COMPA: Detecting Compromised Accounts on Social Networks

Manuel Egele, Gianluca Stringhini,
Christopher Kruegel, Giovanni Vigna
megele@cmu.edu,
{gianluca, chris, vigna}@cs.ucsb.edu
Carnegie Mellon University & UC Santa Barbara

Recently on Twitter …

Why Compromised Accounts?

- Historically, attackers created fake accounts
  - Detection mechanisms proposed
  - Detection implemented by OSNs
  - Identified fake accounts can simply be removed
- Attackers compromise legitimate accounts
  - Leverage existing trust relationships
  - Fake account detection not applicable
  - Cannot be removed easily
  - Involves costly password-reset process

COMPA: Overview

Detect compromised accounts by observing change in behavior

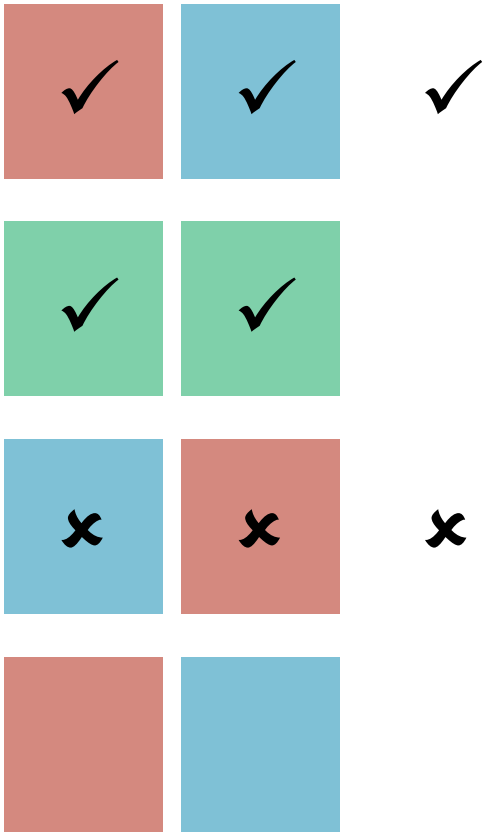
- Statistical modeling
  - Extract behavioral profile for accounts
- Anomaly detection
  - Compare new messages against observed behavior

- Legitimate changes might seem anomalous
  - Identify campaigns by grouping similar messages and look for similar compromises

COMPA: Overview

Step 1: Group similar messages

Step 2: Match messages with behavioral profile



Statistical Modeling

- Behavioral profile: collection of statistical models
- Build statistical models of features to model normal behavior
- Features:
  - Direct User Interaction
  - Message Topic
  - Links in Messages
  - Message Text (language)
  - Time (hour of day)
  - Message Source (application)
  - User Proximity

Statistical Models

- Input: Message stream
  (e.g., Twitter timeline, Facebook posts)
- Extract features for each message
- Train model for each feature
- Model M set of tuples <f_v, c>
  - M_{lang} {<English, 5>,<German, 3>}
- A behavioral profile is a collection of models
- Evaluate new messages by comparing feature values against trained models

Evaluating New Messages (cont.)

- How to compare individual anomaly scores against a behavioral profile?
- Anomaly score: weighted sum of model values
- If anomaly score exceeds threshold → message violates the behavioral profile
- Weights & threshold determined through Weka's SMO on labeled training dataset

Case Study

- July 4th 2011, @foxnewspolitics

BREAKING NEWS: President @BarackObama assassinated, 2 gunshot wounds have proved too much. It's a sad 4th for #america. #obamadead RIP

- Anomaly scores:
  - Time: 1.00 (1:24am EST, usually 8-10am EST)
  - Source: 0.94 (Web, commonly using TweetDeck)
  - Hashtag: 0.88
  - Domain: 0.26
  - Mention: 0.67
  - Lang: 0.00

Detecting Campaigns

- Single profile violation might be due to legitimate change of behavior
- Multiple accounts experience similar violating changes → Campaign
- How to define similarity:
  - Content similarity
  - Similar landing pages

Detecting Similar Messages

- Content similarity
  - Consider two messages similar if they share a common n-gram (e.g., 4-words)
  - Filter template messages, e.g., Foursquare and Nike+
- Link similarity
  - Consider two messages similar if they share a common link or landing-page

Evaluation: Data Sources

- 10% of public Twitter activity (1.4 billion tweets)
  - Individual tweets
  - No direct messages, no protected profile tweets
  - May 13, 2011 – Aug 12, 2011
- 20,000 REST-API requests to Twitter / hour
  - To retrieve message stream (timeline)
  - Max 200 tweets/request
- 106 million Facebook posts
  - Five geographical networks from 2009 (London, NY, LA, Monterey Bay, Santa Barbara)

Evaluation

- Every hour
  - Group similar messages
  - Build behavioral profiles for accounts in groups
  - Compare messages against behavioral profiles
  - If many profiles are violated detect compromise
  - 500,000 distinct users / hour

