Crawling of Online Social Networks

Athina Markopoulou

Joint work with:

Minas Gjoka, Maciej Kurant and Carter T. Butts

1Department of Electrical Engineering and Computer Science
2Department of Sociology
3CalIT2: California Institute of Information Technologies
University of California, Irvine
<table>
<thead>
<tr>
<th>Size</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 million</td>
<td>2</td>
</tr>
<tr>
<td>200 million</td>
<td>9</td>
</tr>
<tr>
<td>130 million</td>
<td>12</td>
</tr>
<tr>
<td>100 million</td>
<td>43</td>
</tr>
<tr>
<td>75 million</td>
<td>10</td>
</tr>
<tr>
<td>75 million</td>
<td>29</td>
</tr>
</tbody>
</table>

(November 2010)

> 1 billion users

>15% of world's population

>50% of world's Internet users

**Rich Activity:** email communication (FB messages), voice and video communication (e.g. skype), photos and videos (flickr, youtube), news, ...
Why study Online Social Networks?

Difference communities have different perspective

- **Social Sciences**
  - Fantastic source of data for studying online behavior
  - Large amount of data, rich interactions, automated collection

- **Large scale data mining**
  - Social influence/marketing
  - User communication patterns
  - Visualization

- **Engineering**
  - **OSN provider**
    - Design high performance/scalable OSNs [e.g. "Little Engines"@Sigcomm2010]
    - Optimize server placement or mapping of users to servers
  - **Network provider**
    - OSN traffic shape the Internet traffic
    - Support mechanisms: e.g. caching, delivery to mobiles
    - Establish trust - network security
  - **Third party services**
    - Establish trust - network security
    - Personalized services
Measuring OSNs

You may have the complete dataset

You may have not the complete dataset
• Let us try to collect the entire social graph of Facebook:
  – >500 million users
  – 130 friends each (on average)
  – 8 bytes (64 bits) per user ID
• Just the connectivity data, without attributes:
  – 500 x 130 x 8B = 520 GB
• To get this data, one would have to download:
  – 260 TB of HTML data!
  – Bandwidth, disk, time constraints
• This is neither feasible nor practical.

You may want to operate on a small dataset

Solution: Sampling!
Population Sampling Unframed

• 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
  – rejection sampling User IDs: may be impossible (not supported by API) or inefficient (rate limited, sparse user ID space).

• Alternative: network-based sampling methods
  – Exploit social ties to draw a probability sample
  – Today: a look at some of our work in link-trace sampling (crawling/exoration/… of the graph)
Many variants depending on the way we choose to follow the links to neighbors.
BFS, Illustrated
BFS, Illustrated
BFS, Illustrated
BFS, Illustrated
BFS, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Random Walk, Illustrated
Goal: to obtain a representative sample of... ....

Graph sampling

- Independent
  - access to all nodes
  - used as a reference

- Exploration
  - access to neighbors only
  - typical of OSNs

- Traceroute
  - sample paths
  - specific to the Internet
  - IP or AS level

- Walks
  - with repetitions
  - RW, RWRW, MHRW, WRW, ...

- Traversals
  - no repetitions
  - BFS, DFS, Snowball...
Challenges in practice

• **Before the crawl**
  – Define the graph (users, relations to crawl)
  – Pick crawling method for lack of bias and efficiency
  – Programming: efficient crawlers, access limitations

• **During the crawl**
  – When to stop? Online convergence diagnostics.

