A Little Bit of Background

• PhD at UC Santa Barbara
  – 2010-2016 (expected)

• Intern at LinkedIn
  – Member reputation (2012)

• Intern at Microsoft Research
  – Drive-by download attack (2011)
  – Insider attack (2014)

• Strong interest in Security and Privacy
  – Security, data mining, online social networks, crowdsourcing, mobile applications
  – Home venues: USENIX Security, NDSS, DSN, IMC, WWW, CSCW, MOBICOM, SIGMETRICS
The State of Internet (In)security

• Data breaches: more often than ever
  – 690 breaches in 2015 ➔ 2.1 per day
  – 430% growth compared to 2005
  – 176 million records, could affect anyone

• Malicious content and attacks
  – Malware, phishing, spam, still problematic
  – Ransomware (encrypt user data, blackmail)
  – Internet of things: new security challenges

The next thing locking you out
Human Factors in Security

- Humans are weak links
  - 95% of all security incidents involve human factors\textsuperscript{[1]}
  - Vulnerable to social engineering, spear phishing
  - Popular targets of today’s attacks

Hi Gang,

I am a recruiter here with Amazon Data Science in Ireland. I am hoping to talk to you about a Systems Engineering role which I am hiring for at the moment.

This position is based on our data science team here in Dublin, Ireland and offers a competitive compensation plan, as well as a fantastic opportunity for continuous career growth and professional development in a challenging work environment. Having reviewed your profile, I think you could be a good match :) 

Please find at the link below some information on the role considering applying. \url{http://tinyurl.com/qxadbqf}

[1] IBM Security Services 2014 Cyber Security Intelligence Index
Questions To Be Answered

1. What are the emerging security threats on the Internet?
   - NSDI’16*
   - IMC’14
   - IMC’13

2. How to understand complex user behavior, and how to use this knowledge to benefit Internet security?
   - CHI’16a*
   - CHI’16b*
   - CSCW’15
   - USENIX Security’13
   - WWW’13
   - MobiCom’11
   - HotMobile’11

3. What’s the impact of attacks with humans in the loop?
   - USENIX Security’14
   - NDSS’13
   - WWW’12

4. How to leverage massive data analytics to build practical security solutions?
   - SIGMETRICS’13
   - DSN’13
   - TON’14

*In submission
1. Understanding User Behaviors
- User behavior modeling → detect malicious users
- Sybil detection in online social networks
- Data-driven, semi-unsupervised learning

2. Emerging Threats from Humans
- Malicious crowdsourcing = Crowdturfing
- Human intelligence to bypass security defenses
- Adversarial machine learning
Lack of Identity and Accountability

- Fake accounts in online social networks
  - 137 Million (Facebook 2014), 20 Million (Twitter 2013)
  - Spread spam and malware

- Fake identities in online financial communities
  - Fake news, misleading stock analysis (SeekingAlpha)
  - “pump and dump” scheme

- Fake (virtual) mobile devices
  - Simulate mobile devices using scripts
  - Attacks mobile apps with 10,000 (virtual) phones
    - Dominating power on Waze, YikYark, Uber, Tinder, etc.

A fundamental problem to Internet services
Sybils in Online Social Networks

- **Sybil**: fake identities in social networks
  - Multiple fake accounts controlled by a single attacker

- Key enabler of malicious attacks
  - Spam, phishing, malware
  - Click fraud, fake impressions
  - Political lobbying efforts (Donald Trump, 2015)

- Malicious URL
  - 50 likes per dollar
  - > 52% of Facebook Likes from Non-US Countries

Known black markets selling fake likes
Sybil Detection: Cat and Mouse Game

- **Graph-based system:** SybilGuard, SybilLimit, SybilInfer, Sumup
  - **Assumption:** Sybils have difficulty “friending” real users
  - Sybils form tight-knit communities

**But Sybils don’t need to form communities in reality**
- Ground-truth Sybil accounts over 6 years [IMC’11]

- **Detection during account registrations**
  - Look for suspicious IPs, bulk of registrations, etc.
  - Deliver CAPTCHA or phone verification

