Social Context And The Web
Everything's Better With Friends...

- “Hyper-presence” of friends
- “networked public spaces”
- All web activity will have social context

Mike Barash: wife just made pancakes and toast...not a bad way to start the day. and it appears we have power again. which is nice.

Griffin Barash: is day 2 ... Options baby!

Ryan van Weezel: at 2:27pm July 7
good luck... better have a red bull at lunch!

Adam Drewry: at 4:36pm July 7
that sounds like a dream of a day

Tyler Redlitz: is celebrating his bday with beautiful weather in NYC!

The Wu likes this.
## Facebook Is Becoming A Second Internet...

<table>
<thead>
<tr>
<th>Function</th>
<th>Internet version</th>
<th>Facebook version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page Markup</td>
<td>HTML, JavaScript</td>
<td>FBML</td>
</tr>
<tr>
<td>DB Queries</td>
<td>SQL</td>
<td>FBQL</td>
</tr>
<tr>
<td>Email</td>
<td>SMTP</td>
<td>FB Mail</td>
</tr>
<tr>
<td>Forums</td>
<td>Usenet, etc.</td>
<td>FB Groups</td>
</tr>
<tr>
<td>Instant Messages</td>
<td>XMPP</td>
<td>FB Chat</td>
</tr>
<tr>
<td>News Streams</td>
<td>RSS</td>
<td>FB Stream</td>
</tr>
<tr>
<td>Authentication</td>
<td>OpenID</td>
<td>FB Connect</td>
</tr>
<tr>
<td>Photo Sharing</td>
<td>Flickr, etc.</td>
<td>FB Photos</td>
</tr>
<tr>
<td>Video Sharing</td>
<td>YouTube, etc.</td>
<td>FB Video</td>
</tr>
<tr>
<td>Blogging</td>
<td>Blogger, etc.</td>
<td>FB Notes</td>
</tr>
<tr>
<td>Microblogging</td>
<td>Twitter, etc.</td>
<td>FB Status Updates</td>
</tr>
<tr>
<td>Micropayment</td>
<td>Peppercoin, etc.</td>
<td>FB Points</td>
</tr>
<tr>
<td>Event Planning</td>
<td>E-Vite</td>
<td>FB Events</td>
</tr>
<tr>
<td>Classified Ads</td>
<td>craigslist</td>
<td>FB Marketplace</td>
</tr>
</tbody>
</table>
Parallel Trend: The Internet is Becoming Social

“Given sufficient funding, all web sites expand in functionality until users can add each other as friends”
“Traditional” Social Network Analysis

- Performed by sociologists, anthropologists, etc. since the 70’s
- Use data carefully collected through interviews & observation
  - Typically < 100 nodes
  - Complete knowledge
  - Links have consistent meaning
- All of these assumptions fail badly for online social network data
Traditional Graph Theory

- Nice Proofs
- Tons of definitions
- Ignored topics:
  - Large graphs
  - Sampling
  - Uncertainty
Models Of Complex Networks From Math & Physics

Many nice models

• Erdos-Renyi
• Watts-Strogatz
• Barabasi-Albert

Social Networks properties:

• Power-law
• Small-world
• High clustering coefficient
Real social graphs are complicated!
When In Doubt, Compute!

We do know many graph algorithms:

- Find important nodes
- Identify communities
- Train classifiers
- Identify anomalous connections

Major Privacy Implications!
Privacy Questions

• What can we infer purely from link structure?
Privacy Questions

What can we infer purely from link structure?

A surprising amount!

- Popularity
- Centrality
- Introvert vs. Extrovert
- Leadership potential
Privacy Questions

• If we know nothing about a node but its neighbours, what can we infer?
Privacy Questions

• If we know nothing about a node but it's neighbours, what can we infer?
  A lot!

• Gender

• Political Beliefs

• Location

• Breed?
Privacy Questions

• Can we anonymise graphs?
Privacy Questions

• Can we anonymise graphs?

Not easily...

• Seminal result by Backstrom et al.: Attack of attack needs just 7 nodes
• Can do even better given user's complete neighborhood
• Also results for correlating users across networks
• Developing line of research...
Privacy Questions

• What can we infer if we “compromise” a fraction of nodes?
Privacy Questions

- What can we infer if we “compromise” a fraction of nodes?
  A lot...
- Common theme: small groups of nodes can see the rest
  - Danezis et al.
  - Nagaraja
  - Korolova et al.
  - Bonneau et al.
Privacy Questions

- Can we defend against crawling in a sound way?

<Work in progress!>
Privacy Questions

• What if we get a subset of neighbours for all nodes?
Privacy Questions

• What if we get a subset of $k$ neighbours for all nodes?

**Emerging question for many social graphs**

• Facebook and online SNS

• Mobile SNS
A Quietly Introduced Feature...

Public Search Listings, Sep 2007
Public Search Listings

- Unprotected against crawling
- Indexed by search engines
- Opt out—but most users don't know it exists!
Entity Resolution
Promotion via Network Effects
“Your name, network names, and profile picture thumbnail will be available in search results across the Facebook network and those limited pieces of information may be made available to third party search engines. This is primarily so your friends can find you and send a friend request.”

