Bot Detection in Social Media: Networks, Behavior, and Evaluation

Fred Morstatter†, Kathleen M. Carley‡, Huan Liu†

†Arizona State University
Tempe, Arizona, USA

‡Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
Tutorial Web Page

• All materials and resources are available online:

# Overview of Today’s Tutorial

- **Introduction**  15 minutes  
- **Network Methods**  40 minutes  
- **Ground Truth Discovery**  20 minutes  
- **Machine Learning Methods**  20 minutes  
- **Real World Example**  10 minutes  
- **Conclusion**  15 minutes
Introduction
Social Media

- Social media connects people.
- It allows people to share and interact anywhere at any time.
- Social media’s explosive growth has brought it to the center of our lives.
Social Media Landscape
• Social media data is big data.
• Twitter is prominent for researchers.
• 500 million tweets/day.
• 100 million users/day.
• Arab Spring, Natural Disasters, etc.
• Twitter shares some of its data.
Bot Definition

• A hijacked or adversary-owned account controlled by software. [22]
  – Compromised accounts.
  – Accounts created automatically.
• Bots can spread spam.
• Bots can be Sybils.
• Cyborgs:
  – Innocuous: Bot-assisted humans.
  – Malicious: Human-assisted bots.
Bots

• Innocuous:

  ![Bia Ben](image1)
  ![RRC Weather Bot](image2)
  ![LA QuakeBot](image3)

• Malicious:

  – Large-scale bot “armies” with simple accounts
  – Few, well-curated accounts
Why Bots are a Problem

• **At least** 7% of Twitter accounts are fake
• Different claims:
  – 50% of accounts created in 2014 have been suspended.*
  – Twitter claims 5% of 230M users are bots.
  – One study found 20M bot accounts = 9%. **
• 24% of all tweets are generated by bots ***
• 5-11% of Facebook accounts are fake ****
• Automated accounts influence discussion, spread disinformation

** [http://www.nbcnews.com/technology/1-10-twitter-accounts-fake-say-researchers-2D1165536](http://www.nbcnews.com/technology/1-10-twitter-accounts-fake-say-researchers-2D1165536)
*** [https://sysomos.com/inside-twitter/most-active-twitter-user-data](https://sysomos.com/inside-twitter/most-active-twitter-user-data)
Why Bots are a Problem

- Bots harm the social media environment:
  - Manipulating discussions
  - Fake connections: follow or like for cash
  - Fake rating and reviews
Why Bots are a Problem

• Bots harm the social media environment:
  – Manipulating discussions
  – Fake connections: follow or like for cash
  – Fake rating and reviews

08 Twitter Bots Drown Out Anti-Kremlin Tweets

Thousands of Twitter accounts apparently created in advance to blast automated messages are being used to drown out Tweets sent by bloggers and activists this week who are protesting the disputed parliamentary elections in Russia, security experts said.

Amid widespread reports of ballot stuffing and voting irregularities in the election, thousands of Russians have turned out in the streets to protest. Russian police arrested hundreds of protesters who had gathered in Moscow’s Triumfalnaya Square, including notable anti-corruption blogger Alexei Navalny. In response, protesters began tweeting their disgust in a Twitter hashtag #триумфальная (Triumfalnaya), which quickly became one of the most-tweeted hashtags on Twitter.

But according to several experts, it wasn’t long before messages sent to that hashtag were drowned out by pro-Kremlin tweets that appear to have been sent by countless Twitter bots. Maxim Gorchakov, a senior
Bots in Politics - Russia

- Influencing discussion on political topics
- Manipulating protesters chats on Twitter*
- Injecting pro-Kremlin rhetoric**

Profile photos of 20500 fake accounts manipulating Russians' discussions on Twitter

“Ukrainians killed him... he was stealing one of their girlfriends”

** https://globalvoicesonline.org/2015/04/02/analyzing-kremlin-twitter-bots/
Bots in Politics - Russia

- More active shortly after political events
- Making connections with other fake accounts
- Lack of location, time zone, and favorites