**But, what if crowdsourcing?**
User Behavior Defines User Identity

• **A new direction**: look at their behaviors!
  – How users browse/click social network pages

• Intuition: Sybil users act differently from normal users
  – **Goal-oriented**: concentrate on specific actions
  – **Time-limited**: fast event generation (small inter-arrival time)

• **Clickstream**: a list of server-side user-generated events
  – Click events: e.g. profile load, photo browse, friend invite
  – Build user behavior models

Analyze ground-truth clickstreams for Sybil detection
Ground-truth Dataset

- Renren Social Network
  - A large online social network in China (280M+ users)
  - Chinese Facebook

- Ground-truth
  - Ground-truth provided by Renren’s security team
  - 16K users, clickstreams over two months in 2011, 6.8M clicks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Users</th>
<th>Sessions</th>
<th>Clicks</th>
<th>Date (2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sybil</td>
<td>9,994</td>
<td>113,595</td>
<td>1,008,031</td>
<td>Feb.28-Apr.30</td>
</tr>
<tr>
<td>Normal</td>
<td>5,998</td>
<td>467,179</td>
<td>5,856,941</td>
<td>Mar.31-Apr.30</td>
</tr>
</tbody>
</table>

*Our study is IRB approved.*
Basic Analysis: Click Transitions

- Normal users use many social network features
- Sybils focus on a few actions (e.g. friend invite, browse profiles)

Sybils and normal users have very different click patterns!
Establishing Identity by Behavior Model

• Goal: quantify the differences in user behaviors
  – Measure the similarity between user clickstreams

• Approach: map user’s clickstreams to a similarity graph
  – Clickstreams are nodes
  – Edge-weights indicate the similarity of two clickstreams

• Clusters in the similarity graph capture user behaviors
  – Each cluster represents certain type of click/behavior pattern
  – Hypothesis: Sybils and normal users fall into different clusters
Model Training

① Clickstream Log

Detection

Unknown User Clickstream

? 

Good Clusters

Sybil Cluster

Model Training

Detection

Unknown User Clickstream

? 

Good Clusters

Sybil Cluster
Clickstream Similarity Functions

- **Similarity of sequences**
  - Common subsequence
    - \( S_1 = \text{AAB} \)
    - \( S_2 = \text{AAC} \)
    - \( \text{ngram}_1 = \{ \text{A, B, AA, AB, AAB} \} \)
    - \( \text{ngram}_2 = \{ \text{A, C, AA, AC, AAC} \} \)
    - \( D_{1,2} = \frac{\text{ngram}_1 \cap \text{ngram}_2}{\text{ngram}_1 | \text{ngram}_2} \)
  - Common subsequence with counts
    - \( S_1 = \text{AAB} \)
    - \( S_2 = \text{AAC} \)
    - \( \text{ngram}_1 = \{ \text{A(2), B(1), AA(1), AB(1), AAB(1)} \} \)
    - \( \text{ngram}_2 = \{ \text{A(2), C(1), AA(1), AC(1), AAC(1)} \} \)

- **Adding “time” to the sequence**
  - Bucketize inter-arrival time, encode time into the sequence
  - An example sequence with time: \( \text{A(t1)}\text{B(t2)}\text{C(t3)}\text{D(t4)}\text{A} \ldots \)

\[ \text{ngram}_1 = \{ \text{A(2), B(1), AA(1), AB(1), AAB(1)} \} \]
\[ \text{ngram}_2 = \{ \text{A(2), C(1), AA(1), AC(1), AAC(1)} \} \]

\[ V_1 = (2, 1, 0, 1, 0, 1, 1, 0) \]
\[ V_2 = (2, 0, 1, 1, 1, 0, 0, 1) \]

Details here
Detection in a Nutshell

- **Sybil detection methodology**
  - Assign the unclassified clickstream to the “nearest” cluster
  - If the nearest cluster is a Sybil cluster, then the user is a Sybil

