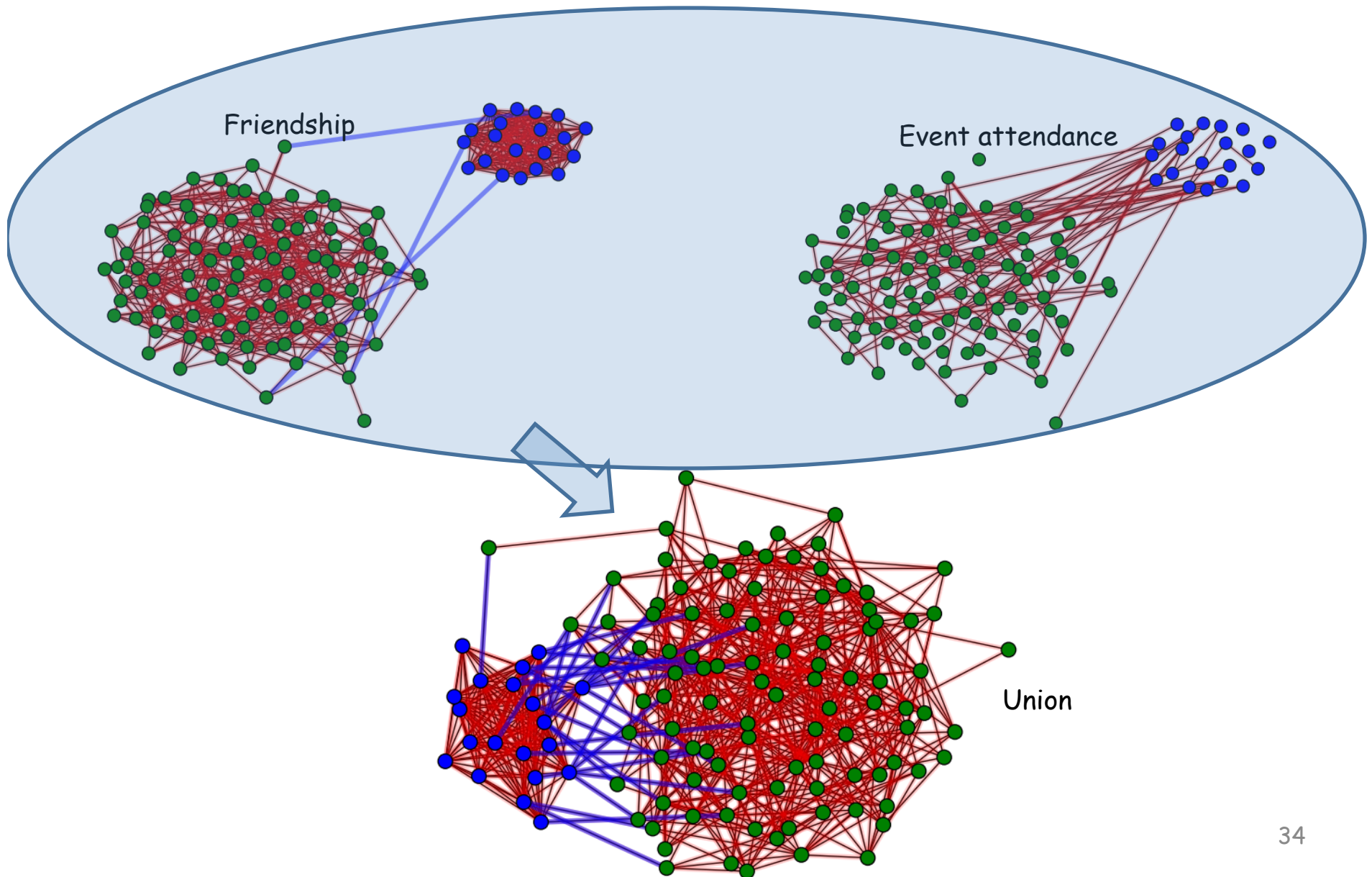
Outline

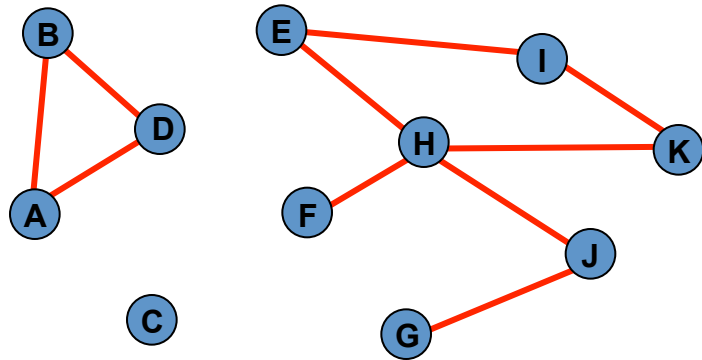
- Introduction
- Sampling Techniques
 - Random Walks/BFS for sampling Facebook
 - **Multigraph Sampling**
 - Stratified Weighted Random Walk
- What can we learn from a sample?
- Conclusion and Future Directions

What if the Social Graph is fragmented?

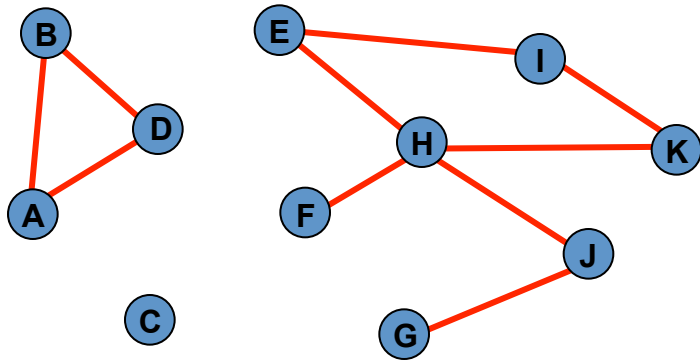


What if the social graph is highly clustered?

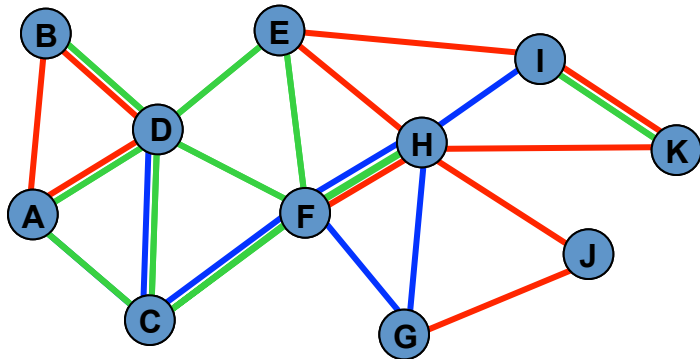




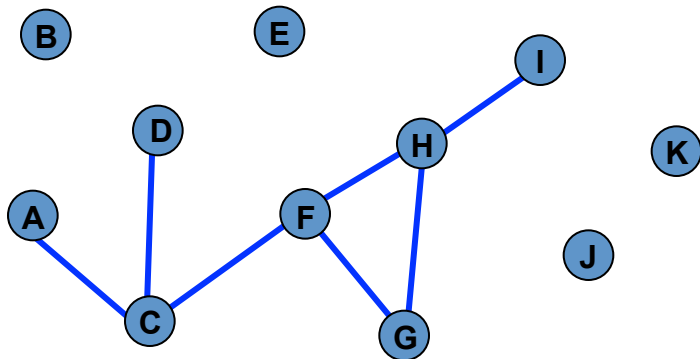
Friends



Friends

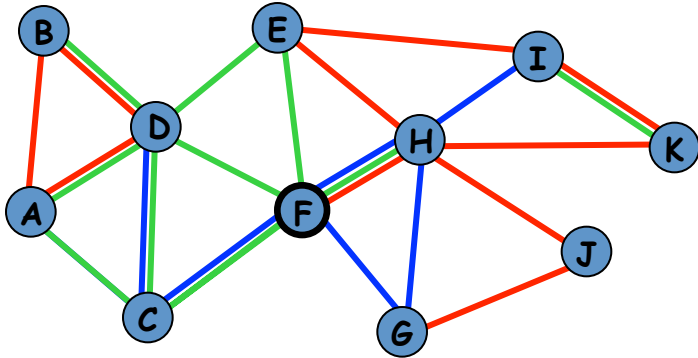


Events

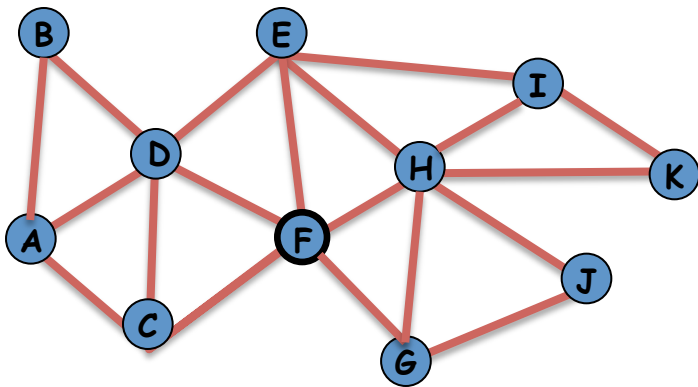


Groups

Combining multiple relations



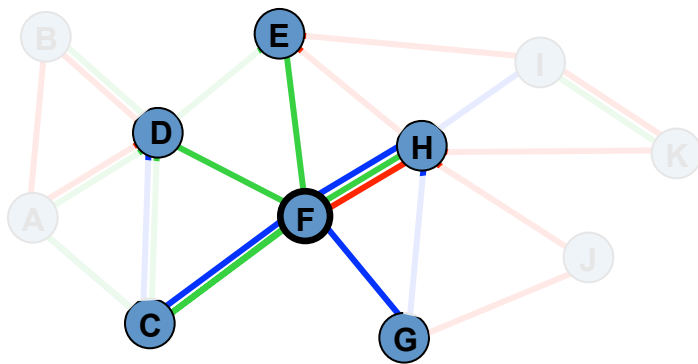
G_* is a union multigraph



G is a union graph

Multigraph Sampling

efficient implementation



Friends + Events + Groups
(G_* is a multigraph)