Evaluation

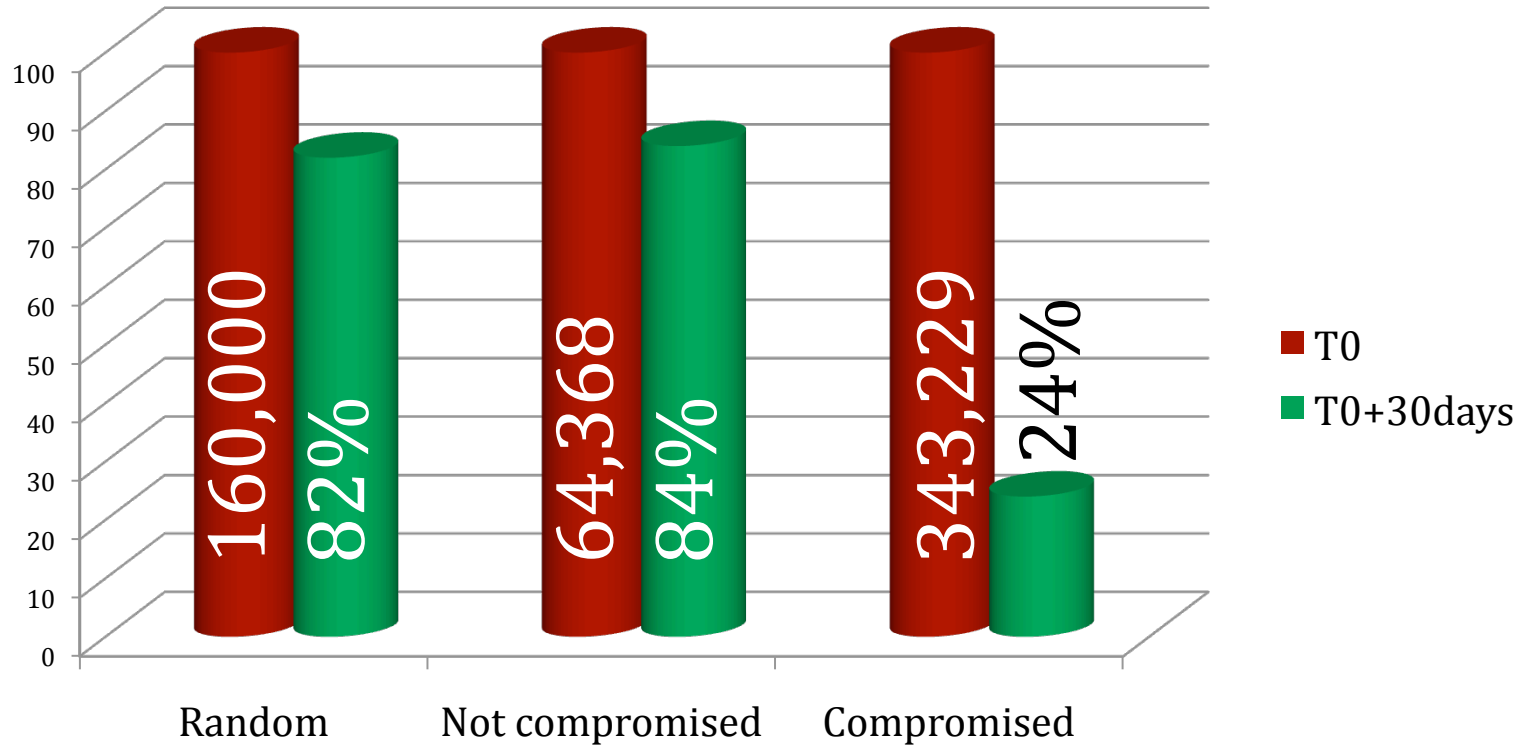
- Text similarity:
  - 374,920 groups identified
  - 9,362 compromised (343,229 accounts)
  - FP: 377 groups (4%), 12,382 accounts (3.6%)
- Landing page similarity:
  - 14,548 groups identified
  - 1,236 compromised (54,907 accounts)
  - FP: 72 groups (5.8%), 2,141 accounts (3.8%)
- Facebook:
  - 48,586 groups identified
  - 671 compromised (11,499 accounts)
  - FP: 22 groups (3.3%), 412 accounts (3.6%)

Case Studies

- Spam is not exclusively using URLs
  Obama is giving FREE Gas Cards Worth $250! Call now-> 1 888-858-5783 (US Only)@@@
- Similar spam applications are used
  [Add Seguidores] 31/03/11
  [Add Seguidores] 01-04
- Similar messages linking to four different "Get More Follower" sites
  - They use the same backend i.e., one cannot sign up at two of the services simultaneously

Message Persistence

- Legitimate tweets are persistent (16% churn)
- Violating tweets are deleted (76% churn)



Evaluation: XSS Worm

http://google.com/@"onmouseover='alert(1)'"

- Choose tweet (t_0) and user (u_0) at random
- Worm propagates iff B follows A and B was active when A posted the worm message
  - User is active if posted +/- 5 minutes using web client
- Worm propagates recursively (e.g., to active friends of A, their active friends, etc.)
- Replace the messages used to determine "active" with worm message
- Compa detects the worm outbreak after 20 minutes or 2,256 infections
- Conservative propagation strategy, real worms spread to up to 40,000 accounts in 10 minutes.

Summary

- Attackers compromise accounts
  - Leverage established trust relationships
  - Cannot easily be removed by OSN
- Build behavioral profiles for accounts
- Compare new messages against profiles
- Group compromised accounts
  - Detect campaigns
- Evaluated on 1.4B tweets and 106M Facebook messages


Questions?

END

Evaluating New Messages

- Extract features from new message
- Compare features with Models
  - Each model returns anomaly score from [0,1]
  - M_{lang} {<English, 5>,<German, 3>}
  - New message is: English, German, or other (e.g., Italian)