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

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

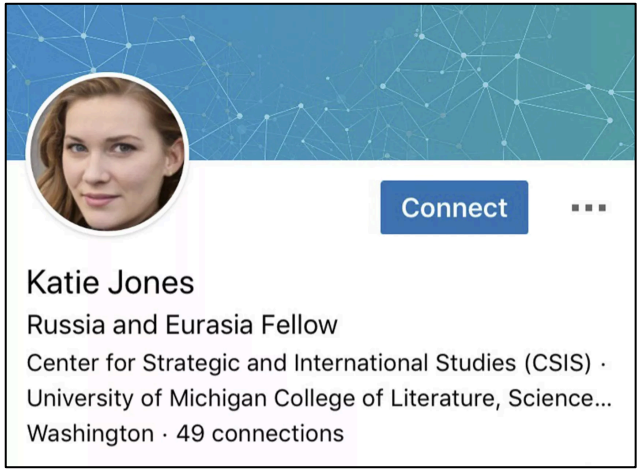
EUROPE

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

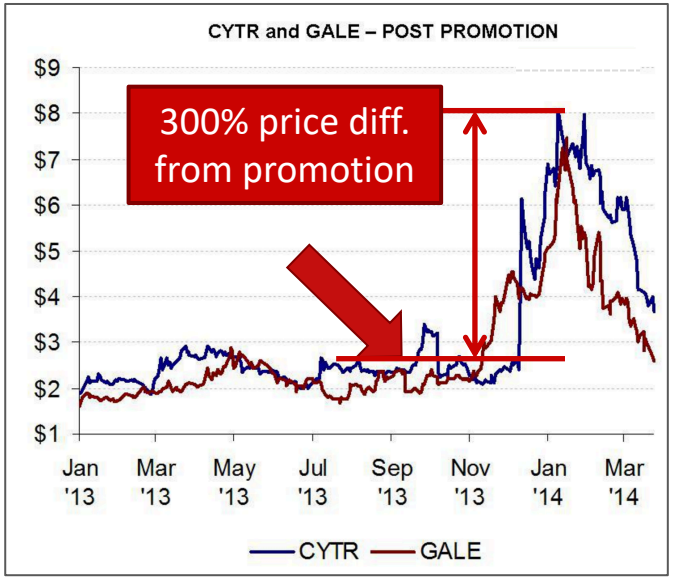


Experts: Spy used AI-generated face to connect with targets

By RAPHAEL SATTER June 13, 2019

Manipulating Stock Markets

Seeking Alpha α



The Problem

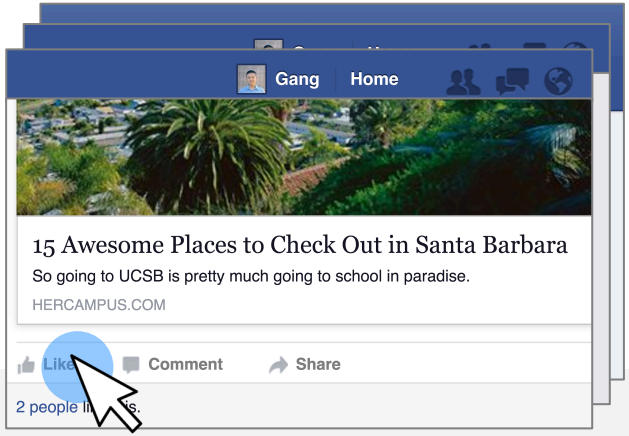
- In large social communities, how to classify bots from normal users?

- 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

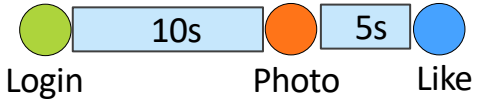
- Key idea of this paper:
  - Reduce the assumption on specific attacker behaviors
  - Semi-supervised method, demanding less labeled data

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)