Approach 1:

- 1) Select edge to follow uniformly at random, *i.e.*, with probability $1 / \deg(F, G_*)$

Approach 2: **does not require listing neighbors from all relations**

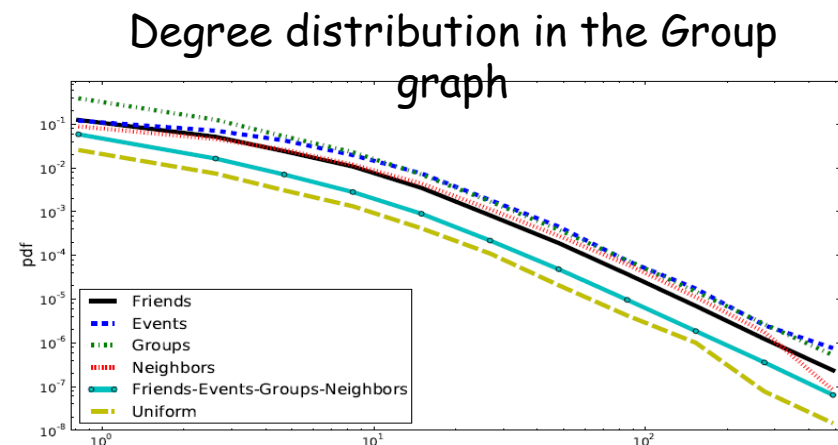
- 1) Select relation graph G_i with probability $\deg(F, G_i) / \deg(F, G_*)$
- 2) Within G_i choose an edge uniformly at random, *i.e.*, with probability $1/\deg(F, G_i)$.

Multigraph Sampling

Evaluation in Last.FM

- Last.FM
 - An Internet radio service with social networking features
 - fragmented graph components and highly clustered users
 - Isolated users (degree 0): 87% in Friendship graph, 94% in Group graph
 - **Solution: exploit multiple relations.**
 - Example: consider the groups graph

Relation to crawl	Isolates discovered
Friends	60.4 %
Events	41.7 %
Groups	0 %
Friend+Events+Groups	85.3 %
True	93.8 %



Multigraph Sampling

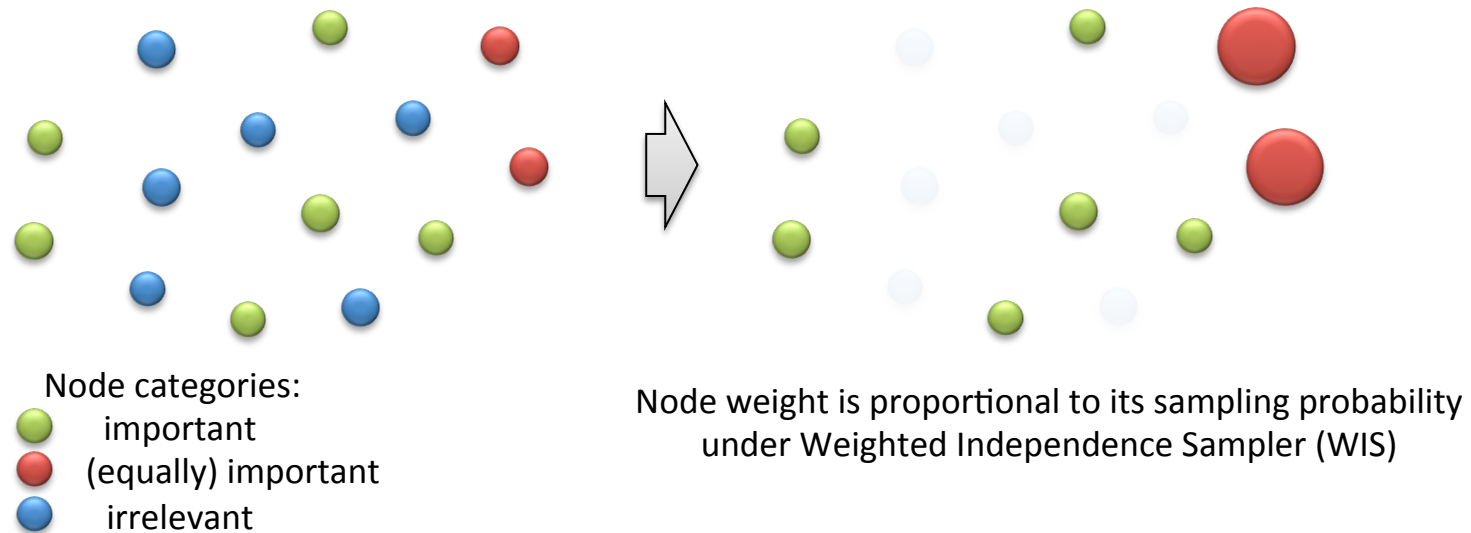
Summary

- simple concept, efficient implementation
 - walk on the union multigraph
 - applied to Last.FM:
 - discovers isolated nodes, better mixing
 - better estimates of distributions and means
 - open:
 - selection and weighting of relations
-
- M. Gjoka, C. T. Butts, M. Kurant, A. Markopoulou, "Multigraph Sampling of Online Social Networks", to appear in *IEEE JSAC on Measurements of Internet Topologies*.

Outline

- Introduction
- Sampling Techniques
 - Random Walks/BFS for sampling Facebook
 - Multigraph Sampling
 - **Stratified Weighted Random Walk**
- What can we learn from a sample?
- Conclusion and Future Directions

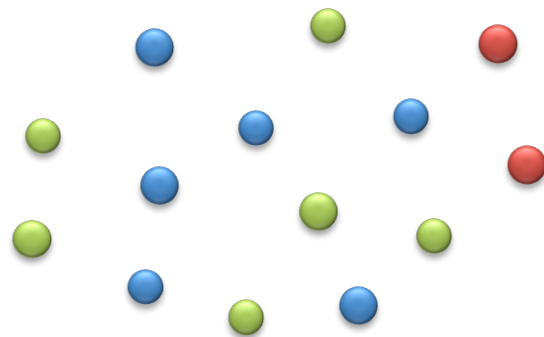
What if *not* all nodes are equally important?



Stratified Independence Sampling:

- Given a population partitioned in non-overlapping categories ("stratas"), a sampling budget and an estimation objective related to categories
- decide how many samples to assign to each category

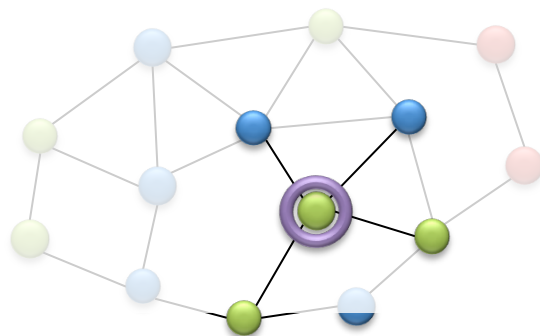
What if *not* all nodes are equally important?



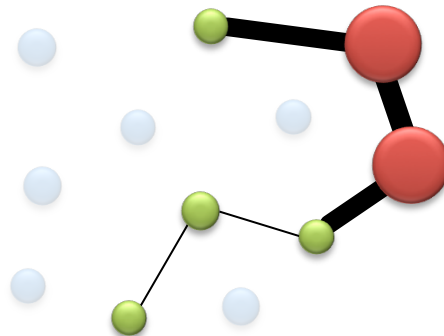
Node categories:
● important
● (equally) important
● irrelevant



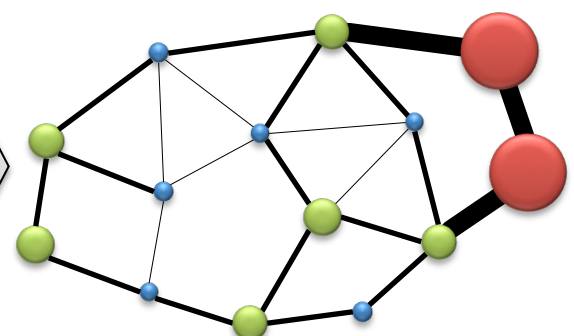
Node weight is proportional to its sampling probability under Weighted Independence Sampler (WIS)



But we sample through crawling!

