Lecture 22: Sample (1/2)

“How to measure social networks?”
“Why I have less friends than my friends?”
“How to avoid bot on Facebook?”

COMS 4995-1: Introduction to Social Networks
Tuesday, December 4th
A sampling problem

How to measure a social graph?
  – What is crawling?
  – How to sample peers?
  – Why does it work?

How to avoid bots in Facebook?
Are you a popular person?

Never been so easy to figure out:
1. Take FB profile
2. Count # friends
3. Compare it with your friends

Larger data sets

- Same effect
- Known as the “chosen class” effect

How many students in a class?
How many people go to a park?
Assume that individual $i$ has $x_i$ friendship ties

What’s the mean # friend?

What’s the mean # of friends’s friend?
Outline

- A sampling problem

- How to measure a social graph?
  - What is crawling?
  - How to sample peers?
  - Why does it work?

- How to avoid bots in Facebook?
Crawler

* Initially invented to collect web-data
* Main idea: start from a point, follow connections
* In practice a crawler is usually
  1. A method to traverse a graph links,
  2. A method to retrieve contents from ID
  3. A method to parse the content into standard form

Typically, python http gets, javascript, legal issues
* A sampling problem

* How to measure a social graph?
  – What is crawling?
  – How to sample peers?
  – Why does it work?

* How to avoid bots in Facebook?
How to walk on a graph

* Breadth-First Search, Depth-First Search
* Probabilistic version:
  – BFS: *Forest Fire*
  – Every edge traversed with proba p, with restart
* Fixed scope version:
  – BFS with a fixed fanout: *Snowball sampling*
  – Lookup n neighbors from each node
What happens when you sample?

BFS on FB
- Avg deg = 324
- Real avg deg = 94

Why
1. A small fraction
2. Highest degree overrepresented!

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2010). Walking in Facebook: A case study of unbiased sampling of OSNs. INFOCOM
A random walk:
  - Look at the list of neighbors, pick one at random

It follows a Markovian evolution
  “Given present, future does not depend on past”
... so it converge to the stationary distribution!
Comparison RW BFS

Histogram (\# of peers)

(a) Erdős–Rényi
(b) Gnutella
(c) Watts–Strogatz (small world)
(d) Barabási–Albert (power-law)
Correcting the bias

* The rejection method: to sample measure \( \mu \)

\[
Q(x, y) = \begin{cases} 
  P(x, y) \min \left( \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}, 1 \right) & \text{if } x \neq y, \\
  1 - \sum_{x \neq y} Q(x, y) & \text{if } x = y
\end{cases}
\]

– equiv.: “accept the move” with proba \( \min \left( \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}, 1 \right) \)
* A sampling problem

* How to measure a social graph?
  – What is crawling?
  – How to sample peers?
  – Why does it work?

* How to avoid bots in Facebook?
First, what is the stationary distribution?

Second, how fast is it reached?
- This is denoted the mixing time: $T$ such that
  \[ T(\epsilon) = \max_i \min_t \{ t : |\pi - \pi(i) P^t|_1 < \epsilon \}, \]
- In many situations, $T(\epsilon) = O(\log n*(1/\epsilon))$ (fast mixing)
Theorem 2 (Second largest eigenvalue [21]). Let $P$ be the transition matrix of $G$ with ergodic random walk, and $\lambda_i$ for $1 \leq i \leq n$ be the eigenvalues of $P$. Then all of $\lambda_i$ are real numbers. If we label them in decreasing order, $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{n-1} \geq \lambda_n > -1$ holds. We define the second largest eigenvalue $\mu$ as $\mu = \max(\{\lambda_2, |\lambda_{n-1}|\})$. Then, the mixing time $T(\epsilon)$ is bounded by:

$$\frac{\mu}{2(1-\mu)} \log\left(\frac{1}{2\epsilon}\right) \leq T(\epsilon) \leq \frac{\log(n) + \log\left(\frac{1}{\epsilon}\right)}{1 - \mu}$$

(4)
A sampling problem

How to measure a social graph?
  – What is crawling?
  – How to sample peers?
  – Why does it work?

How to avoid bots in Facebook?
“Open a blog on my site. If you blog and receive at least 100 FB likes, I give you an iPad”.

Or … “the blog with the most vote will get an iPad”

What should you do?
The Sybil attack

* Multiplication of identities to increase weight
  - Quorum in distributed computing
  - Manipulate vote process

* Solutions:
  - Centralized certification: costly, single point of failure
  - History based
Outline

* The Sybil attack

* How to mitigate Sybils?
  - A technique using random walks
Assumes that social networks represent trust
- Weak trust ... “At least not to launch a Sybil attacks”

Assumptions:
- Attacker cannot have many attack edges
Perform a random route of size $w$
  - Random walk with local memory
  - $P[\text{routes goes through an attack edge}] < \frac{gw}{n}$

If $w = n^{0.5} \log(n)$, two routes from honest nodes
  - Intersect with large probability