• **After the walk**
  – What to drop?
  – Correct for the bias
  – Evaluate success? ground truth?
Related Work

- **MCMC literature**

- **Graph sampling techniques**
  - BFS/traversal
    - [Mislove et al. '07, Ahn et al. '07, Wilson et al. '09, Ye et al. '10, Leskovec et al. '06]
  - Random walks
    - [Henzinger et al. '00, Stutbach et al. '06, Leskovec et al. '06, Rasti et al. '09]

- **Characterization studies of OSNs**
  - Cyworld, Orkut, Myspace, Flickr, Youtube [...]
  - Facebook [Wilson et al. '09, Krishnamurthy et al. '08]
Outline

• Single graph: comparative study + implementation

• Multigraph sampling

• Weighted random walk

• Quantifying and correcting for the bias of BFS

• From nodes to topologies
Outline

• Single graph
  – Goal: obtain a uniform sample of FB users via crawling
  – Assumptions: a connected, static, undirected graph
  – Approach: apply MCMC techniques + provide practical recommendations

• Multigraph sampling

• Weighted random walk

• Correcting for the bias of BFS

• From nodes to topologies
Method 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

• Early measurement studies of OSNs use BFS as primary sampling technique
  i.e [Mislove et al], [Ahn et al], [Wilson et al.]
Method 2: Random Walk (RW)

- Explores graph one node at a time with replacement

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

Degree of node \( v \)

- In the stationary distribution

\[ \pi_v = \frac{k_v}{2 |E|} \]

Number of edges

Next candidate

Current node
Method 3: Re-Weighted Random Walk (RWRW)

• Corrects for degree bias at the end of collection

• Property estimation without re-weighting:

\[ p(u \text{ from Greece}) = \frac{\sum_{u \text{ from Greece}} 1}{\sum_{u \in V} 1} = \frac{|\text{users from Greece}|}{|\text{all users}|} \]

• Property estimation with re-weighting:
  - Hansen-Hurwitz estimator

\[ p(u \text{ from Greece}) = \frac{\sum_{u \text{ from Greece}} 1/k_u}{\sum_{u \in V} 1/k_u} = \]

Degree of node u
Method 4: Metropolis-Hastings Random Walk (MHRW)

- Random walk with:
  \[ P_{v,w}^{MH} = \begin{cases} 
  \frac{1}{k_v} \min(1, k_w) & \text{if } w \text{ neighbor of } v \\
  1 - \sum_{y \neq v} P_{v,y}^{MH} & \text{if } w = v
  \end{cases} \]

- In the stationary distribution:
  \[ \pi_v = \frac{1}{|V|} \]
(5) Uniform userID Sampling (UNI)

- As a basis for comparison, we collect a uniform sample of Facebook userID (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
Data Collection
Sampled Node Information

- UserID
- Name
- Networks
- Privacy settings

Friend List

- UserID
- Name
- Networks
- Privacy settings

Regional
School/Workplace

1111
Send Message
View Friends
Profile Photo
Add as Friend
Data Collection

Challenges

• Facebook is not easy to crawl
  – rich client side Javascript
  – stronger than usual privacy settings
  – limited data access when using API
  – unofficial rate limits that result in account bans
  – large scale
  – growing daily

• Designed and implemented efficient OSN crawlers
  • API + HTML scraping
Distributed Crawling

- Careful implementation is important

- Decreased time to crawl ~1million users from ~2 weeks to <2 days.
Speeding up Crawling
Parallelization

Seed nodes

Queue

Visited

User Account Server

Pool of threads

Distributed data fetching
- cluster of 50 machines
- coordinated crawling

Multiple walks/traversals
Per walk
- multiple threads
- limited caching (usually FIFO)

RW, MHRW, BFS
## Dataset

### April-May 2009

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>MHRW</th>
<th>RW</th>
<th>BFS</th>
<th>UNI</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Sampled Users</td>
<td>28x81K</td>
<td>28x81K</td>
<td>28x81K</td>
<td>984K</td>
</tr>
<tr>
<td># Unique Users</td>
<td>957K</td>
<td>2.19M</td>
<td>2.20M</td>
<td>984K</td>
</tr>
</tbody>
</table>

**First dataset:**
- Publicly available at: [http://odysseas.calit2.uci.edu/research/osn.html](http://odysseas.calit2.uci.edu/research/osn.html)
- ~300 requests in the last 6 months

**Latest dataset:**
- Oct. 2010: ~2 days, 25 independent walks for a RW sample of 1M users
- will be posted soon
Detecting Convergence