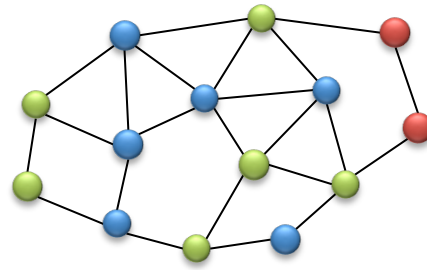


Enforcing WIS weights may lead to slow (or no) convergence



We have to trade off between fast convergence and ideal (WIS) node sampling probabilities

Measurement objective

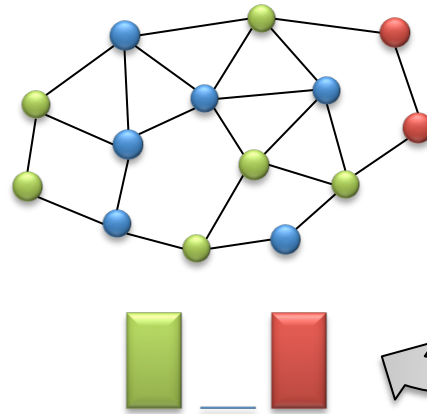


E.g., compare the size of red and green categories.

Measurement objective

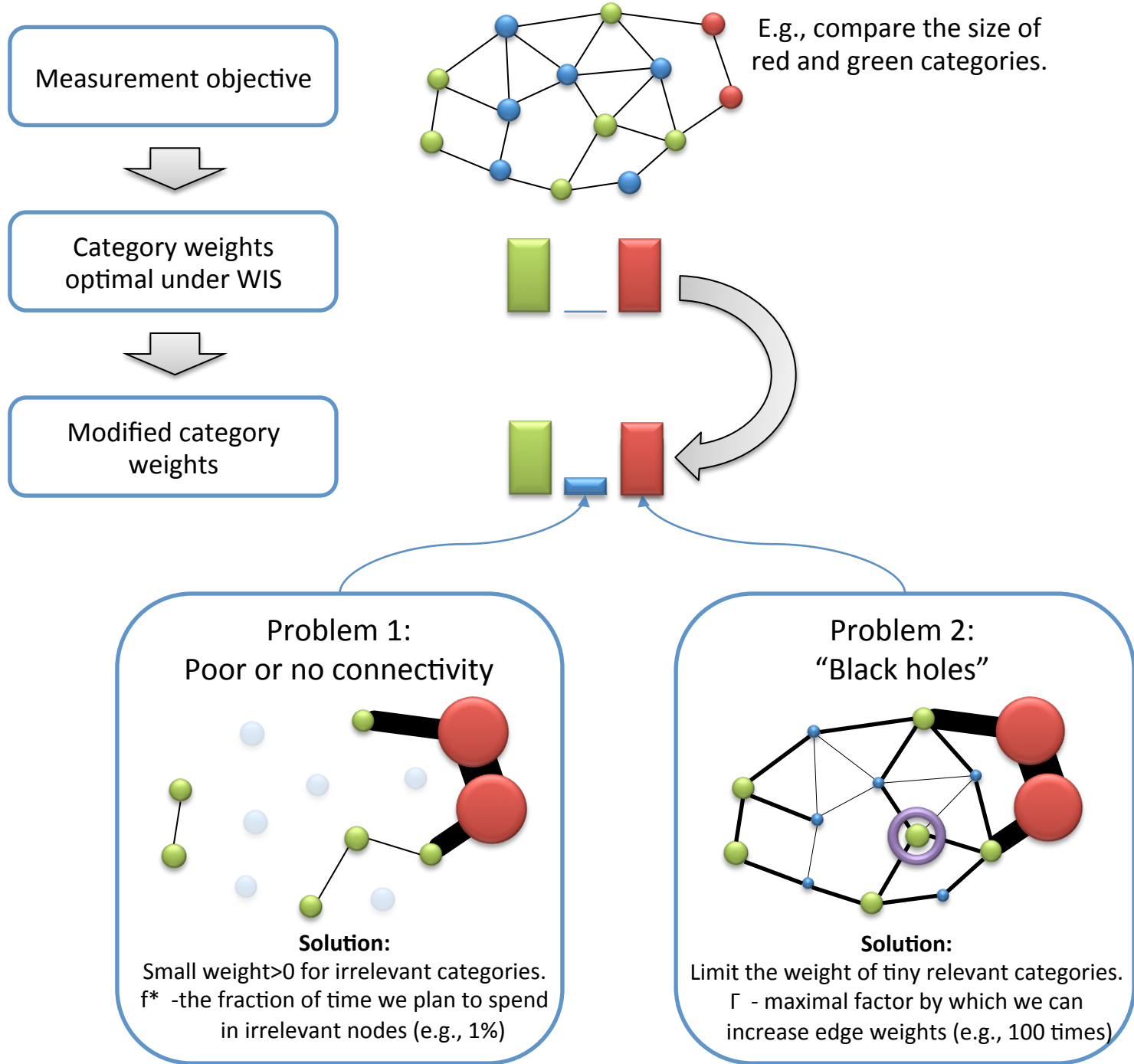


Category weights
optimal under WIS



E.g., compare the size of
red and green categories.

Warm-up crawl:
• category relative volumes



Measurement objective



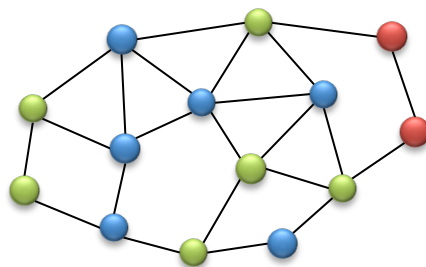
Category weights
optimal under WIS



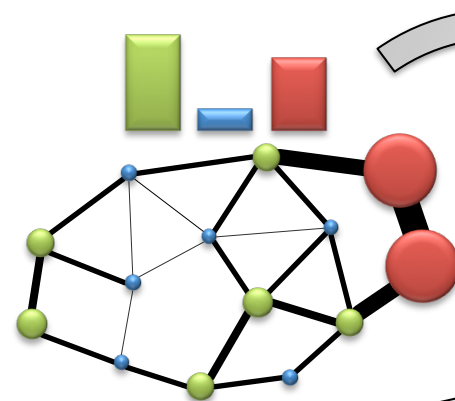
Modified category
weights



Edge weights in G

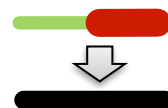
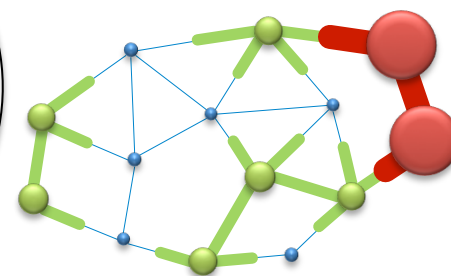


E.g., compare the size of
red and green categories.



Target edge weights

$$\frac{\text{Green bar}}{20} = \text{Green line} \quad \frac{\text{Blue bar}}{22} = \text{Blue line} \quad \frac{\text{Red bar}}{4} = \text{Red line}$$



Resolve conflicts:

- arithmetic mean,
- geometric mean,
- max,
- ...

Measurement objective



Category weights
optimal under WIS



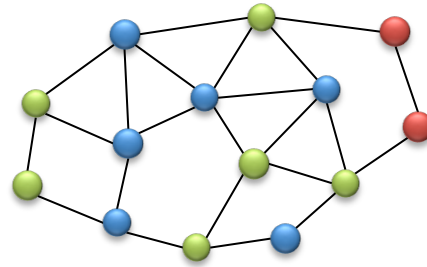
Modified category
weights



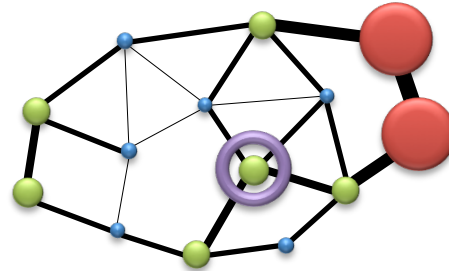
Edge weights in G



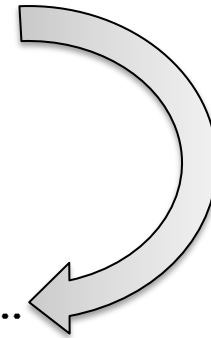
WRW sample



E.g., compare the size of
red and green categories.



$$S = (v_1, w_{v_1}), (v_2, w_{v_2}), (v_3, w_{v_3}), \dots$$



Measurement objective



Category weights optimal under WIS



Modified category weights



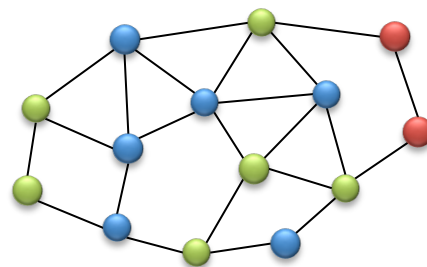
Edge weights in G



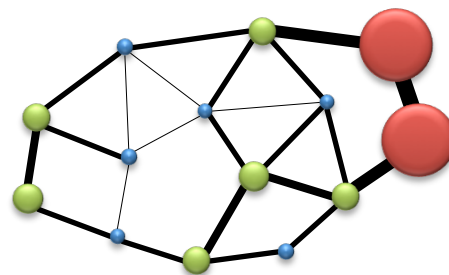
WRW sample



Final result

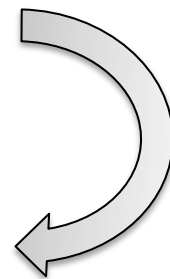


E.g., compare the size of red and green categories.