Regular Follower Network

Bot Social Network
Bots in Politics

• Massachusetts 2010 Special Senate Election
  – Bots generated messages in support of a Senate candidate
  – Candidate gained 60,000 followers due to retweets [21]

• Mexican Elections
  – 2012 General Election
  – Used 10,000 automated accounts to promote Institutional Revolutionary Party

• Venezuela
  – Twitter bots used to inflate popularity of policies.
Why Bots are a Problem

• Bots harm the social media environment:
  – Manipulating discussions
  – Fake connections: follow or like for cash
  – Fake rating and reviews
Why Bots are a Problem

• Bots harm the social media environment:
  – Manipulating discussions
  – Fake connections: follow or like for cash
  – Fake rating and reviews

One bot farm operated 750,000 bot accounts
to promote diet pills.
Bot Characteristics

• Many accounts required
  – Twitter statistics.
  – Trending topics.

• Bots are made quickly
  – Automated account creation
  – Can be bought
  – At scale

• Now let’s discuss how to deal with bots
Bots & Network Analytics

Prof. Kathleen M. Carley
kathleen.carley@cs.cmu.edu
The Network Perspective

- It’s not just the elements (composition) of a system, but how they are put together
- Non-reductionist, holistic

- Networks are everywhere and they are connected
- Systems thinking
Social Networks

A social network is a description of the social structure at a particular point in time in terms of the actors, mostly individuals or organizations and the links among them. A social network indicates the ways in which the actors are connected through various social familiarities ranging from casual acquaintance to close familiar bonds.
Networks are Ubiquitous

**Nodes**
- People
- Topics
- Countries
- Hashtags
- Departments
- Resources
- Events

**Ties Between Nodes (links)**
- Transfer of resources
- Authority lines
- Affiliation
- Alliance
- Substitution
- Proximity
- What do you do
- Who do you like

- Claire Le Goues
- Bill Scherlis
- Christian Kaestner
- Jonathan Aldrich
- Joshua Sunshine
- David Garlan
- Mary Shaw
- Zico Kolter
- Kathleen M. Carley
- Lorrie Cranor
- Norman Sadeh
- Raj Reddy
- Angel Jordan
- Jim Herbsleb
- Jürgen Pfeffer
- Yuvraj Agarwal
- Travis Breaux

- software development tools
- software analysis and quality assurance
- modularity
- software tools and human cognition
- language design
- applied formal methods
- software design
- cyber-physical systems
- cybersecurity
- concurrency
- pervasive & mobile computing
- sustainability
- hardware
- big data & scalability
- complex socio-technical systems
- social media
- collaboration & coordination
- decision making
- computational methods
- organizations
- social network analysis
- computational modeling and simulation
- usable privacy and security
- hardware
- hardware
- hardware
- hardware
So – why is this hard?

- **The Network**
  - Vast quantities of data
  - Multi-mode – people, events, etc.
  - Multi-plex – many connections e.g. financial and authority
  - Geo-temporal

- **The Information**
  - Intentional misinformation – e.g., aliases
  - Inaccurate information – e.g., typos
  - Out-of-date information
  - Incomplete information
  - Inconsistent information
  - Varying levels of resolution in spatially and temporally

- **Dynamic**
  - Learning
  - Recruitment
  - Attrition
  - Daily activity...
Padang Indonesia
Local Twitter Network

Size = Out-Degree

1. McriskaMomo 29
2. editriaa 15
3. Aybekazta 13
4. vionaretha 10
5. Qwertyudii 8
Individual v. Network Level

- Individual behaviors are not independent of the network within which the behaviors occur.
- Individual network position is not independent of the network structure – being central in a centralized network is different than being central in a decentralized network.
Network Level Metrics Commonly Used

Metric
- Size
- Link count
- Density
- Isolate count
- Component count
- Reciprocity
- Characteristic path length
- Clustering coefficient