• MCMC literature can be applied here
• Online diagnostics:
  • no ground truth available in practice

• What to drop:
  • number of samples to lose dependence from seed nodes
• When to stop:
  • Number of samples to declare convergence and stop
Online Convergence Diagnostics

Geweke

- Detects (lack of) convergence for a single walk.
- Let $X$ be a sequence of samples for metric of interest i.e. node degree

\[
Z = \frac{E(X_a) - E(X_b)}{\sqrt{Var(X_a) - Var(X_b)}}
\]

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 across several (m>1) walks

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

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

Overall: acceptable convergence between 500 and 3000 iterations (depending on property)
Comparison in terms of Bias

Node Degree

- MHRW - Metropolis-Hastings Random Walk
- RWRW - Re-Weighted Random Walk
- RW - Random Walk
- BFS - Breadth First Search
MHRW vs. RWRW

efficiency comparison: error vs. #samples

\[ D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \]
MHRW vs. RWRW

<table>
<thead>
<tr>
<th>Network</th>
<th>degree distribution (CDF)</th>
<th>KS distance ALL degrees</th>
<th>mean degree error</th>
<th>fraction 0.5 error all</th>
<th>fraction 0.5 error lowest degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cond-mat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND_WWW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- degree
- walk length (unique nodes)

Values:
- 5.47
- 1.06
- 4.83
- 1.72
- 1.56
- 5.9
- 1.23
- 4.83
- 2.13
- 2.14
- 6.95
- 2.23
- 2.23
- 2.09
- 2.09
- 3.15
- 1.31
- 5.34
- 1.31
- 3.15
- 2.09

~3.0
MHRW vs. RWRW

• RWRW converges faster than MHRW
  – for all practical purposes (1.5-8 times faster)
    • Why? repetition + avoiding high degree nodes (mixing)
  – pathological counter-examples exist.

• Both do the job: they yield an unbiased sample

• MHRW easy/ready to use - does not require reweighting
What can we learn from the dataset?

Characterization (1)

- Based on a uniform node sample, estimate any node property:
  - Degree distribution
  - Membership information
  - Privacy settings:
    - It ranges from 1111 (all privacy settings on) to 0000 (all privacy settings off)

...
E.g. Privacy Awareness in Facebook

\[ PA = \text{Probability that a user changes the default (off) privacy settings} \]

<table>
<thead>
<tr>
<th>( PA )</th>
<th>Network ( n )</th>
<th>( PA )</th>
<th>Network ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.08</td>
<td>Iceland</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0.11</td>
<td>Denmark</td>
<td>0.22</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>0.11</td>
<td>Provo, UT</td>
<td>0.23</td>
<td>Hamilton, ON</td>
</tr>
<tr>
<td>0.11</td>
<td>Ogden, UT</td>
<td>0.23</td>
<td>Calgary, AB</td>
</tr>
<tr>
<td>0.11</td>
<td>Slovakia</td>
<td>0.23</td>
<td>Iran</td>
</tr>
<tr>
<td>0.11</td>
<td>Plymouth</td>
<td>0.23</td>
<td>India</td>
</tr>
<tr>
<td>0.11</td>
<td>Eastern Idaho, ID</td>
<td>0.23</td>
<td>Egypt</td>
</tr>
<tr>
<td>0.11</td>
<td>Indonesia</td>
<td>0.24</td>
<td>United Arab Emirates</td>
</tr>
<tr>
<td>0.11</td>
<td>Western Colorado, CO</td>
<td>0.24</td>
<td>Palestine</td>
</tr>
<tr>
<td>0.11</td>
<td>Quebec City, QC</td>
<td>0.25</td>
<td>Vancouver, BC</td>
</tr>
<tr>
<td>0.11</td>
<td>Salt Lake City, UT</td>
<td>0.26</td>
<td>Lebanon</td>
</tr>
<tr>
<td>0.12</td>
<td>Northern Colorado, CO</td>
<td>0.27</td>
<td>Turkey</td>
</tr>
<tr>
<td>0.12</td>
<td>Lancaster, PA</td>
<td>0.27</td>
<td>Toronto, ON</td>
</tr>
<tr>
<td>0.12</td>
<td>Boise, ID</td>
<td>0.28</td>
<td>Kuwait</td>
</tr>
<tr>
<td>0.12</td>
<td>Portsmouth</td>
<td>0.29</td>
<td>Jordan</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>0.30</td>
<td>Saudi Arabia</td>
</tr>
</tbody>
</table>
What can we learn from the dataset?
Characterization (2)