$$S = (v_1, w_{v_1}), (v_2, w_{v_2}), (v_3, w_{v_3}), \dots$$

$$\frac{\text{size}(\text{green})}{\text{size}(\text{red})} = \frac{\sum_{v \in S} 1_{\{v \text{ is green}\}} / w_v}{\sum_{v \in S} 1_{\{v \text{ is red}\}} / w_v}$$



Hansen-Hurwitz estimator

Measurement objective



Category weights
optimal under WIS



Modified category
weights



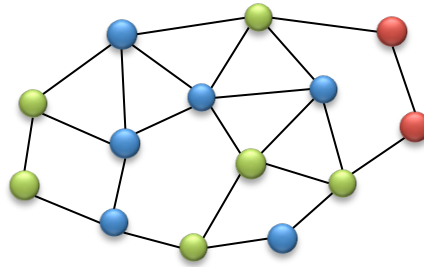
Edge weights in G



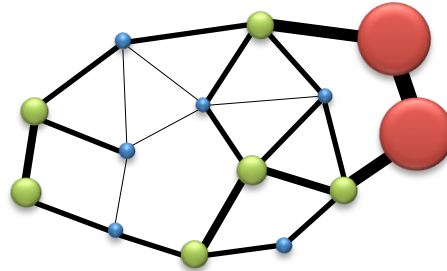
WRW sample



Final result



E.g., compare the size of
red and green categories.



Stratified Weighted Random
Walk (S-WRW)
[Sigmetrics '11]

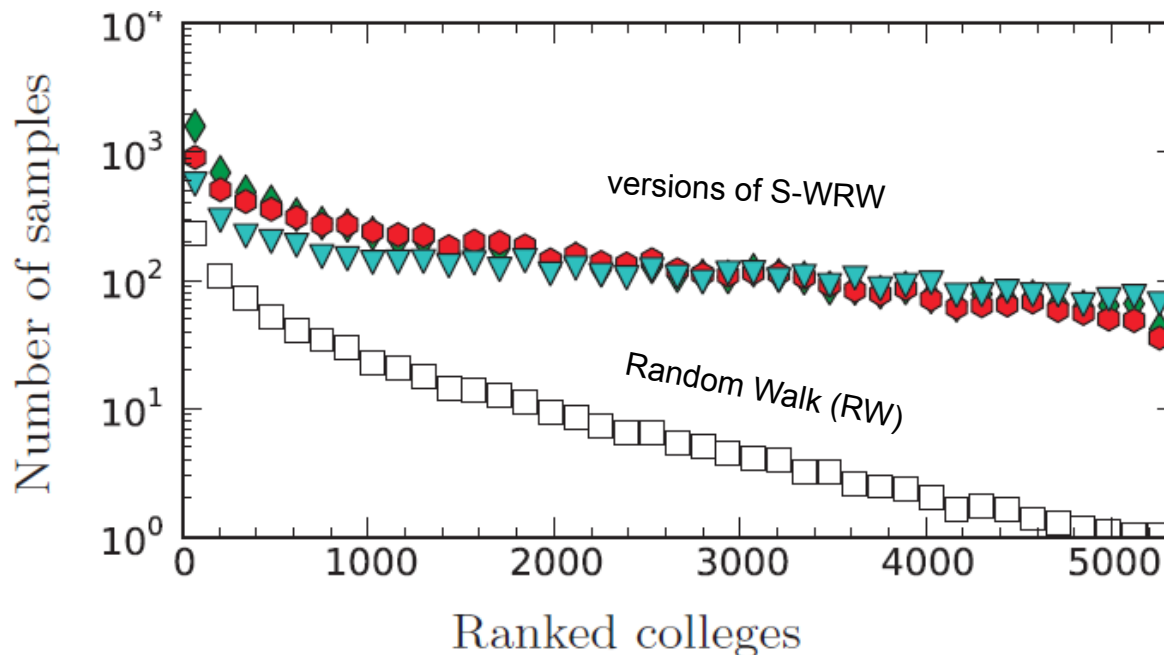
$$S = (v_1, w_{v_1}), (v_2, w_{v_2}), (v_3, w_{v_3}), \dots$$

$$\frac{\text{size}(\text{green})}{\text{size}(\text{red})} = \frac{\sum_{v \in S} 1_{\{v \text{ is green}\}} / w_v}{\sum_{v \in S} 1_{\{v \text{ is red}\}} / w_v}$$

Example:

colleges in Facebook

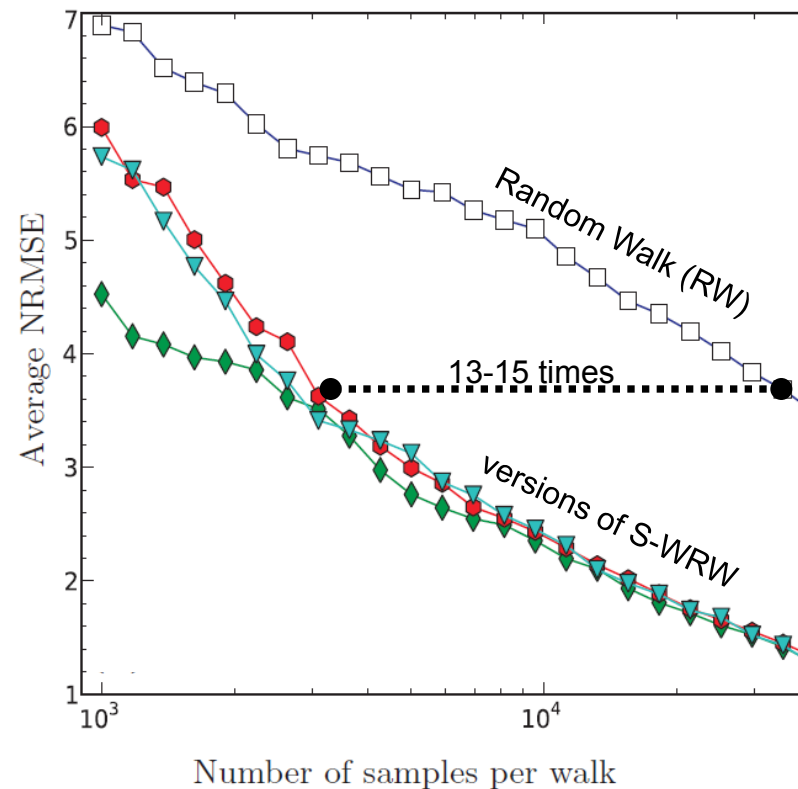
3.5% of Facebook users declare memberships in colleges



Let's compare a sample of 1M nodes collected by RW vs S-WRW

- RW visited college users in 9% of samples; S-WRW in 86% of samples
- RW discovered 5,325 and S-WRW 8,815 unique colleges
- S-WRW collects 10-100 times more samples per college (avoid irrelevant categories)
- This difference is larger for small colleges (because of stratification)

Example continued: colleges in Facebook



RW needs 13-15 times more samples to achieve the same error

S-WRW: Stratified Weighted Random Walk

Summary

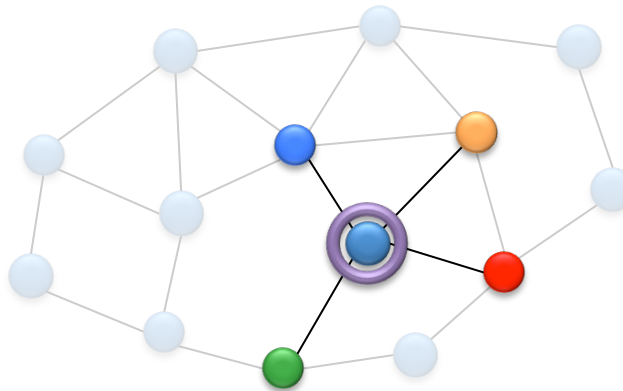
- Walking on a weighted graph
 - weights control the tradeoff between stratification and convergence
- Unbiased estimation
- Setting of weights affects efficiency
 - currently heuristic, optimal weight setting is an open problem
 - S-WRW “conservative”: between RW and WIS
 - Robust in practice
- Does not assume a-priori knowledge of graph or categories
- M. Kurant, M. Gjoka, C. T. Butts and A. Markopoulou, “Walking on a Graph with a Magnifying Glass”, to appear in *ACM SIGMETRICS*, June 2011.

