

# PRYING DATA FROM A SOCIAL NETWORK

**Joseph Bonneau**

jcb82@cl.cam.ac.uk

**Jonathan Anderson**

jra40@cl.cam.ac.uk

**George Danezis**

gdane@microsoft.com



UNIVERSITY OF  
CAMBRIDGE

Computer Laboratory

ASONAM CONFERENCE

ATHENS, GREECE

JULY 20, 2009

# I. Research Question

How can we extract data from a social network on an large scale?

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  
Computer Science, Mathematics

High School: Redwood High '02

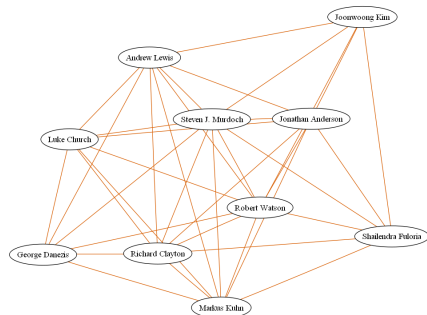
Employer: Cryptography Research Inc.

Position: Cryptographic Scientist

Time Period: April 2007 - May 2008

# 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** 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!  
5 hours ago · [Comment](#) · [Like](#)

 The Wu likes this.

# 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 M profiles**
  - See Wilson et al. User Interactions in Social Networks and their Implications. *EuroSys 2009*.



## II. Data Extraction Techniques

- Public Listings
- False Profiles
- Malicious Applications
- Phishing
- Facebook API

# 1.) Public Listings

**facebook**

☐ Remember Me

Forgot your password?

Login

Sign Up

Sign up for Facebook to connect with Josh Morris.



Not the **Josh Morris** you were looking for? Search more »

**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:



Julien Folstrom



Jay Meistrich



Seth Leslie



Cori Gale



Britt Johnson



Dejan Cikarovski



Johan Barros



Gil Stark

**Josh Morris is on Facebook.**

Sign up for Facebook to connect with Josh Morris.

Sign Up

It's free and anyone can join. Already a Member? Login to contact Josh Morris.

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  
Oreos!!!!!!  
JACK DANIEL'S  
In-N-Out

#### Restaurants

In-N-Out  
Rubio's  
Rubio's

# 1.) Public Listings

## 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



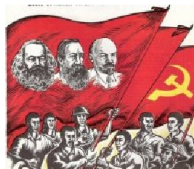
Austin



Irvin



Lahiru



### 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

# 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 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



[View Updates](#)

### Information

Affiliation:  
Scraps.TV  
Location:  
Boston, MA  
Birthday:  
July 6, 1987

### Fans

6 of 186 fans

[See All](#)



David  
Hamlyn



Tessa  
Santana



Racing  
Camel



Sammy  
Lad



Joseph  
Peachen  
cream



Michael  
Phoenix  
Alva Word

## Freddy J. Frog

Wall

Info

Photos

Boxes

Freddy J. Frog Just Fans



**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

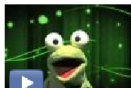
Source: [www.youtube.com](http://www.youtube.com)

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](http://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

#### **F** Allow Access?

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!

**+1 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 [What's this?](#)
  - ✓ 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 [Why can't I select this?](#)

### 3.) Top Applications

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

Source: InsideFacebook.com, 7/7/09

### 3.) Top Developers

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

Source: InsideFacebook.com, 7/7/09

### 3.) Weekly Application Churn

	<b>Application</b>	<b># Users</b>
1.	MindJolt Games	+2,444,470
2.	We're Related	+1,291,531
3.	Quizzer	+959,600
4.	Farm Town	+953,428
5.	Pet Society	+840,296
6.	MyCalendar	+820,085
7.	What Type Of Girl Are you?	+743,560
8.	FARKLE	+731,537
9.	Food Fling!	+713,604
10.	Music	+621,588
11.	Barn Buddy	+600,105
12.	What Era Should You Time Travel To?	+558,301
13.	Texas HoldEm Poker	+490,325
14.	Cities I've Visited	+488,831
15.	Waka-Waka	+486,538

Source: InsideFacebook.com, 7/7/09

## 4.) Profile Compromise & Phishing

**tagged 3 photos of you on Facebook** Inbox | X Print all Turn on highlighting

★ **Facebook** to me show details Jun 28 (8 days ago) Reply

tagged 3 photos of you in the album " ".

