Consequences of Compromise: Characterizing Account Hijacking on Twitter

Frank Li
UC Berkeley

With: Kurt Thomas (UCB → Google), Chris Grier (UCB/ICSI → Databricks), Vern Paxson (UCB/ICSI)
Accounts on Social Networks

• Accounts are valuable!
  – Precursor for abuse (spam, phishing, malware)
  – Twitter accounts are attractive
Accounts on Social Networks

• Accounts are valuable!
  – Precursor for abuse (spam, phishing, malware)
  – Twitter accounts are attractive

• Two ways for attackers to get accounts:
  – Fraudulent accounts
  – Compromised accounts
Prior Works

- Fraudulent accounts
  - Lots of prior work on detecting and preventing fake accounts
- Compromise accounts
  - COMPA (NDSS '13)
  - PCA-based Anomaly Detection (USENIX Security '14)
Compromise on Social Networks

• Is compromise occurring at large scales?
• What do miscreants do with compromised accounts?
• Who are being victimized?
• How do users react to compromise?
• What is causing compromise?
Detecting Compromise

• We take an external perspective of Twitter

• Looked at 8.7B tweets with URLs gathered from Jan – Oct 2013
  – 168M users in data set
Spam Tweets

AwesomEEE! I made $171.50 this week so far taking a couple of surveys.
http://t.co/cwG67lh4

10:20 AM - 19 Nov 13

Awesome! I made $106.03 this week so far just filling out a couple of surveys.
http://t.co/PoHBayLz

8:44 AM - 5 Dec 12
SEMrush @semrush · Oct 4
Hahaha! I didn’t know Harry spoke Python :D  #fun #itjokes #python
Analysis Pipeline

1. Data Collection
   - Twitter Stream
   - Tweets (HDFS)

2. Cluster Identification
   - URL Clustering
   - Minhash Clustering
   - User Clustering
   - Clusters (HDFS)

3. Classification
   - User Status
   - Tweet Status
   - Training
   - Labeling

4. Graph Crawling
   - User Followers
   - Graph (HDFS)

5. User Labeling
   - Label Hierarchy
   - User Labels (HDFS)
Identifying Compromised Users

1. Data Collection
   - Twitter Stream
   - Tweets (HDFS)

2. Cluster Identification
   - URL Clustering
   - Minhash Clustering
   - User Clustering
   - Clusters (HDFS)

3. Classification
   - User Tweet
Identifying Compromised Users

1. Data Collection
   - Twitter Stream
   - Tweets (HDFS)

2. Cluster Identification
   - URL Clustering
   - Minhash Clustering
   - User Clustering
   - Clusters (HDFS)

3. Classification
   - User
   - Tweet
   - Label
   - Training
White people ...
Twitter Stream Data

- **created_at** (UTC, seconds)
- **id** (>53 bits)
- **text** (UTF-8, <140 char)
- **source**
- **lang** (machine-detected, BPC-47)
- **in_reply_to_status_id**
- **in_reply_to_user_id**
- **in_reply_to_screen_name**
- **entities**
  - hashtags
  - **urls** (both URL and domain)
  - **user_mentions**
- **user**
  - **id** (>53 bits)
  - **name** (<=20 char)
  - **screen_name** (<=15 char)
  - **description** (<=160 char)
  - **protected**
  - **verified**
  - **followers_count**
  - **friends_count**
  - **statuses_count**
  - **created_at** (UTC, seconds)
  - **lang** (user self-declared, BPC-47)
Infrastructure
Infrastructure

Tweets from Twitter Stream

EC2
Infrastructure

Tweets from Twitter Stream

Upload to S3

Download to our cluster

EC2

AMAZON S3

HDFS

Hadoop

Spark
Filtered Stream

- Access to a filtered stream of URLs
- ~200 GB of data per day, compressed to ~20 GB per day
- In total, 4.1 TB of compressed data for 2013.
Data Collection
Infrastructure Issues

- Twitter feed outage
- EC2 reboot
- EC2 feed application crash
- Low disk space
- Disk failures
- Updates break things
Filtered Stream

Roughly 61% of all Tweets with URLs
Sampling Error

- Under-estimate size of clusters

- Any graph analysis will under-represent social connectivity
Identifying Compromised Users