Outline

- Introduction
- Sampling Techniques
 - Random Walks/BFS for sampling Facebook
 - Multigraph Sampling
 - Stratified Weighted Random Walk
- What can we learn from a sample?
- Conclusion and Future Directions

Information Collected - revisited

at each sampled node



Always:

- Observe attributes of sampled node
- Observe (number of) edges incident to node

Usually, possible through HTML scraping:

- Observe ids and attributes of neighbors

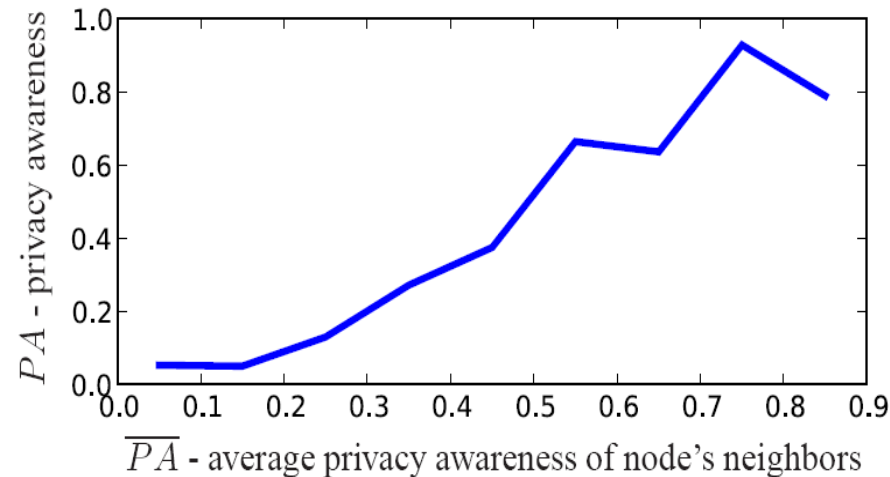
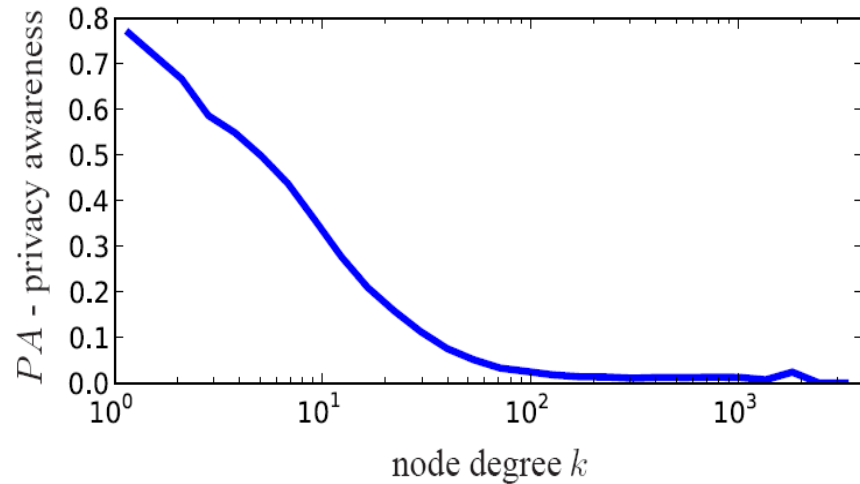
Can also collect complete egonet (=neighbors of neighbors)

What can we estimate

based on sample of nodes?

- Frequency of nodal attributes
 - Personal data: gender, age, name etc...
 - Privacy settings : it ranges from 1111 (all privacy settings on) to 0000 (all privacy settings off)
 - Membership to a "category": university, regional network, group
- Local topology properties
 - Degree distribution
 - Assortativity
 - Clustering coefficient
- Global structural properties???
- Edges or nodal attributes not observed in the sample?
- M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, "Practical Recommendations for Sampling OSN Users by Crawling the Social Graph", to appear in *IEEE JSAC on Measurements of Internet Topologies 2011*.

Privacy Awareness in Facebook

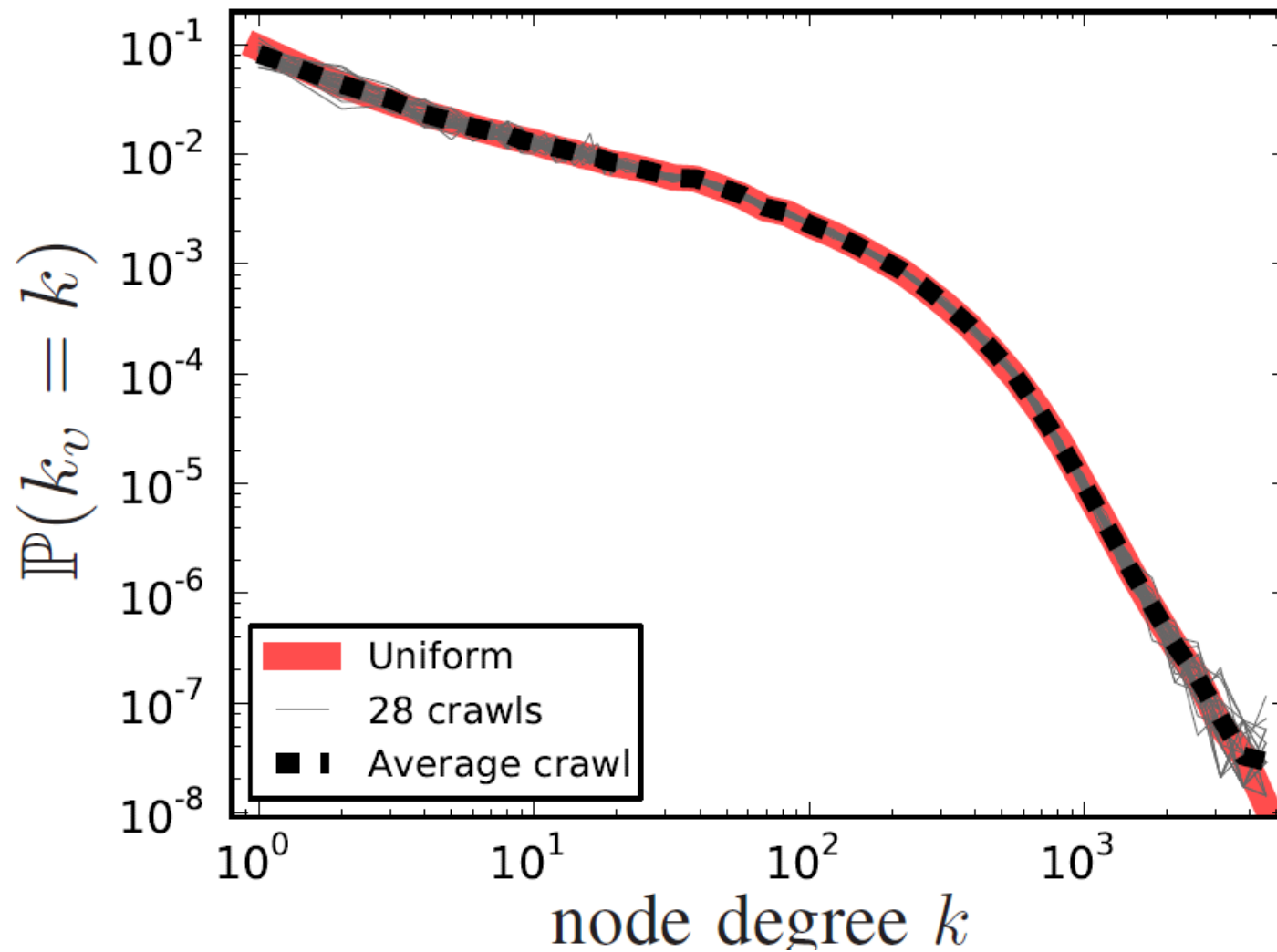


PA = Probability that a user changes the default (off) privacy settings

PA	Network n	PA	Network n
0.08	Iceland
0.11	Denmark	0.22	Bangladesh
0.11	Provo, UT	0.23	Hamilton, ON
0.11	Ogden, UT	0.23	Calgary, AB
0.11	Slovakia	0.23	Iran
0.11	Plymouth	0.23	India
0.11	Eastern Idaho, ID	0.23	Egypt
0.11	Indonesia	0.24	United Arab Emirates
0.11	Western Colorado, CO	0.24	Palestine
0.11	Quebec City, QC	0.25	Vancouver, BC
0.11	Salt Lake City, UT	0.26	Lebanon
0.12	Northern Colorado, CO	0.27	Turkey
0.12	Lancaster, PA	0.27	Toronto, ON
0.12	Boise, ID	0.28	Kuwait
0.12	Portsmouth	0.29	Jordan
...	...	0.30	Saudi Arabia

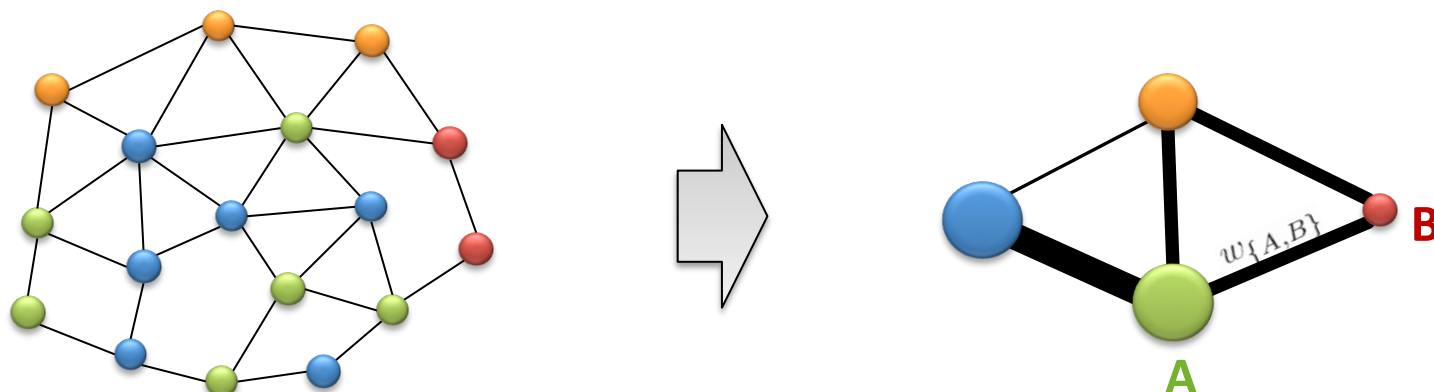
Degree Distribution

or the frequency of any node attribute frequency



What about network structure based on sample of nodes?