- Also collected extended egonets for a subsample of MHRW
  - 37k egonets with ~6 million neighbors
  - Computed assortativity, clustering
What can we learn from the dataset?

Characterization (3)

- From a random sample of nodes, we can also estimate the topology at coarse granularity (membership declared by user)
  - country-to-country graph
  - university-to-university graph
  - community-to-community graph
Some observations:

- Clusters with strong ties in Middle East and South Asia
- Inwardness of the US
- Many strong and outwards edges from Australia and New Zealand
Strong clusters among middle-eastern countries
Single Graph Crawling

Summary

• Compared graph crawling methods
  – MHRW, RWRW performed remarkably well
  – RWRW (more efficient) vs. MHRW (ready to use)
  – BFS, RW lead to substantial bias

• Practical recommendations
  – usage of online convergence diagnostics
  – proper use of multiple chains
  – Implementation matters

• Obtained the first unbiased sample of Facebook
  – http://odysseas.calit2.uci.edu/research/osn.html

• Characterization of some FB properties
  – Node properties
  – Egonets
  – Topology at coarse granularity

• Reference
Outline

• Single graph: comparative study + implementation

• Multigraph sampling

• Weighted random walk

• Correcting for the bias of BFS

• From nodes to topologies
Outline

• Single graph: comparative study + implementation

• Multigraph sampling

• Weighted random walk

• Correcting for the bias of BFS

• From nodes to topologies
From a Single to Multiple Graphs

• **Motivation**
  – Often, no single network on a given population supports sampling
  – May be fragmented or clustered/heterogeneous

• **Idea: multigraph sampling**
  – Consider several relations\(\rightarrow\) graphs on the same set of nodes
  – Union of graphs has better properties than individual graphs
  – Walk on union graph or on multiple graphs
Fragmented social graph

Friendship

Event attendance

Group membership

Union
Highly clustered social graph

Friendship

Event attendance

Union
Combining of multiple relations

$G_*$ is a union multigraph

$G$ is a union graph
Multigraph sampling

Friends + Events + Groups

($G_*$ is a multigraph)

Approach 1:
1) Select edge to follow uniformly at random, i.e., with probability $1 / \text{deg}(F, G_*)$

Approach 2: \textbf{does not require listing all neighbours, may save some bandwidth}
1) Select relation graph $G_i$ with probability $\text{deg}(F, G_i) / \text{deg}(F, G_*)$
2) Within $G_i$, choose an edge uniformly at random, i.e., with prob $1 / \text{def}(F, G_i)$. 
Multigraph sampling
Implemented efficiently

\[ p(\text{Friends}) = \frac{1}{8} \]
\[ p(\text{Events}) = \frac{4}{8} \]
\[ p(\text{Groups}) = \frac{3}{8} \]

Degree information available without enumeration
Take advantage of pages functionality
Multigraph sampling

Case study: Last.fm

• Last.fm, an Internet radio service
  – social networking features
  – multiple relations
  – fragmented graph components and highly clustered users expected

• Last.fm relations used
  – Friends
  – Groups
  – Events
  – Neighbors
## Summary of datasets