1. Data Collection
   - Twitter Stream
   - Tweets (HDFS)

2. Cluster Identification
   - URL Clustering
   - Minhash Clustering
   - User Clustering
   - Clusters (HDFS)

3. Classification
Awesomеее! I made $171.50 this week so far taking a couple of surveys. http://t.co/cwG67lh4

Awesome! I made $106.03 this week so far just filling out a couple of surveys. http://t.co/PoHBayLz

Near duplicate text
Different URL
Clustering Tweets

• Cluster on same URLs

• Cluster on similar content
  – Split text into n-grams
  – Want Jaccard similarity coefficient:
    \[ J(M_i, M_j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j|} \]
  – To avoid \(O(n^2)\), where \(n = O(\text{billion})\), use minhash estimation
Minhash Estimation

- Set \( A = \{a_1, \ldots, a_N\} \)  \quad \text{Set} \ B = \{b_1, \ldots, b_N\}
Minhash Estimation

- Set $A = \{a_1, \ldots, a_N\}$
- Set $B = \{b_1, \ldots, b_N\}$
- Hash all elements:
  - $A' = \{h(a_1), \ldots, h(a_N)\}$
  - $B' = \{h(b_1), \ldots, h(b_N)\}$
Minhash Estimation

- Set $A = \{a_1, \ldots, a_N\}$  Set $B = \{b_1, \ldots, b_N\}$
- Hash all elements:
  $A' = \{h(a_1), \ldots, h(a_N)\}$  $B' = \{h(b_1), \ldots, h(b_N)\}$
- Sort hashes for each set:
  $A'' = \{h(a_3), h(a_7), \ldots\}$  $B'' = \{h(b_9), h(b_2), \ldots\}$
Minhash Estimation

- Set $A = \{a_1, \ldots, a_N\}$  \hspace{1cm}  Set $B = \{b_1, \ldots, b_N\}$

- Hash all elements:
  
  $A' = \{h(a_1), \ldots, h(a_N)\}$  \hspace{1cm}  $B' = \{h(b_1), \ldots, h(b_N)\}$

- Sort hashes for each set:

  $A'' = \{h(a_3), h(a_7), \ldots\}$  \hspace{1cm}  $B'' = \{h(b_9), h(b_2), \ldots\}$

- Key for each set is the $k$ smallest hashes:

  $\text{Key}_A = h(a_3)||h(a_7)$  \hspace{1cm}  $\text{Key}_B = h(b_9)||h(b_2)$
Minhash Estimation

- Set $A = \{a_1, \ldots, a_N\}$  \hspace{1cm} Set $B = \{b_1, \ldots, b_N\}$
- Hash all elements:
  
  $A' = \{h(a_1), \ldots, h(a_N)\}$  \hspace{1cm} $B' = \{h(b_1), \ldots, h(b_N)\}$
- Sort hashes for each set:
  
  $A'' = \{h(a_3), h(a_7), \ldots\}$  \hspace{1cm} $B'' = \{h(b_9), h(b_2), \ldots\}$
- Key for each set is the $k$ smallest hashes:
  
  Key$_A = h(a_3)||h(a_7)$  \hspace{1cm} Key$_B = h(b_9)||h(b_2)$
- The probability keys are equal for two sets is proportional to their Jaccard similarity.
Minhash Parameters

Grid search on sample of 19 M tweets
Classifying a Group of Tweets
Classifying a Group of Tweets

- Observation 1: Users delete tweets from compromise.
Classifying a Group of Tweets

- Observation 1: Users delete tweets from compromise.
- Observation 2: Twitter suspends fraudulent accounts.
Classifying a Group of Tweets

- Observation 1: Users delete tweets from compromise.
- Observation 2: Twitter suspends fraudulent accounts.
Deletions and Suspensions as Features

- Manually labeled 1700 random clusters
Deletions and Suspensions as Features

- Manually labeled 1700 random clusters
Other Features

- Fraction of tweets in a cluster that were retweets
- Average # of tweets per user in the cluster
- # of distinct tweet sources per cluster
- # of distinct languages per cluster
Classification

- Multi-class logistic regression
- 10-fold cross-validation: 99.4% accuracy
- Most important features:
  - Ratio of suspended users, ratio of deleted tweets, number of distinct languages
Identifying Compromised Users
Identifying Compromised Users
Analyzing Compromised Users