- A coarse-grained topology: category-to-category graph.



- Categories are declared by user/node (not inferred via community detection)
- Weight of edge between categories can be defined in a number of ways

$$w(A, B) = \frac{|E_{A,B}|}{|A| \cdot |B|}$$

- Probability that a random node in **A** is a friend of a random node in **B**
- Trivial to compute if the graph known. Must be estimated based on sample..

Estimating category size and edge weights

Uniform sample of nodes, induced subgraph

$$|\hat{A}| = N \cdot \frac{|S_A|}{|S|},$$

$$\hat{w}(A, B) = \frac{\sum_{a \in S_A} \sum_{b \in S_B} 1_{\{\{a,b\} \in E\}}}{|S_A| \cdot |S_B|}.$$

Uniform sample of nodes, star sampling

$$|\hat{A}| = N \cdot \hat{f}_A^{\text{vol}} \cdot \frac{\bar{k}_V}{\bar{k}_A},$$

$$\hat{w}(A, B) = \frac{\sum_{a \in S_A} |E_{a,B}| + \sum_{b \in S_B} |E_{b,A}|}{|S_A| \cdot |\hat{B}| + |S_B| \cdot |\hat{A}|}.$$

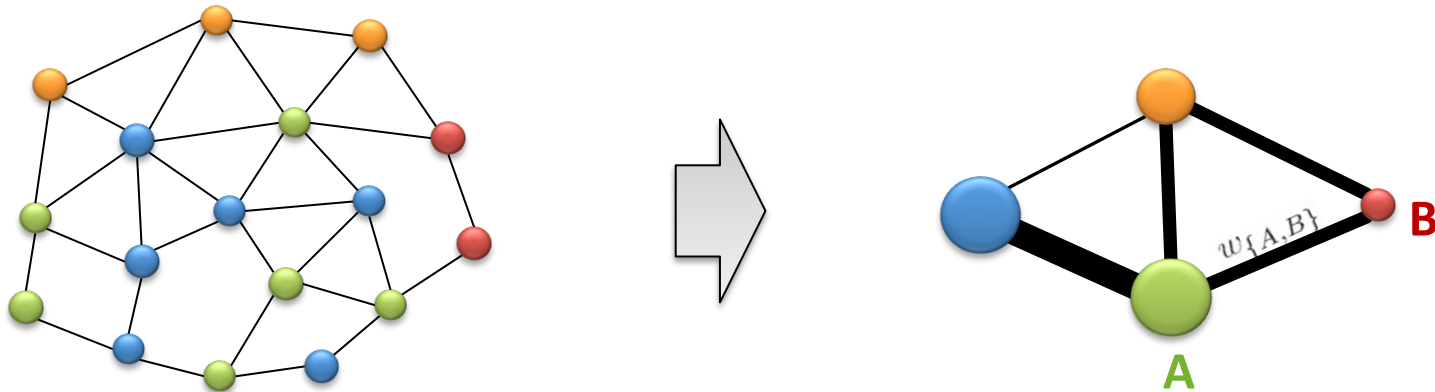
Non-uniform sample of nodes, star sampling

$$\hat{w}(A, B) = \frac{\sum_{a \in S_A} \frac{|E_{a,B}|}{w(a)} + \sum_{b \in S_B} \frac{|E_{b,A}|}{w(b)}}{w_{\cdot 1}(S_A) \cdot |\hat{B}| + w_{\cdot 1}(S_B) \cdot |\hat{A}|}.$$

- Weighting guided by Hansen-Hurwitz
- Showed consistency

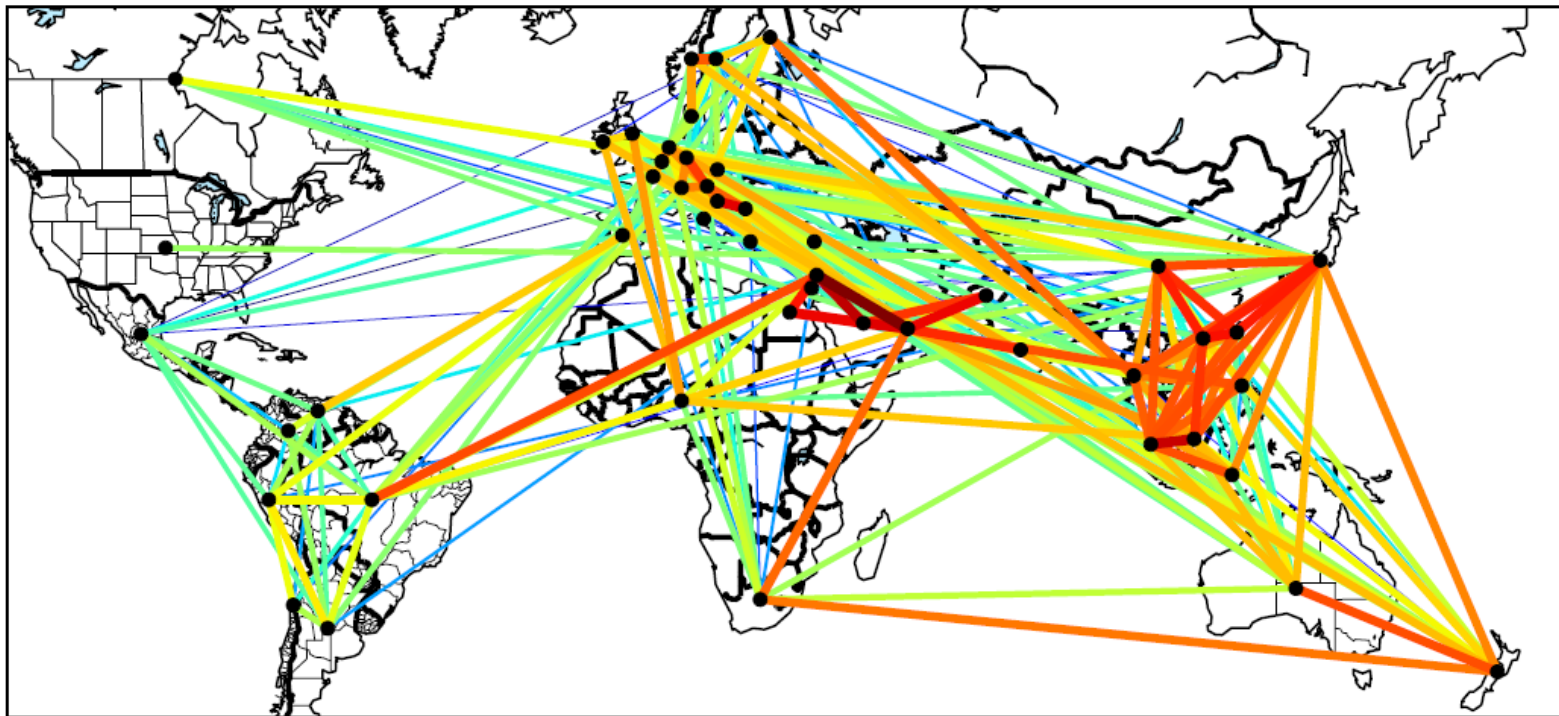
What about network structure based on sample of nodes?

- A coarse-grained topology: category-to-category graph.

