I. Research Question

How can we extract data from a social network on a large scale?
Our Case Study

Why Facebook is interesting:

- **Size:** 225 M users
- **Complexity**
  - Third-Party Applications
  - Public Listings
  - FB Connect
- **Accurate Profiles**
Data of Interest

- User Profiles
- Social Graph
- Traffic Data

Basic Information

Networks: Cambridge Grad Student '11
Stanford Alum '06
San Francisco, CA

Sex: Male
Birthday: July 17, 1984

Contact Information

Email: jbonneau@gmail.com
jcb82@cam.ac.uk

Mobile Number: UK 044.07590.677117
Other: US 01.650.804.6934
Skype: joseph.bonneau
Website: http://www.jbonneau.com
http://picasaweb.google.com/jbonneau
http://joecambridge.blogspot.com

Education and Work

Grad Schools:
- Stanford '07
  - Master of Science, Cryptography
  - Cambridge '11
    - PhD, Computer Science

College: Stanford '06

High School: Redwood High '02

Employer: Cryptography Research Inc.

Position: Cryptographic Scientist

Data of Interest

- User Profiles
- Social Graph
- Traffic Data
Data of Interest

- User Profiles
- Social Graph
- Traffic Data

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.
about an hour ago · Comment · Like

Griffin Barash: Is day 2 ... Options baby!
3 hours ago · Comment · Like

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

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

Tyler Redlitz: is celebrating his bday with beautiful weather in NYC!
5 hours ago · Comment · Like

The Wu: likes this.

Write a comment...
Potential Adversaries

- Advertisers
- Marketers
- Data Aggregators
  - Credit Ratings Agencies
  - Insurance Companies
- Law Enforcement
- Intelligence
- Employers
- Educators
- Online Scammers
- Research Community
What This Talk is Not

- Mechanics of large-scale parallelized web crawling
  - Largest academic crawls: \( \sim 10 \text{ M profiles} \)
II. Data Extraction Techniques

- Public Listings
- False Profiles
- Malicious Applications
- Phishing
- Facebook API
Advocates of Communism

Global

Basic Info
Type:
Description:

Common Interest - Politics
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.

Group Type
This is an open group. Anyone can join and invite others to join.

Officers

Sim
Party Philosopher

Nils
Questioner of Party Authority

Aaron
Official Representative of Anti-Revisionist Socialism

Aziz
Official Representative of U.C.Y

William
Official Representative of Moderate Trotskyist Party

Pmk
Official Representative of Communist Revolution Party

Will
Official Representative of Utopia Party

Adem
Vice Representative of the Downtrodden

Members
Displaying 8 of 3,513 members

Logan  Akosua  James  Thomas  Raph  Austin  Irvin  Lahiru
1.) Public Listings

**Search Result Content**
People who can find you in search can click through to a very limited version of your profile. Use these checkboxes to control what people can see in addition to your name.

People who can see me in search can see:

- ✓ My profile picture
- ✓ My friend list
- ✓ A link to add me as a friend
- ✓ A link to send me a message
- □ Pages I am a fan of

**Public Search Listing**
Use this setting to control whether your search result is available outside of Facebook.

- ✓ Create a public search listing for me and submit it for search engine indexing (see preview)
1.) Public Listings

- Not protected from crawling
  - Able to extract \( \sim 500 \text{k} \) per day, desktop PC
  - Extract entire network in \( \sim 500 \) machine-days

- Get only 8 links per listing

- Can still extract many useful features (Bonneau et al. 2009)
  - High Degree Nodes
  - Small Dominating Sets
  - Highly Central Nodes
  - Communities
2.) False Profiles

Freddy J. Frog

Freddy J. Frog we just planned to super awesome roadtrip now all we need is a few more if now one giant sponsor...and of course we will visit anyone who wants us to swing by

June 17 at 6:01am · Share

Freddy J. Frog

YouTube - PuppetsonScraps’s Channel
Source: www.youtube.com
Welcome JR Enterprises is proud to bring you ScrapsTV featured only on Youtube, be sure to SUSCRIBE...here you will see as the scraps puppet group comes together to bring you all forms of entertainment, ...