- **Assigning clickstreams to clusters**
  - $K$ nearest neighbor (KNN)
  - Nearest cluster (NC)
  - Nearest cluster with center (NCC)

NEW CLICKSTREAMS \[ightarrow\] CLUSTERED SIMILARITY GRAPH

- Fastest, scalable
Evaluation using Ground-truth

- Split 12K clickstreams into training and testing datasets
  - Train initial clusters with 3K Sybil + 3K normal users
  - Classify remaining 6K testing clickstreams

![Error Rate Chart]

- K-nearest neighbor: False Positive < 0.7%
- Nearest Cluster: False Positive < 0.7%
- Nearest Cluster (center): False Positive < 0.7%

NCC (fastest) is as good as the others
(Semi) unsupervised Approach

- What if we don’t have a big ground-truth dataset?
  - Need a method to label clusters
- Use a (small) set of known-good users to **color** clusters
  - Adding known users to existing clusters
  - Clusters that contain good users are “good”

- 400 random good users are enough to color all behavior clusters
- For unknown dataset, add good users until diminishing returns
- Still achieve high detection accuracy (1% fp, 4% fn)
Real-world Experiments

- Deploy system prototypes onto social networks
  - Shipped our prototype code to LinkedIn renren
  - Positive feedbacks, detected previously unknown Sybils

- Key insight: force Sybils to mimic normal users
  - Slowdown click speed, generate normal clicks as cover traffic

“Image” Spammers
- Embed spam content in images
- Easy to evade text/URL based detectors

= Win
What Is Next?

Open problems

– Behavior analysis beyond Sybil detection
– Understand the meaning of behavioral models
– Capture behavior change over time

• Clickstream analysis + visualization

– Unsupervised, capture natural behavioral types
– Interactive, easy to interpret

Hierarchical Clusters

What Behavior IS This?

The first column shows the Rank of the Action Pattern, with a higher ranking means this pattern is more prevalent among users in this cluster.

This is an example Action Pattern:

```
Cluster ID: 106 | Number of Users: 45747
```

<table>
<thead>
<tr>
<th>Rank</th>
<th>Action Pattern</th>
<th>Frequency Distribution</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>View Whisper 1M View Whisper 1M View Whisper</td>
<td><img src="#" alt="Frequency Distribution" /></td>
<td>192360.58</td>
</tr>
<tr>
<td>2</td>
<td>View Whisper 1M View Whisper</td>
<td><img src="#" alt="Frequency Distribution" /></td>
<td>162735.17</td>
</tr>
<tr>
<td>3</td>
<td>View Whisper 1S View Whisper 1M View Whisper</td>
<td><img src="#" alt="Frequency Distribution" /></td>
<td>131239.25</td>
</tr>
</tbody>
</table>

Get: Read Whispers Sequentially

Example: The red distribution is more skewed to the left, indicating users in this cluster perform this action more frequently.
Talk Outline

1. Understanding User Behavior

2. Emerging Threats from Humans
   - Malicious crowdsourcing = Crowdturfing
   - Human intelligence to bypass security defense
   - Adversarial machine learning
High-quality Spam, Fake Accounts

• Review posted on Yelp
  – Detailed content
  – Even has a personal touch

•

Been B.
IN, USA

11/02/2015  Review for New Mongolian BBQ

Really great BBQ, we had such a great time. Kind of noisy when the line was long, but the food was great to wait for. Loved the way they cook the food on an open table. You can watch the food being cooked and it smells so good. Would recommend this place. They have ice cream after the meal and that is a good treat, soft ice cream, love it!
Malicious Crowdsourcing = Crowdturfing

• Malicious crowdsourcing: real users carry out attacks
  – Fake reviews, fake accounts, fake ads, rumors, etc.
  – Easy to bypass existing defense (e.g. CAPTCHA)

• Crowdturfing campaign workflow

  Customer

  Crowd workers...