*Density & Size are Negatively Correlated*
Size

- Number of nodes (people) in the network
- Matters because as size increases
  - Density decreases
  - Clustering increases
- Reflects network boundary
- Should always be included as a covariate
Density

- Number of ties, expressed as percentage of the number of ordered/unordered pairs
- Number of ties / Number of possible ties
- If number of nodes = N and number of ties is M, then $M/(N*(N-1))$ if directed and $M/((N*(N-1))/2)$ if undirected

Low Density (25%)
Avg. Dist. = 2.27

High Density (39%)
Avg. Dist. = 1.76
Reciprocity (Mutuality, Symmetry)

- Mutual ties: \( A \rightarrow B \) then \( B \rightarrow A \)
- Some relations are inherently symmetric or asymmetric
  - Who did you have lunch with?
  - Who did you go to for advice?
- Reciprocity is calculated as the percent of ties that are reciprocated:

\[
R = \frac{(A_{ij} = 1) \& (A_{ji} = 1)}{(A_{ij} = 1) or (A_{ji} = 1)}
\]

- Reciprocal mentions is a measure of trust
Network Elite

- Nodes that stand out as high/low on some measure

- Power
  - Access to resources, information, people
  - Ability to mobilize others (reach)
  - Ability to control the flow of information
  - Ability to broker between groups
  - Ability to give orders
## Simple SNA Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Meaning</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree Centrality</strong></td>
<td>Node with the most connections</td>
<td>In the know</td>
<td>Identifying sources for intel; Reducing information flow</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Betweenness</strong></td>
<td>Node in the most best paths</td>
<td>Connects groups</td>
<td>Typically has political influence, but may be too constrained to act</td>
</tr>
<tr>
<td></td>
<td>Needs symmetric data</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Eigenvector centrality</strong></td>
<td>Node most connected to other highly connected nodes</td>
<td>Strong social capital</td>
<td>Identifying those who can mobilize others</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Closeness</strong></td>
<td>Node that is closest to all other nodes</td>
<td>Rapid access to all information</td>
<td>Identifying sources to acquire/transmit information</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Betweenness - Centrality</strong></td>
<td>High in betweenness but not degree centrality</td>
<td>Connects disconnected groups</td>
<td>Go-between; Reduction in activity by disconnecting groups</td>
</tr>
</tbody>
</table>
Degree Centrality

- Number of edges incident upon a vertex
  - \( d_8 = 6 \), while \( d_{10} = 1 \)
- Sum of degrees of all nodes is twice the number of edges in graph
- Average degree = density times \((n-1)\)
- Index of exposure to what is flowing through the network
  - Gossip network: central actor more likely to hear a given bit of gossip
- Interpreted as opportunity to influence & be influenced directly
- Predicts variety of outcomes from virus resistance to power & leadership to job satisfaction to knowledge
Degree Centrality

- **Degree** – total number of edges/nodes ego is connected to
  - Commonly thought of as a measure of influence or importance
- **In Degree** – total number of nodes that send edge to ego (column)
- **Out Degree** – total number of nodes that receive edge from ego (row)
- **Sink** – 0 in degree; **Source** – 0 out degree

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>In</th>
<th>Out</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

```
0 1 0 1 0
1 0 0 1 0
1 0 0 0 1
0 0 1 0 1
0 1 1 0 0
```
Betweenness Centrality

- How often a node lies along the shortest path between two other nodes
- Computed as:
  \[ b_k = \sum_{i,j} \frac{g_{ikj}}{g_{ij}} \]
  where \( g_{ij} \) is the number of geodesic paths from \( i \) to \( j \) and \( g_{ikj} \) is number of those paths that pass through \( k \)
- Index of potential for
  - gate-keeping, brokering, controlling the flow, and liaising between disparate parts of network – “connects groups”
- Indicates power, access to diversity of flows, potential for synthesizing
- Very “expensive” to compute
Closeness Centrality