June 15 at 8:35pm · Share

Freddy J. Frog

Freddy J Frog Dating Video
Source: www.youtube.com
see our own freddy j frog when he was first starting out and looking for love
2.) False Profiles

- **80%** of users will befriend a frog (Krishmanurthy and Wills, 2008)
  - Can then crawl profiles with Friend-of-Friend Privacy
- **70-90%** of users viewable within a sub-network
  - Regional networks being phased out
3.) Malicious Applications

Allowing Farm Town access will let it pull your profile information, photos, your friends' info, and other content that it requires to work.

Farm Town
In the world of Farm Town you and your friends can have a great time! You can play games, design, grow and maintain your own farm and even send gifts to your friends. Play now and share the fun with everyone!

[Allow] or cancel

By proceeding, you are allowing Farm Town to access your information and you are agreeing to the Facebook Terms of Use in your use of Farm Town.
3.) Malicious Applications

- Share my name, networks, and list of friends, as well as the following information:
  - Profile picture
  - Basic info
  - Personal info (activities, interests, etc.)
  - Current location (what city I'm in)
  - Education history
  - Work history
  - Profile status
  - Wall
  - Notes
  - Groups I belong to
  - Events I'm invited to
  - Photos taken by me
  - Photos taken of me
  - Relationship status
  - Online presence
  - What type of relationship I'm looking for
  - What sex I'm interested in
  - Who I'm in a relationship with
  - Religious views

- Do not share any information about me through the Facebook API.
## 3.) Top Applications

<table>
<thead>
<tr>
<th>Application</th>
<th># Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How Well Do You Know Me?</td>
<td>28,074,528</td>
</tr>
<tr>
<td>2. Causes</td>
<td>25,508,174</td>
</tr>
<tr>
<td>3. MyCalendar</td>
<td>18,403,878</td>
</tr>
<tr>
<td>4. We’re Related</td>
<td>16,860,948</td>
</tr>
<tr>
<td>5. LivingSocial</td>
<td>16,618,043</td>
</tr>
<tr>
<td>6. Movies</td>
<td>16,128,539</td>
</tr>
<tr>
<td>7. RockYou Live</td>
<td>14,931,229</td>
</tr>
<tr>
<td>8. Texas HoldEm Poker</td>
<td>14,594,931</td>
</tr>
<tr>
<td>9. Pet Society</td>
<td>12,743,918</td>
</tr>
<tr>
<td>10. Mafia Wars</td>
<td>12,694,729</td>
</tr>
<tr>
<td>11. MindJolt Games</td>
<td>12,346,549</td>
</tr>
<tr>
<td>12. Top Friends</td>
<td>12,144,263</td>
</tr>
<tr>
<td>13. MyCalendar</td>
<td>12,128,128</td>
</tr>
<tr>
<td>14. Slide FunSpace</td>
<td>11,088,636</td>
</tr>
<tr>
<td>15. Farm Town</td>
<td>11,001,529</td>
</tr>
</tbody>
</table>

Source: InsideFacebook.com, 7/7/09
3.) Top Developers

<table>
<thead>
<tr>
<th>Application</th>
<th># Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Zynga</td>
<td>54,778,127</td>
</tr>
<tr>
<td>2. RockYou!</td>
<td>37,783,778</td>
</tr>
<tr>
<td>3. Playfish</td>
<td>33,030,872</td>
</tr>
<tr>
<td>4. How Well Do You Know Me?</td>
<td>28,074,528</td>
</tr>
<tr>
<td>5. Slide, Inc.</td>
<td>27,149,377</td>
</tr>
<tr>
<td>6. Causes</td>
<td>25,508,174</td>
</tr>
<tr>
<td>7. MyCalendar</td>
<td>18,403,878</td>
</tr>
<tr>
<td>8. LivingSocial</td>
<td>17,543,375</td>
</tr>
<tr>
<td>9. FamilyLink.com</td>
<td>17,299,316</td>
</tr>
<tr>
<td>10. Flixster</td>
<td>16,128,539</td>
</tr>
<tr>
<td>11. MindJolt</td>
<td>12,346,549</td>
</tr>
<tr>
<td>12. My Calendar</td>
<td>12,128,128</td>
</tr>
<tr>
<td>13. Slashkey</td>
<td>11,001,529</td>
</tr>
<tr>
<td>14. 6 waves</td>
<td>10,809,797</td>
</tr>
<tr>
<td>15. Zwigglers</td>
<td>10,006,859</td>
</tr>
</tbody>
</table>