Clickstream:
A sequence of click events and time gaps



Ground-truth Dataset

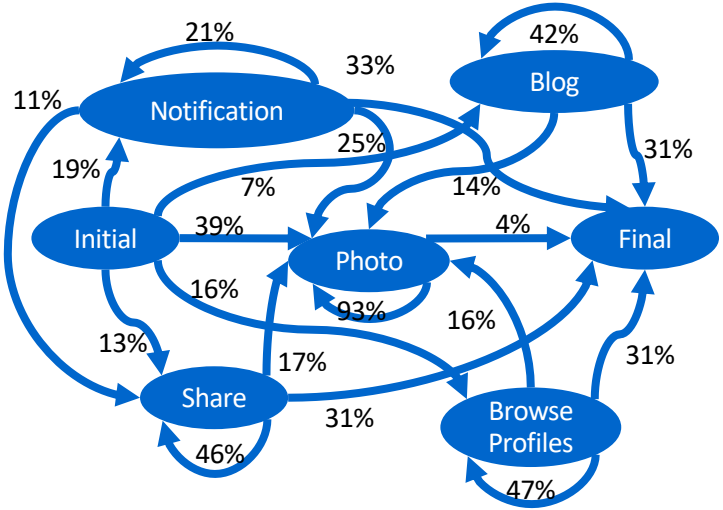
- Renren Social Network
  - Large online social network in China (220M+ users)
  - Chinese equivalent of Facebook

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

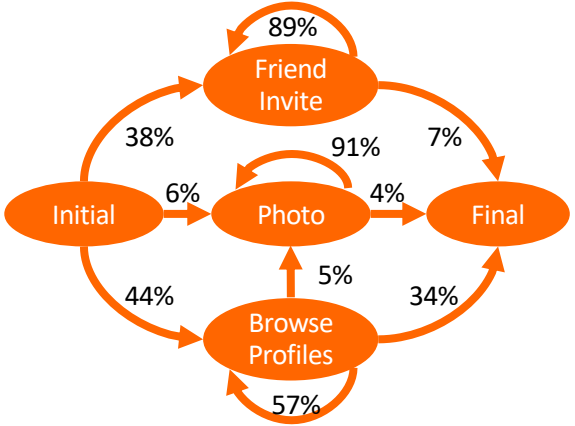
Dataset	Users	Sessions	Clicks	Date (2011)
Sybil	9,994	113,595	1,008,031	Feb.28-Apr.30
Normal	5,998	467,179	5,856,941	Mar.31-Apr.30

Different Behaviors: Normal vs. Sybils

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



Normal Clickstream

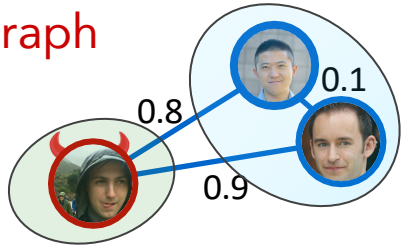


Sybil Clickstream

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



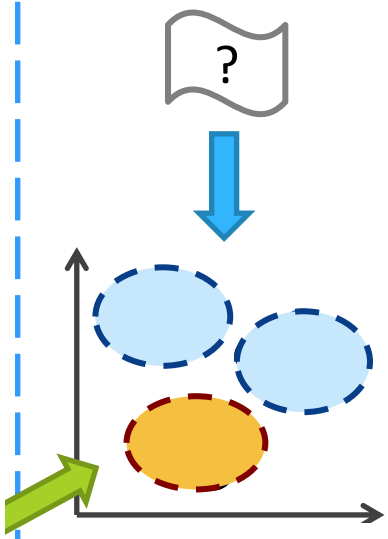
- Clusters in the similarity graph capture user behaviors
  - Each cluster represents certain type of click/behavior pattern
  - Hypothesis: Sybils and normal users fall into different clusters

Model Training

① Clickstream Log

Detection

Unknown
User Clickstream



Clickstream Similarity Functions

- **Similarity of sequences**
  - Common subsequence

S_1= AAB
S_2= AAC

→

$ngram_1$= {A, B, AA, AB, AAB}
$ngram_2$= {A, C, AA, AC, AAC}

→

$$D_{1,2} = \frac{ngram_1 \,\&\, ngram_2}{ngram_1 \mid ngram_2}$$

  - Common subsequence with counts

S_1= AAB
S_2= AAC

→

ngram1= {A(2), B(1), AA(1), AB(1), AAB(1)}
ngram2= {A(2), C(1), AA(1), AC(1), AAC(1)}

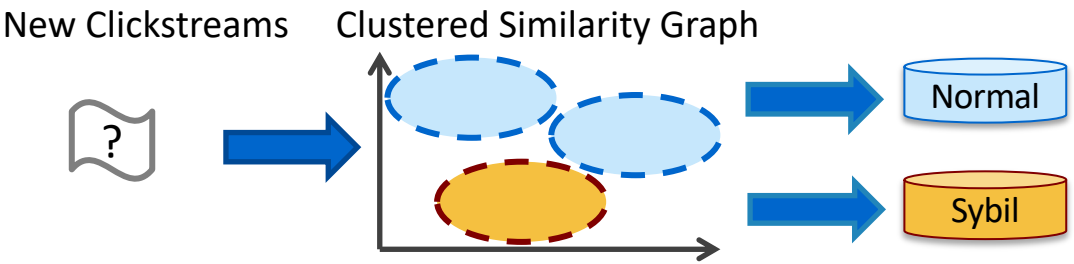
→

Euclidean Distance

V_1=(2,1,0,1,0,1,1,0)
V_2=(2,0,1,1,1,0,0,1)