  Crowdturfing site

  Target Network

  Facebook

  Campaign Likes
  From
  Real Users

  Cannot Be Detected

**Get Facebook Likes.**

<table>
<thead>
<tr>
<th>Bids</th>
<th>Avg Bid (USD)</th>
<th>Project Budget (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>$115</td>
<td>$10 - $30</td>
</tr>
</tbody>
</table>

3 days, 0 hours left

**Project Description**

We need the 500 - 700 Likes spread over 4 days averaging 150 Likes Per Day

Criteria for our Project:

1. NO Spam, Bots or Fake Accounts!
2. NO tactics that will result in suspension of our FB account/imagw.
3. The "Likes" you get us must be from real people in "India" with:
   a) 2 or more photos in their profile
   b) at least 20+ friends for each profile and,
   c) status updates that go back at least 2 weeks
   d) no admin access to our profile will be given
   e) Target cities: anywhere throughout India
A Fast Growing Market

• Measurement study on crowdturfing sites
  – Two largests sites ZhuBajie (ZBJ), SanDaHa (SDH)
  – Historical transaction records over 3 years
  – 80K campaigns, 180K workers, 7.7 million tasks

• Similar sites in US and India
  – MinuteWorkers, MyEasyTasks, Microworkers, ShortTasks
  – Poultry Markets ($20 for 1000 followers)

• Other studies confirm our results
  – Freelancer: 28% spam jobs (fake reviews, fake accounts)
  – Fiverr: a seller driven market (recently sued by Amazon)
Detecting Crowdturfing

• Machine learning (ML) to detect crowdturfing workers
  – Simple Turing tests fail on real users
  – Machine learning: sophisticated behavioral models for detection

• Focus on campaigns on Weibo (Chinese Twitter)

---

**Experiment Summary**

- Ground-truth Data from ZBJ and SDH
  - 28K workers, 317K benign users
  - 35 behavioral features

- Different machine learning classifiers
  - Decision Tree, SVM, Bayes, Random Forests

• Results: 95% - 99% accuracy
• Winners: Random Forests, Decision Tree

Not Yet ...
Adversarial Machine Learning

- **Problems:** Humans are intelligent and capable of changes
  - Motivated workers/crowdturf admins will attack ML classifiers

**Our Questions**
- What’s the impact adversarial attacks in practice?
- Which ML classifiers are more robust?
Evasions by Changing Behaviors

- Individual workers evade detection of a classifier
  - Identify a key set of behavioral features
  - Mimic normal users on these features

- Optimal evasion scenarios
  - **Per-worker optimal:** perfect knowledge
  - **Global optimal:** knows direction of the boundary
  - **Feature-aware evasion:** knows feature ranking

- Practical evasion scenario
  - Only knows normal users statistics
  - Estimate which of their features are most “abnormal”
Evasion Attack Results

- Highly effective with **perfect** knowledge, less effective in practice
- Most classifiers are vulnerable to evasion
  - Random Forests are slightly more robust (Decision Tree the worst)
Poisoning Attacks

- Temper with training data, manipulate classifier training
  - E.g., crowdturfing admins publish false records on their websites
  - Injecting benign accounts as “workers” into training data

- No single classifier is robust against all attacks
- More accurate classifier are more vulnerable (Decision Tree)

Decision Tree is the most vulnerable model

10% of poison samples ➞ boost false positives by 5%
Discussion

• Identified an emerging threat: **crowdturfing**
  – Growing exponentially in size and revenue
  – $1 million per month on just one site

• Huge problem for existing security systems
  – Little to no automation to detect
  – Turing tests fail

• Machine learning as defense
  – Effective on current workers, but vulnerable to adversarial attacks
  – **Happening now**: worker training for evasion, reverse-engineer behavioral thresholds
Summary

• Online communities are key battleground for spam, phishing, malware, and opinion manipulation
  – Cat and mouse game in attacks and defenses
  – A deep understanding on user behavior helps

• Attacks with humans in the loop
  – Strong adversaries to existing security mechanisms
  – Security systems must improve to handle human factors

• Big data analytics and measurement
  – Provide new insights to emerging threats
  – Data-driven security systems: scalable, robust, usable
Thank You!

http://www.cs.ucsb.edu/~gangw/