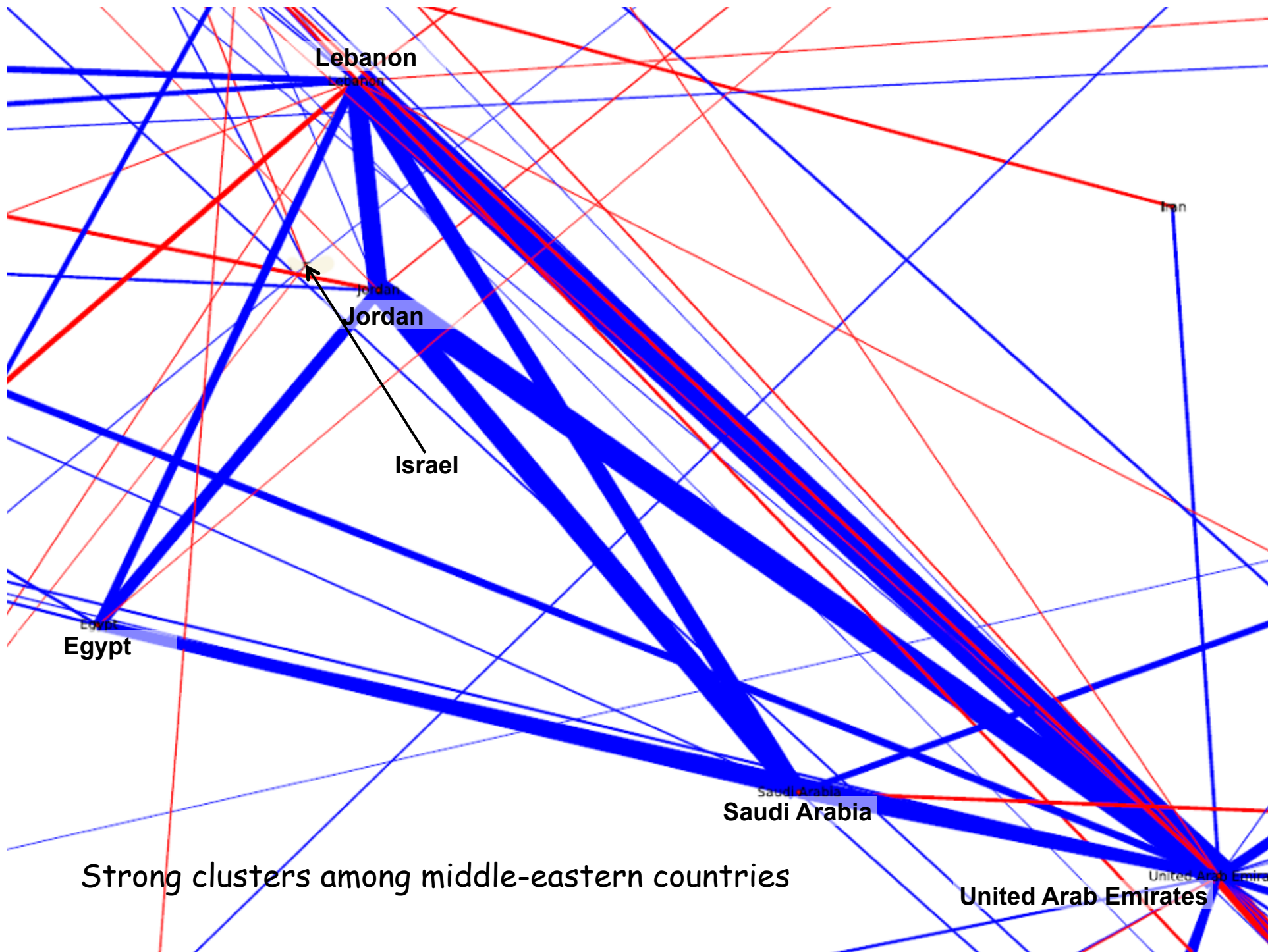


- M.Kurant, M.Gjoka, Y.Wang, Z.Almquist, C.T.Butts, A. Markopoulou, "Coarse-grained Topology Estimation via Graph Sampling", on *arXiv.org*, May 2011.
- Visualization available at: <http://www.geosocialmap.com>

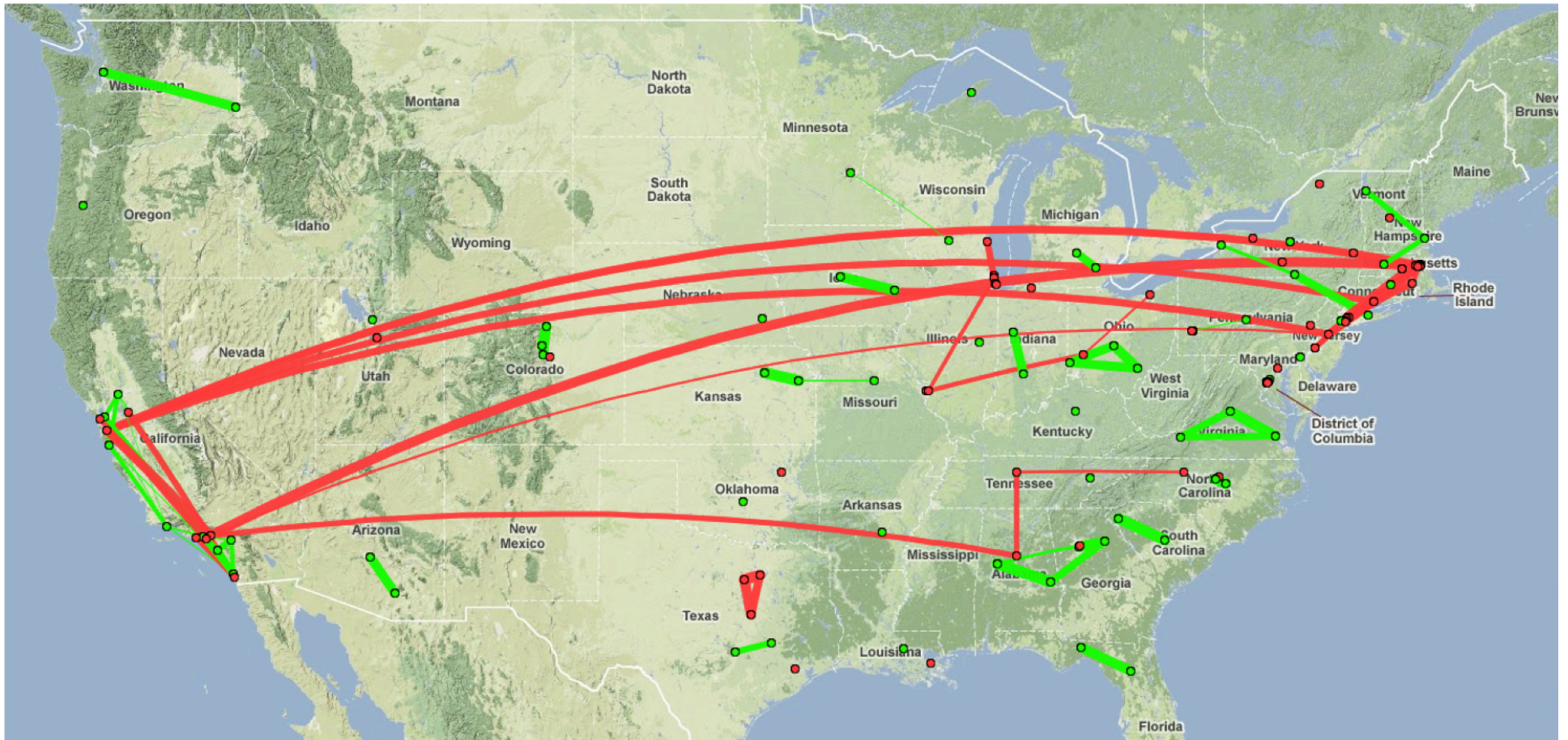
Country-to-country FB graph



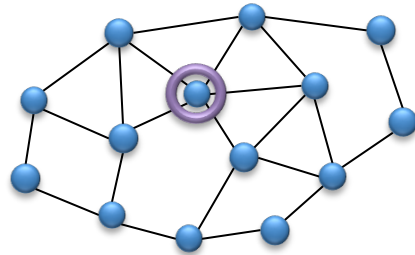
- Some observations (300 strongest edges between 50 countries)
 - Clusters with strong ties in Middle East and South Asia
 - Inwardness of the US
 - Many strong, outwards edges from Australia and New Zealand



Top US Colleges: public vs. private

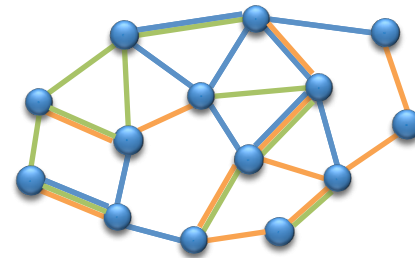


Physical distance is a major factor in ties between public (green), but not between private schools (red)
More generally, potential applications: descriptive uses, input to models

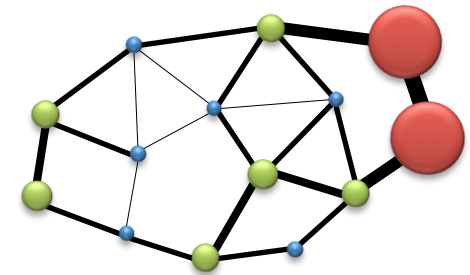


Random Walks [1,2,3]

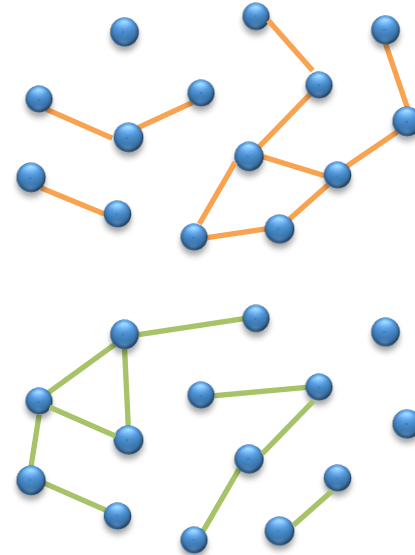
- RWRW > MHRW [1]
- vs. BFS, RW, uniform
- Bias, efficiency
- Convergence diagnostics [1]
- First unbiased sample of Facebook nodes [1,6]



Multigraph sampling [2,6]

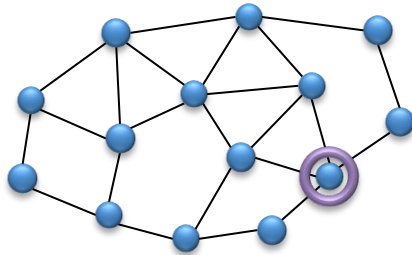


Stratified WRW [3,6]