To see the photos, follow the link below:  
<http://www.facebook.com/n/?photo.php&pid=43006381&op=1&view=all&subj=210132&id=14801089&mid=b17f50G334d4G21c51b3G5>

Thanks,  
The Facebook Team

---

This message was intended for [jbonneau@gmail.com](mailto:jbonneau@gmail.com). Want to control which emails you receive from Facebook? Go to:  
<http://www.facebook.com/editaccount.php?notifications&md=cGhvdG9fdGFuO2Zyb209MTQ4MDEwODk7dWlkPTE0ODAxMDg5O3BpZD00MzAwNjM4MT10bz0yMTAxMzI=&mid=b17f50G334d4G21c51b3G5>  
Facebook's offices are located at 1601 S. California Ave., Palo Alto, CA 94304.

**sent you a message on Facebook...** Inbox | X facebook | X Print all Turn on highlighting

★ **Facebook** to me show details 2008-12-05 Reply

sent you a message.

Subject: Nice ass! But why you put them in the internet?

"YAYYYYY"  
<http://www.facebook.com/1.php?u=http://geocities.com/%2Frubingallegos09%2F%3Fdcbb850%3D13191be140046e6d498e1ac0d07d218c>

To reply to this message, follow the link below:  
<http://www.facebook.com/n/?inbox/readmessage.php&t=1061058890741>


---

Want to control which emails you receive from Facebook? Go to:  
<http://www.facebook.com/editaccount.php?notifications&md=bXNnO2Zyb209MTAyMzAwMjY7dD0xMDYxMDU4ODkwNzQxO3RvPTIxMDEzMg==>

## Email Phishing

## 4.) Profile Compromise & Phishing

### Invite Your Friends

 **Web Email** (Hotmail, Gmail, Yahoo, etc.)


Invite contacts from your email account.

**Your Email:**

**Password:**

[Find Your Friends](#)

We won't store your password or contact anyone without your permission.

 **Find People You Email**

[Upload Contact File](#)


Searching your email account is the fastest and most effective way to find your friends on Facebook.

**Your Email:**

**Password:**

[Find Friends](#)

We won't store your password or contact anyone without your permission.

 **Valid webmail address**

## Password Sharing

## 4.) Profile Compromise & Phishing



### Facebook Connect

## 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:

$$\binom{\frac{N}{1,000}}{2} \approx \binom{200,000}{2} \approx 2 \cdot 10^{10}$$

- Still, useful to fill in gaps from other methods

# III.) Simulation

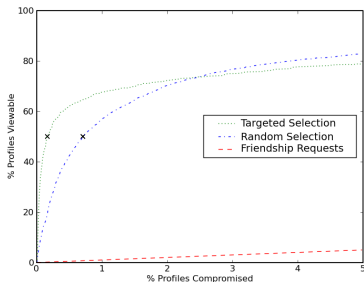
- How many nodes must be “compromised” to view a large portion of the network?
- Assume all nodes have **friends-only** or **friend-of-friend** privacy
- Test growth of **node coverage** and **edge coverage**

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

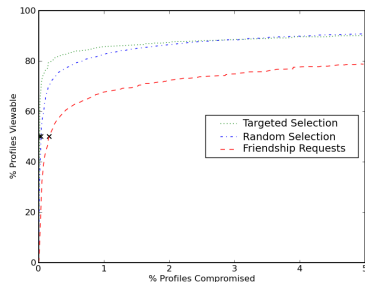
# Experimental Results

N  
o  
d  
e  
s

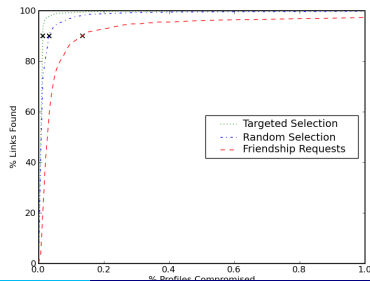
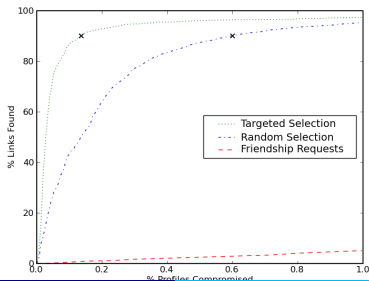
## Friends-Only



## Friend-of-Friend



L  
i  
n  
k  
s



# Experimental Results

	50% profiles	90% links
Targeted compromise, friend-only	0.16%	0.14%
Random compromise, friend-only	0.71%	0.60%
Friend requests, friend-only	50.0%	19.6%
Targeted compromise, friend-of-friend	0.01%	0.01%
Random compromise, friend-of-friend	0.04%	0.03%
Friend requests, friend-of-friend	0.16%	0.14%

# 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



# Thank You

Questions?