CS 598: Machine Learning for Sys, Networks, and Security

Gang Wang
University of Illinois Urbana-Champaign
Fall 2020
ML for defense (spam, phishing)

- You are How You Click: Clickstream Analysis for Sybil Detection USENIX Security 2013
- Detecting Credential Spearphishing in Enterprise Settings USENIX Security 2017 (Mark Cockburn)
Example Paper Review

- https://gangw.cs.illinois.edu/class/cs598/sample-review.txt
Bots, Fake Identities, Misinformation

- **Sybil** (*stɪbəl*): multiple identities controlled by single attacker
- Spread misinformation, manipulate public opinion

**Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate**

**Europe**

Senate Finds Russian Bots, Bucks Helped Push Brexit Vote Through

January 19, 2019 • 7:01 AM ET

Heard on Weekend Edition Saturday
More Sophisticated Bots

Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International Studies (CSIS) · University of Michigan College of Literature, Science...
Washington · 49 connections

Experts: Spy used AI-generated face to connect with targets
By RAPHAEL SATTER June 13, 2019
Manipulating Stock Markets

300% price diff. from promotion
The Problem

• In large social communities, how to classify bots from normal users?

• Existing approaches:
  – Graph based method: assuming fake accounts friend with each other
  – Behavior based approach: supervised learning, require large set of labeled data, cannot handle new attacks

• Key idea of this paper:
  – Reduce the assumption on specific attacker behaviors
  – Semi-supervised method, demanding less labeled data
User Behavior Defines User Identity

• Clickstream based Sybil detection
  – Look at how users browse/click social network pages

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

- Renren Social Network
  - Large online social network in China (220M+ users)
  - Chinese equivalent of Facebook

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

<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>
Different Behaviors: Normal vs. Sybils

Normal users use many social network features
Sybils focus on a few actions (e.g. friend invite, browse profiles)
Methodology: Building Behavior Model

• Goal: quantify the differences in user behaviors
  – Measure the similarity between user clickstreams
  – Clickstream similarity: \( F(u_x, u_y) \in [0,1] \)

• Approach: map user’s clickstreams to a similarity graph
  – Clickstreams are nodes
  – Edge weighted by the similarity of 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
    
    \[
    S_1 = \text{AAB} \quad S_2 = \text{AAC}
    \]
    
    \[
    \text{ngram}_1 = \{\text{A, B, AA, AB, AAB}\} \quad \text{ngram}_2 = \{\text{A, C, AA, AC, AAC}\}
    \]
    
    \[
    D_{1,2} = \frac{\text{ngram}_1 \& \text{ngram}_2}{\text{ngram}_1 | \text{ngram}_2}
    \]

  – Common subsequence with counts
    
    \[
    S_1 = \text{AAB} \quad S_2 = \text{AAC}
    \]
    
    \[
    \text{ngram}_1 = \{\text{A(2), B(1), AA(1), AB(1), AAB(1)}\} \quad \text{ngram}_2 = \{\text{A(2), C(1), AA(1), AC(1), AAC(1)}\}
    \]
    
• Adding “time” to the sequence
  – Bucketize inter-arrival time, encode time into the sequence
  – An example sequence with time: \(A(t_1)B(t_2)C(t_3)D(t_4)A\ ...\)
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)
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

<table>
<thead>
<tr>
<th>Detection Algorithm</th>
<th>False Positive Rate</th>
<th>False Negative Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-nearest neighbor</td>
<td>&lt; 0.7%</td>
<td></td>
</tr>
<tr>
<td>Nearest Cluster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nearest Cluster (center)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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)
Discussion Questions

• How can attackers evade the detection system?