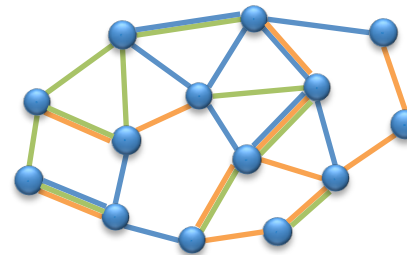
References

- [1] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs", in *INFOCOM 2010 and JSAC 2011*
- [2] M. Gjoka, C. T. Butts, M. Kurant and A. Markopoulou, "Multigraph Sampling of Online Social Networks", to appear in *IEEE JSAC 2011*
- [3] M. Kurant, M. Gjoka, C. T. Butts and A. Markopoulou, "Walking on a Graph with a Magnifying Glass", in *ACM SIGMETRICS 2011*.
- [4] M. Kurant, A. Markopoulou and P. Thiran, "On the bias of BFS (Breadth First Search)", *ITC 22, 2010 and IEEE JSAC 2011*.
- [5] M. Kurant, M. Gjoka, Y. Wang, Z. Almquist, C. T. Butts and A. Markopoulou, "Coarse Grained Topology Estimation via Graph Sampling", in *arxiv.org*
- [6] Facebook datasets: <http://odysseas.calit2.uci.edu/research/osn.html>
- [7] Visualization of Facebook category graphs: www.geosocialmap.com

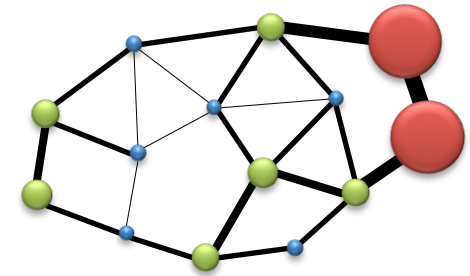


Random Walks [1,2,3]

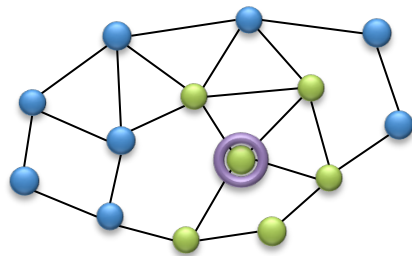
- RWRW > MHRW [1]
- vs. BFS, RW, uniform
- Bias, efficiency
- Convergence diagnostics [1]
- First unbiased sample of Facebook nodes [1,6]



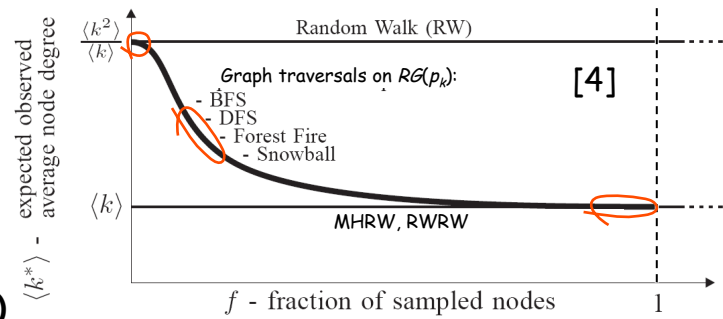
Multigraph sampling [2,6]



Stratified WRW [3,6]

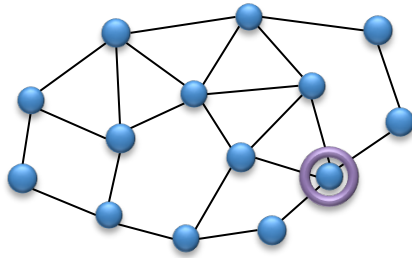


Sampling w/o replacement (BFS)



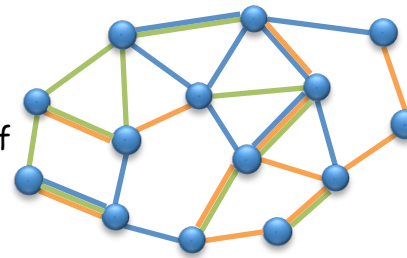
References

- [1] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs", in *INFOCOM 2010 and JSAC 2011*
- [2] M. Gjoka, C. T. Butts, M. Kurant and A. Markopoulou, "Multigraph Sampling of Online Social Networks", to appear in *IEEE JSAC 2011*
- [3] M. Kurant, M. Gjoka, C. T. Butts and A. Markopoulou, "Walking on a Graph with a Magnifying Glass", in *ACM SIGMETRICS 2011*.
- [4] M. Kurant, A. Markopoulou and P. Thiran, "On the bias of BFS (Breadth First Search)", *ITC 22, 2010 and IEEE JSAC 2011*.
- [5] M. Kurant, M. Gjoka, Y. Wang, Z. Almquist, C. T. Butts and A. Markopoulou, "Coarse Grained Topology Estimation via Graph Sampling", in *arxiv.org*
- [6] Facebook datasets: <http://odysseas.calit2.uci.edu/research/osn.html>
- [7] Visualization of Facebook category graphs: www.geosocialmap.com

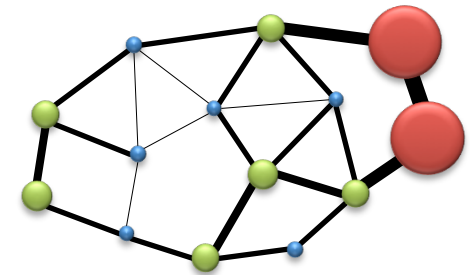


Random Walks [1,2,3]

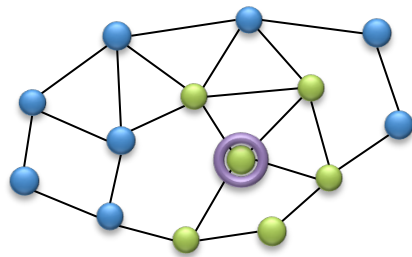
- RWRW > MHRW [1]
- vs. BFS, Uniform
- The first unbiased sample of Facebook nodes [1,6]
- Convergence diagnostics [1]



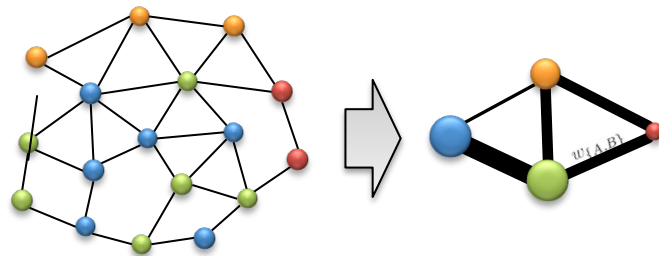
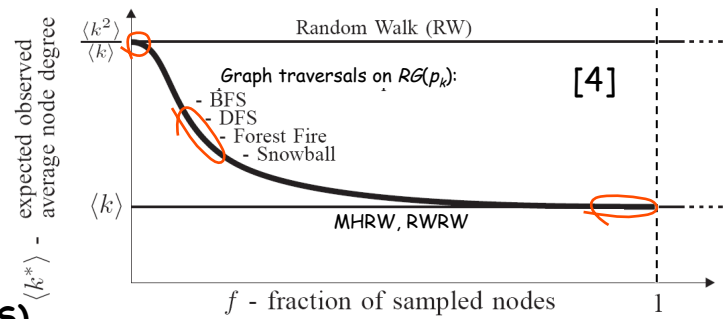
Multigraph sampling [2]



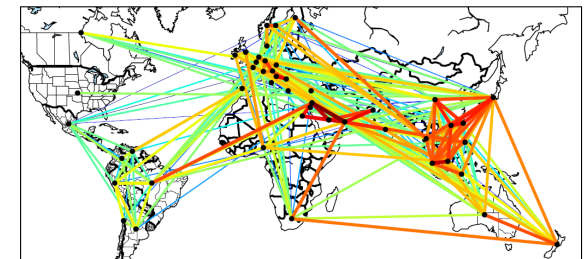
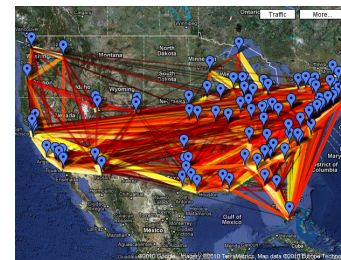
Stratified WRW [3]



Sampling w/o replacement (BFS)



Coarse-grained topologies [5,7]



References

- [1] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs", in *INFOCOM 2010 and JSAC 2011*
- [2] M. Gjoka, C. T. Butts, M. Kurant and A. Markopoulou, "Multigraph Sampling of Online Social Networks", to appear in *IEEE JSAC 2011*
- [3] M. Kurant, M. Gjoka, C. T. Butts and A. Markopoulou, "Walking on a Graph with a Magnifying Glass", in *ACM SIGMETRICS 2011*.
- [4] M. Kurant, A. Markopoulou and P. Thiran, "On the bias of BFS (Breadth First Search)", *ITC 22, 2010 and IEEE JSAC 2011*.
- [5] M. Kurant, M. Gjoka, Y. Wang, Z. Almquist, C. T. Butts and A. Markopoulou, "Coarse Grained Topology Estimation via Graph Sampling", in *arxiv.org*
- [6] Facebook datasets: <http://odysseas.calit2.uci.edu/research/osn.html>
- [7] Visualization of Facebook category graphs: www.geosocialmap.com