- **Adding "time" to the sequence**
  - Bucketize inter-arrival time, encode time into the sequence
  - An example sequence with time: $A(t_1)B(t_2)C(t_3)D(t_4)A$ …

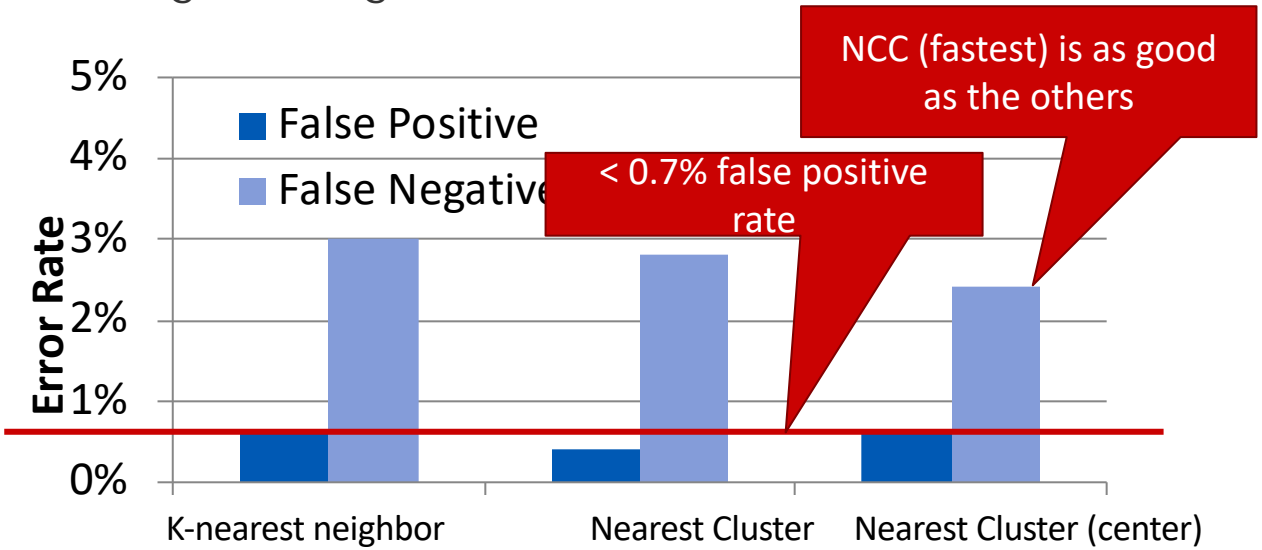
Detection in a Nutshell



- 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)

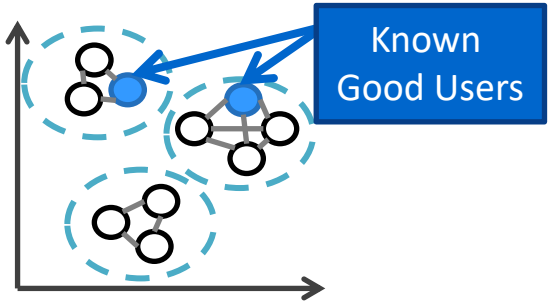
Evaluation Using Ground-truth

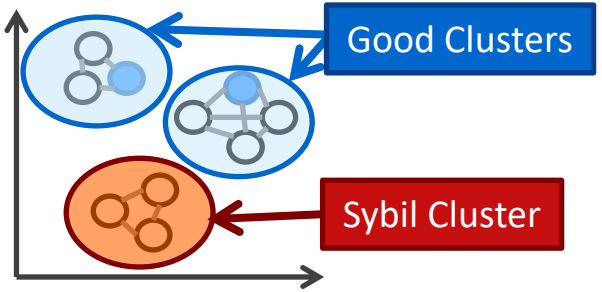
- Split 12K clickstreams into training and testing datasets
  - Train initial clusters with 3K Sybil + 3K normal users
  - Classify remaining 6K testing clickstreams



(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?