_Last.fm - July 2010_

<table>
<thead>
<tr>
<th>Crawl type</th>
<th># Total Users</th>
<th>% Unique Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>5x50K</td>
<td>71%</td>
</tr>
<tr>
<td>Events</td>
<td>5x50K</td>
<td>58%</td>
</tr>
<tr>
<td>Groups</td>
<td>5x50K</td>
<td>74%</td>
</tr>
<tr>
<td>Neighbors</td>
<td>5x50K</td>
<td>53%</td>
</tr>
<tr>
<td>Friends-Events-Groups-Neighbors</td>
<td>5x50K</td>
<td>76%</td>
</tr>
<tr>
<td>UNI</td>
<td>500K</td>
<td>99%</td>
</tr>
</tbody>
</table>
## Isolated users

*By relation type*

<table>
<thead>
<tr>
<th>Crawl type</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>60.4%</td>
</tr>
<tr>
<td>Events</td>
<td>41.7%</td>
</tr>
<tr>
<td>Groups</td>
<td>0%</td>
</tr>
<tr>
<td>Neighbors</td>
<td>62.4%</td>
</tr>
<tr>
<td>Friends-Events-Groups-Neighbors</td>
<td>86.3%</td>
</tr>
<tr>
<td><strong>UNI</strong></td>
<td><strong>93.8%</strong></td>
</tr>
</tbody>
</table>
## Isolated users

By relation type

<table>
<thead>
<tr>
<th>Crawl type</th>
<th>Groups</th>
<th>Future Events</th>
<th>Past Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>60.4%</td>
<td>93.7%</td>
<td>73.2%</td>
</tr>
<tr>
<td>Events</td>
<td>41.7%</td>
<td>78.2%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Groups</td>
<td>0%</td>
<td>89.9%</td>
<td>62.0%</td>
</tr>
<tr>
<td>Neighbors</td>
<td>62.4%</td>
<td>89.5%</td>
<td>71.2%</td>
</tr>
<tr>
<td>Friends-Events-Groups-Neighbors</td>
<td>86.3%</td>
<td>98.3%</td>
<td>86.7%</td>
</tr>
<tr>
<td>UNI</td>
<td>93.8%</td>
<td>99.2%</td>
<td>96.1%</td>
</tr>
</tbody>
</table>
# Isolated users

By relation type

<table>
<thead>
<tr>
<th>Crawl type</th>
<th>Groups</th>
<th>Future Events</th>
<th>Past Events</th>
<th>Friends</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>60.4%</td>
<td>93.7%</td>
<td>73.2%</td>
<td>0%</td>
</tr>
<tr>
<td>Events</td>
<td>41.7%</td>
<td>78.2%</td>
<td>4.5%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Groups</td>
<td>0%</td>
<td>89.9%</td>
<td>62.0%</td>
<td>21.2%</td>
</tr>
<tr>
<td>Neighbors</td>
<td>62.4%</td>
<td>89.5%</td>
<td>71.2%</td>
<td>40.4%</td>
</tr>
<tr>
<td>Friends-Events-Groups-Neighbors</td>
<td>86.3%</td>
<td>98.3%</td>
<td>86.7%</td>
<td>7.4%</td>
</tr>
<tr>
<td>UNI</td>
<td>93.8%</td>
<td>99.2%</td>
<td>96.1%</td>
<td>87.9%</td>
</tr>
</tbody>
</table>
Comparison to Uniform

% of Subscribers
Related Work

• Sampling in fragmented graphs

• Last.fm studies
  – [Konstas et al. ’09] Track recommendation system
  – [Schifanella et al. ’10] Prediction of social links
Multigraph Sampling

Summary

• Multigraph sampling
  – simple concept, efficient implementation
  – discovers isolated nodes
  – Better estimates of distributions and means

• M. Gjoka, C. T. Butts, M. Kurant, A. Markopoulou, “Multigraph Sampling of Online Social Networks”

• Future direction
  – selection and weighting of relations
Outline

• Single graph: comparative study + implementation

• Multigraph sampling

• **Weighted random walk**

• Correcting for the bias of BFS

• From nodes to topologies
Blue and black nodes are equally important. We are not interested in white nodes.

Node weight is proportional to its sampling probability under WIS.

But the exploration techniques have to follow the links!