-Facebook Privacy Policy
Legal Status

Josh Morris
Add Josh Morris as Friend | Send Josh Morris a Message | View Josh Morris’s Friends

Here are some of Josh Morris’s friends:

Josh Morris is a fan of:

Celebrities / Public Figures
- Bruno
- “The Dude”
- Karl Marx, philosopher
- San Diego
- Iraq Veterans Against the War
- Music
- Metallica
- System of a Down
- Tool Band
- Les Pauls
- Products
- REESE’S
- Oreoos!!!!!!
- JACK DANIEL’S
- In-N-Out
- Restaurants
- In-N-Out
- Rubio’s

Much More Info Now Included...
Advocates of Communism

Global

Basic Info
Type: Common Interest - Politics
Description: The working class has nothing to lose but their chains. They have the world to win.

We have seen above that the first step in the revolution by the working class is to raise the proletariat to the position of ruling class to win the battle of democracy.

The proletariat will use its political supremacy to wrest, by degree, all capital from the bourgeoisie, to centralize all instruments of production in the hands of the state, i.e., of the proletariat organized as the ruling class, and to increase the total productive forces as rapidly as possible.

Of course, in the beginning, this cannot be effected except by means of despotic inroads on the rights of property, and on the conditions of bourgeois production; by means of measures, therefore, which appear economically insufficient and untenable, but which, in the course of the movement, outstrip themselves, necessitate further inroads upon the old social order, and are unavoidable as a means of entirely revolutionizing the mode of production.

These measures will, of course, be different in different countries.

Nevertheless, in most advanced countries, the following will be pretty generally applicable.

1. Abolition of property in land and application of all rents of land to public purposes.
2. A heavy progressive or graduated income tax.

Members
Displaying 8 of 3,513 members

Logan Akosua James Thomas Raph Austin Irvin Lahiri
Obvious Attack

• Initially returned new friend set on refresh

• Can find all $n$ friends in $O(n \cdot \log n)$ queries
  • The Coupon Collector's Problem
  • For 100 Friends, need 65 page refreshes

• As of Jan 2009, friends fixed per IP address
Fun with Tor

UK
- David Cottingham
- Eirik George
- Emma Alden
- Luke Church
- Stella Nordhagen
- David J Hornsby
- Justin Palfreyman
- Jillian Sullivan

Germany
- Shoshana Freisinger
- Lauren Duffey
- Conor Loftus-Sweetland
- Will Cordingley
- Srilakshmi Raj
- Sarita Kristina Sylvester
- Brian Brown
- Gary Champagne

USA
- Melanie Kannokada
- Shoshana Freisinger
- Russ Hedleston
- Conor Loftus-Sweetland
- Gustav Rydstedt
- Seth Ort
- Cameron Lochte
- Ben Skolnik

Australia
- Shoshana Freisinger
- Federico Baradello
- Lauren Duffey
- Adrian Boscolo-Hightower
- Justin David Carl
- Katie Gunderson
- Ankit Garg
- Srilakshmi Raj
Attack Scenario

- Spider all public listings
  - Our experiments crawled 250 k users daily
  - Implies ~800 CPU-days to recover all users
Abstraction

- Take a graph $G = \langle V,E \rangle$
- Randomly select $k$ out-edges from each node
- Result is a sampled graph $G_k = \langle V,E_k \rangle$
- Try to approximate $f(G) \approx f_{\text{approx}}(G_k)$
Approximable Functions

- Node Degree
- Dominating Set
- Betweenness Centrality
- Path Length
- Community Structure
Experimental Data

- Crawled networks for Stanford, Harvard universities
- Representative sub-networks

<table>
<thead>
<tr>
<th></th>
<th># Users</th>
<th>Mean $d$</th>
<th>Median $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>15043</td>
<td>125</td>
<td>90</td>
</tr>
<tr>
<td>Harvard</td>
<td>18273</td>
<td>116</td>
<td>76</td>
</tr>
</tbody>
</table>
• Take a graph $G = <V,E>$

• Randomly select $k$ out-edges from each node

• Result is a sampled graph $G_k = <V,E_k>$

• Try to approximate $f(G) \approx f_{\text{approx}}(G_k)$
Estimating Degrees

- Convert sampled graph into a directed graph
  - Edges originate at the node where they were seen
- Learn exact degree for nodes with degree < $k$
  - Less than $k$ out-edges
- Get random sample for nodes with degree $\geq k$
  - Many have more than $k$ in-edges
Estimating Degrees

Average Degree: 3.5
Estimating Degrees

Sampled with $k=2$
Degree known exactly for one node
Estimating Degrees

Naïve approach: Multiply in-degree by average degree / $k$
Estimating Degrees

Raise estimates which are less than $k$
Nodes with high-degree neighbors underestimated
Estimating Degrees

