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
Our Contributions

• Compare various crawling techniques in FB’s social graph
  – Breadth-First-Search (BFS)
  – Random Walk (RW)
  – Metropolis-Hastings Random Walk (MHRW)

• Practical recommendations
  – online convergence diagnostic tests
  – proper use of multiple parallel chains
  – methods that perform better and tradeoffs

• Uniform sample of Facebook users
  – collection and analysis
  – made available to researchers
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

- 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

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

- In the stationary distribution

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

Degree of node \( u \)

Number of edges

Next candidate

Current node
(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:

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

- Re-Weighted probability distribution:

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

Subset of sampled nodes with value $i$

All sampled nodes

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

- Explore graph one node at a time with replacement

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

- In the stationary distribution

\[ \pi_v = \frac{1}{|V|} \]

Next candidate

Current node

\[ \frac{P_{AA}^{MH}}{3} = \frac{1}{3} + \frac{31}{55} = \frac{1}{5} \]

\[ \frac{5}{15} \]
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

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>MHRW</th>
<th>RW</th>
<th>BFS</th>
<th>UNI</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Valid Users</td>
<td>28×81K</td>
<td>28×81K</td>
<td>28×81K</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>

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

- Datasets available at: [http://odysseas.calit2.uci.edu/research/osn.html](http://odysseas.calit2.uci.edu/research/osn.html)
Data Collection

Basic Node Information

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

- UserID
- Name
- Networks
- Privacy settings

- UserID
- Name
- Networks
- Privacy settings

- UserID
- Name
- Networks
- Privacy settings

- UserID
- Name
- Networks
- Privacy settings

- Regional
  - School/Workplace

- 1111
  - Send Message
  - View Friends
  - Profile Photo
  - Add as Friend
Detecting Convergence

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

![Graph showing convergence](image)
Online Convergence Diagnostics

Geweke

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

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

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

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

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
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
• Degree distribution not a power law
FB Social Graph
Topological Characteristics

• Our MHRW sample
  – Assortativity Coefficient = 0.233
  – range of Clustering Coefficient [0.05, 0.35]

• [Wilson et al, Eurosyst 09]
  – Assortativity Coefficient = 0.17
  – range of Clustering Coefficient [0.05, 0.18]

• More details in our paper and technical report
Conclusion

• Compared graph crawling methods
  – MHRW, RWRW performed remarkably well
  – BFS, RW lead to substantial bias

• Practical recommendations
  – correct for bias
  – usage of online convergence diagnostics
  – proper use of multiple chains

• Datasets publicly available
  – http://odysseas.calit2.uci.edu/research/osn.html
Thank you

Questions?