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

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

• Single graph: comparative study + implementation

• Multigraph sampling

• Weighted random walk

• Correcting for the bias of BFS

• From nodes to topologies
Why BFS?

- **BFS sample reveals parts of the graph**
  - E.g., BFS of a lattice is a lattice
  - We can study its topological characteristics (e.g., shortest path lengths, clustering coefficients, community structure), which is (in general) not possible with random walks

- **It is widely used in practice:**
  - ......
Real average node degree:
\[ \langle k \rangle = \sum_k k p_k \approx 94 \]

Observed average node degree:
\[ \sum_k k q_k \approx 338 \]

Why?!
Goal: Quantify BFS bias

The BFS bias has been empirically observed in the past, but not formally analyzed.
Graph model $\textit{RG}(p_k)$

- Random graph $\textit{RG}(p_k)$ with a given node degree distribution $p_k$
- Can be generated by \textit{configuration model}:

\begin{align*}
\text{Example:} \quad p_1 = p_2 = p_3 = p_4 = 0.25, \quad |V| = 4
\end{align*}
Solution (very briefly)

- Use The Principle of Deferred Decisions to generate $RG(\rho_k)$ “on the fly”
- Index stubs at random, uniformly and iid from [0,1]. This breaks the correlations between stubs and allows for analytical formulas.
- This powerful trick was originally proposed in a different context in [J. H. Kim, 2004], and next developed in [D. Achlioptas et al.]

Results
- Compare real degree distribution ($p_k$) to the expected degree distribution ($q_k$) seen by BFS (bias)

\[
q_k^{BFS}(f) = \frac{p_k(1-(1-t(f))^k)}{\sum_l p_l(1-(1-t(f))_l^k)}
\]

\[
\hat{p}_k^{BFS} = \frac{\hat{q}_k}{1-(1-t(f))^k} \cdot \left( \sum_i \frac{\hat{q}_i}{1-(1-t(f))^i} \right)^{-1}
\]
Results

exact for $RG(p_k)$

\[ \langle k \rangle - \text{real average node degree} \]
\[ \langle k^2 \rangle - \text{real average squared node degree}. \]
Results

exact for $RG(p_k)$

$\langle k^2 \rangle$ - real average squared node degree.

$\langle k \rangle$ - real average node degree
Results

exact for $RG(p_k)$

For small sample size (for $f \to 0$), BFS has the same bias as RW.

(observed in our Facebook measurements)

Graph traversals on $RG(p_k)$:

- BFS
- DFS
- Forest Fire
- Snowball

For large sample size (for $f \to 1$), BFS becomes unbiased.

This bias monotonically decreases with $f$.
We found analytically the shape of this curve.

$\langle k^2 \rangle$ - real average squared node degree.

$\langle k \rangle$ - expected observed average node degree

$\langle k^*_c \rangle$ - expected observed average node degree

Random Walk (RW)

MHRW, RWRW

$f$ - fraction of sampled nodes

$\langle k \rangle$ - real average node degree
Theory vs. Practice

Simulations on a power law random graph with 10K nodes

Exact for $RG(p_k)$
What if the graph is not random?
What if the graph is \textit{not} random?
On the bias of BFS
Summary

• We computed analytically the node degree bias of BFS in $R\mathcal{G}(p_k)$
• Can be used to correct for the bias
  – Exactly $R\mathcal{G}(p_k)$
  – Well enough in real-life topologies

References:
• M. Kurant, A. Markopoulou, P. Thiran, "On the bias of BFS (Breadth-First-Search) Sampling", in Proc. of ITC'22, Sept. 7-9, Amsterdam, Netherlands
• The python code can be downloaded from:
  – http://mkurant.com/maciej/publications
Outline

• Single graph: comparative study + implementation

• Multigraph sampling

• Weighted random walk

• Correcting for the bias of BFS

• From nodes to topologies
Thank you!

http://odysseas.calit2.uci.edu/wiki/doku.php

http://odysseas.calit2.uci.edu/research/osn.html