Iteratively scale by current estimate / k in each step
Estimating Degrees

After 1 iteration
Estimating Degrees

Normalise to estimated total degree
Estimating Degrees

Convergence after n > 10 iterations
Estimating Degrees

- Converges fast, typically after 10 iterations
- Absolute error is high—38% average
  - Reduced to 23% for nodes with $d \geq 50$
- Still accurately can pick high degree nodes
Aggregate of $x$ highest-degree nodes
Comparison of sampling parameters

![Cumulative Degree Diagram]

- Complete
- $k = 50$
- $k = 20$
- $k = 8$
- $k = 4$
- $k = 2$
- $k = 1$
- $k = 0$
Dominating Sets

- Set of Nodes $D \subseteq V$ such that
  \[ D \cup \text{Neighbours}(D) = V \]
- Set allows viewing the entire network
- Also useful for marketing, trend-setting
Dominating Sets

Trivial Algorithm: Select High-Degree Nodes in Order
In fact, finding minimal dominating set is NP-complete
Dominating Sets

Greedy Algorithm: select for maximal coverage
Dominating Sets

Greedy Algorithm: select for maximal coverage
Shown to perform adequately in practice
Works Well on Sampled Graph
Insensitive to Sampling Parameter!

Surprising: Even $k = 1$ performs quite well.
Centrality

• A measure of a node's importance

• *Betweenness centrality*: 

\[ C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \]

• Measures the shortest paths in the graph that a particular vertex is part of
Centrality

Message Interception Probability

- Maximum Centrality
- Max Centrality (k=10)
- Max Centrality (k=8)
- Max Centrality (k=5)
- Max Centrality (k=2)
- Max Centrality (k=1)
- Random Selection

Probability of Intercept

Compromised Nodes

0 200 400 600 800 1000 1200 1400 1600 1800

0.0 0.1 0.2 0.3 0.4 0.5 0.6
Community Detection

- Goal: Find highly-connected sub-groups
- Measure success by high *modularity*:

\[
Q = \frac{1}{2m} \sum_{v,w} \left[ A_{vw} - \frac{d(v)d(w)}{2m} \right]
\]

- Ratio of intra-community edges to random
- Normalised to be between -1 and 1
Community Detection

- Clausen et. al 2004 – find maximal modularity in $O(n \lg^2 n)$
- Track marginal modularity, update neighbours on each merge
Community Detection

Q=0.04
Community Detection

Q=0.08
Community Detection

\[ Q = 0.14 \]
Community Detection

Q=0.175
Community Detection

\[ Q = 0.2125 \]
Community Detection

Q = 0.2225
Community Detection

![Graph showing community detection](image-url)
Conclusions

- $k$-sampling of each edge gives away a lot
Conclusions

- $k$-sampling of each edge gives away a lot

Can we fix it?
Can we find a 2-regular subgraph?
Regular subgraph extraction

Step 1: Remove edges, weight by smallest attached node
Regular subgraph extraction

Step 1: Remove edges, weight by smallest attached node
Step 1: Remove edges, weight by smallest attached node
Regular subgraph extraction

Step 1: Remove edges, weight by smallest attached node
Regular subgraph extraction

Step 1: Remove edges, weight by smallest attached node
Regular subgraph extraction

Step 2: Remove further edges to force all degrees $\leq k$
Regular subgraph extraction

Step 3: Randomly add edges between pairs of edges below $k$
Regular subgraph extraction

Step 3: Randomly add edges between pairs of edges below $k$
Regular subgraph extraction

(note: producing a cycle is atypical!)
How well have we done?

• Recall original goal of showing $k$-sample
  • Promotion, identification

• Two measures:
  • *Precision*: Percentage of edges shown which are real
  • * Recall*: Percentage of real edges which are shown
    (normalise recall to showing a max of $k$ per node)
How well have we done?

• Recall original goal of showing \( k \)-sample
  • Promotion, identification

• Two measures:
  • *Precision*: Percentage of edges shown which are real
  • *Recall*: Percentage of real edges which are shown

  (normalise recall to showing a max of \( k \) per node)
Regular subgraph extraction

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.90</td>
</tr>
<tr>
<td>Recall</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Regular subgraph extraction
Drawbacks

• Requires complete graph knowledge
• Graph frequently changes!
Drawbacks

- Requires complete graph knowledge
- Graph frequently changes!

**Alternative: Random Sampling**

- Weight selection towards low-degree neighbours
- Computable locally, incrementally
- (much weaker...
Random Sampling

The diagram illustrates the fraction of covered edges in a graph as a function of the number of target nodes. The graph shows four different cases:

- $G_8$: Dotted line
- $G_1$: Dashed line
- $G_{\text{weight}}$: Solid line
- Random: Black line with circles

The x-axis represents the number of target nodes, ranging from 0 to 2000, while the y-axis shows the fraction of covered edges, ranging from 0 to 0.7. Each line represents a different scenario, demonstrating how the fraction of covered edges increases with the number of target nodes.
Caveats

- Can gain some protection against degree estimation
  - With a lot of work
- Doesn't prevent inference of dominating sets, centrality!
Conclusions

- Availability of social graphs raises serious privacy concern
  - The blueprint of our society...
- Very fragile to many attacks
- Right now, we're choosing utility over privacy

Thank You!  
jcb82@cl.cam.ac.uk