Source: InsideFacebook.com, 7/7/09
## 3.) Weekly Application Churn

<table>
<thead>
<tr>
<th>Application</th>
<th># Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>MindJolt Games</td>
<td>+2,444,470</td>
</tr>
<tr>
<td>We’re Related</td>
<td>+1,291,531</td>
</tr>
<tr>
<td>Quizzer</td>
<td>+959,600</td>
</tr>
<tr>
<td>Farm Town</td>
<td>+953,428</td>
</tr>
<tr>
<td>Pet Society</td>
<td>+840,296</td>
</tr>
<tr>
<td>MyCalendar</td>
<td>+820,085</td>
</tr>
<tr>
<td>What Type Of Girl Are you?</td>
<td>+743,560</td>
</tr>
<tr>
<td>FARKLE</td>
<td>+731,537</td>
</tr>
<tr>
<td>Food Fling!</td>
<td>+713,604</td>
</tr>
<tr>
<td>Music</td>
<td>+621,588</td>
</tr>
<tr>
<td>Barn Buddy</td>
<td>+600,105</td>
</tr>
<tr>
<td>What Era Should You Time Travel To?</td>
<td>+558,301</td>
</tr>
<tr>
<td>Texas HoldEm Poker</td>
<td>+490,325</td>
</tr>
<tr>
<td>Cities I’ve Visited</td>
<td>+488,831</td>
</tr>
<tr>
<td>Waka-Waka</td>
<td>+486,538</td>
</tr>
</tbody>
</table>

Source: InsideFacebook.com, 7/7/09
4.) Profile Compromise & Phishing

Email Phishing

Joseph Bonneau (University of Cambridge)  Prying Data From a Social Network  July 20, 2009  16 / 1
4.) Profile Compromise & Phishing

Password Sharing
4.) Profile Compromise & Phishing

Facebook Connect

Facebook Connect

By proceeding, you are allowing The Run Around to access your information and you are agreeing to the Facebook Terms of Use in your use of The Run Around. By using The Run Around, you also agree to the The Run Around Terms of Service.
5.) Facebook Query Language

```
SELECT uid, name, affiliations FROM user
WHERE uid IN (X,Y, ... Z);
```

Step 1: Fetch Name/UID pairs
5.) Facebook Query Language

```
SELECT uid1, uid2 FROM friend
    WHERE uid1 IN (X,Y, ... Z)
    AND uid2 IN (U,V, ... W);
```

Step 2: Fetch Friendships
5.) Facebook Query Language

- Can query sets of \( \sim 1,000 \) users at a time
- Can fetch all Name/UID pairs in \( \sim 600 \) machine-days
- Exponential blowup in friendship queries:
  \[
  \left( \frac{N}{1,000} \right) \approx \left( \frac{200,000}{2} \right) \approx 2 \cdot 10^{10}
  \]
- Still, useful to fill in gaps from other methods
III.) Simulation

- How many nodes must be “compromised” to view a large portion of the network?
- Assume all nodes have friends-only or friend-of-friend privacy
- Test growth of node coverage and edge coverage
Data Set

- Crawled ~ 15,000 users from Stanford University
  - Used FQL method, took < 12 hours.
Experimental Results

Friends-Only

Nodes

Links

Friend-of-Friend

% Profiles Viewable

% Profiles Compromised

% Links Found

% Profiles Compromised

% Links Found

Targeted Selection
Random Selection
Friendship Requests
## Experimental Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>50% profiles</th>
<th>90% links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted compromise, friend-only</td>
<td>0.16%</td>
<td>0.14%</td>
</tr>
<tr>
<td>Random compromise, friend-only</td>
<td>0.71%</td>
<td>0.60%</td>
</tr>
<tr>
<td>Friend requests, friend-only</td>
<td>50.0%</td>
<td>19.6%</td>
</tr>
<tr>
<td>Targeted compromise, friend-of-friend</td>
<td>0.01%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Random compromise, friend-of-friend</td>
<td>0.04%</td>
<td>0.03%</td>
</tr>
<tr>
<td>Friend requests, friend-of-friend</td>
<td>0.16%</td>
<td>0.14%</td>
</tr>
</tbody>
</table>
Simulation Conclusions

- Only need to compromise a small fraction of network
  - Initial gains very fast
- Friends-of-friend makes discovery **10-20** times faster
- Targeted compromise *doesn’t* help much
- Phishing needs to be taken seriously...
General Conclusions

- Many ways to get data out of a modern SNS
- Most users unaware of these methods
- Data collection practical for many motivated parties
Questions?