1. Collection
2. Cluster Identification
3. Classification
4. Graph
Analyzing Compromised Users
Analyzing Compromised Users

3 Classification

4 Graph Crawling

5 User Labeling
Scale of Compromise
<table>
<thead>
<tr>
<th>Measurement</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meme clusters</td>
<td>10,792</td>
</tr>
<tr>
<td><strong>Compromise clusters</strong></td>
<td>2,661</td>
</tr>
<tr>
<td>Fraudulent account clusters</td>
<td>2,753</td>
</tr>
<tr>
<td>Meme participants</td>
<td>17.3 million</td>
</tr>
<tr>
<td><strong>Compromised victims</strong></td>
<td>13.9 million</td>
</tr>
<tr>
<td>Fraudulent accounts</td>
<td>4.7 million</td>
</tr>
<tr>
<td>Meme tweets</td>
<td>130 million</td>
</tr>
<tr>
<td><strong>Spam tweets via compromised accounts</strong></td>
<td>81 million</td>
</tr>
<tr>
<td>Spam tweets via fraudulent accounts</td>
<td>44 million</td>
</tr>
</tbody>
</table>
Monetizing Compromised Accounts
Monetizing Compromised Accounts

• Largest single campaign advertised Garcinia
  – 1.1M accounts
  – 70k distinct URLs
  – Lasted 23 days

• Nutraceutical campaigns were largest source
  – 4.7M accounts total (34% of all we detect)

_sid bishop_ @sustainablesid · 7h
Dr. Oz **Garcinia** Cambogia Where To Buy Natural And Organic Food That Burns ... - Amersham People tinyurl.com/lq9wa5l
Other Leading Monetization Vectors

- **Gain followers and retweets**
  - 3.7M users
  - 779 distinct clusters advertising free followers

- **Generating Leads**
  - 1M users, 1 cluster, lasting 31 days

---

**benny blanco** @bennyblanco523 · Mar 21
AwesomEEE! I earned $102.46 this week just doing a couple of surveys.
apps.facebook.com/162827083864702
Compromise Demographics
Compromise Demographics

(a) account age [days]
(b) followers
(c) followings
(d) tweets

CDF

- compromise
- fraudulent
- meme
- random
Accounts After Compromise
Accounts After Compromise

CDF

Change in followers

-200 -100 0 100 200

compromise random
Accounts After Compromise

57% of compromise victims lost followers!
Accounts After Compromise

CDF

Days since last tweet

compromise random
Accounts After Compromise

21% of compromised victims no longer tweet!
Sources of Compromise

• Potential sources
  – Password brute-force
  – Database dumps
  – Social contagion (i.e. spread via your friends)
  – External contagion (i.e. driveby download site)
Compromise Can Spread

- Probability of adoption
- Number of $k$ influencing neighbors

**Label**
- Black: compromise
- Red: meme
Compromise Can Spread

Observed >100X Increase in Rate of Compromise

Probability of adoption

Number of $k$ influencing neighbors

label

- black: compromise
- red: meme
Sources of Compromise

- Potential sources
  - Password brute-force
  - Database dumps
  - Social contagion (i.e. spread via your friends)
  - External contagion (i.e. driveby download site)
Sources of Compromise

• Potential sources
  – Password brute-force
  – Database dumps
  – Social contagion (i.e. spread via your friends)
  – External contagion (i.e. driveby download site)

• Defense: Early victims are indicators. If spread is on Twitter, quarantining can help.
Summary
Summary

- Is compromise occurring at a large scale?

  YES! 14 million victims!
Summary

• Is compromise occurring at a large scale?

   YES! 14 million victims!

• What do miscreants do with compromised accounts?

   $$$ Profit! $$$
Summary

• Is compromise occurring at a large scale?
  YES! 14 million victims!

• What do miscreants do with compromised accounts?
  $$$ Profit! $$$

• How do users react to compromise?
  Bad! 21% of victims quit, 57% lost followers
Summary

• Is compromise occurring at a large scale?
  YES! 14 million victims!

• What do miscreants do with compromised accounts?
  $$$ Profit! $$$

• How do users react to compromise?
  Bad! 21% of victims quit, 57% lost followers

• How might compromise be occurring?
  Highly potent social contagions
I has a question...

frankli@cs.berkeley.edu