• Measured as:
  – Sum of distances to all other nodes
  – Computed as marginals of symmetric geodesic distance matrix

• Closeness is an inverse measure of centrality

• Index of expected time until arrival for given node of whatever is flowing through the network
  – Gossip network: central player hears things first
Eigenvector Centrality

- Node has high score if connected to many nodes that are themselves well connected
- Computed as:

where $A$ is adjacency network and $V$ is eigenvector centrality. $V$ is the principal eigenvector of $A$

- Indicator of popularity, “in the know”
- Index of exposure, risk
- Tends to identify centers of large cliques
  - Often identified as leader of self-contained group
  - Leader of Leaders
Clustering Coefficient
(Carley et al., 2013: 845-847)

NODE LEVEL –
Measures the degree of clustering in a network by averaging the clustering coefficient of each node, which is defined as the density of the node's ego network.

Let $G=(V, E)$ be the graph representation of a square network. Define for each node $v \in V$ its Clustering Coefficient $cc(v)$:

let $G(v) = \text{ego network of entity } v$

Then Clustering Coefficient for entity: entity $v = cc(v) = \text{density } (G(v))$

GRAPH LEVEL -
Then Clustering Coefficient for graph: graph $= \sum_{v \in V} cc(v) / |V|$
Moving Beyond Single Measures

**Issue:** Centrality Measures are highly correlated

- Betweenness
- Degree

A Bridge!

Sink? Or Source?
Overall Tweet Network
Note there are a few sources that are picked up
Most Retweeted Actors are often news agencies
Hashtag network

Note – no direct linkage between Arabic topic-group and English
Most Critical Hashtags
These are ones that co-occur with other tags the most

![Graph showing the centrality and column degree of hashtags.](chart.png)
Estimates – 25 to 50% of tweeters not human

"AT FIRST, BOB WAS THRILLED WITH ALL THE ATTENTION AND FOLLOWERS..."
Number of Actors – Changes Radically Based on Suspensions

- Synthetic actors
  - If you want only human’s then remove bots
  - Estimates – 25 to 50% of tweeters not human

- Data is not stable
  - Twitter deletes accounts
  - Some are bots and some people
  - For some time periods and countries can be 25%
Change of Degree after Suspended Tweeters are Removed

- Degree centrality of remaining nodes may go down substantially
- Oct 2012 has the most change
Network impact is mixed

avg closeness and diameter can increase by 50% or decrease by 50%;
avg degree and so density decreases.

Change in Closeness

Change in Diameter
Impact on Network Metrics

- 25% nodes removed – chance you predicted top 3 correct drops to 60%
- 25% extra nodes (undetected bots) – chance you predicted top 3 correct drops to 80%
Bots and Network Statistics

- Bots tend to be network outliers
- Bots tend to be higher in degree centrality than normal tweeters
- Network results will vary based on whether or not you remove bots
  - If you don’t remove bots will often be network elite
  - If you rely on twitter they may miss some
  - If your could remove all bots
    - Avg degree will decrease
    - Other network metrics may increase or decrease
    - Will have less impact on hashtag network than social network
- But, there are classes of Bots and some look more human
Outline

Finding Ground Truth
- Manual Labeling
- Twitter as a Labeling Mechanism
- Employing Honeypots

Algorithms for Detecting Bots
- Bot Detection Heuristics
- Classification Algorithms

Applications
- Current Systems
- Resources
Finding Ground Truth

• What is ground truth?
  – We know there are many bots, but who are they?
• Why is ground truth important?
• Why is ground truth hard to obtain in bot detection?
**Manual Labeling**

- **Idea:** Humans are good at detecting bots.
- **Crowdsourcing**
- **Advantages**
  - Reliable labels
- **Drawbacks**
  - Expensive (time and money)

![CrowdFlower](image)

![SamaSource](image)

![Amazon Mechanical Turk](image)
Manual Labeling

• Example: Chu et al. 2010 [1]

• Features of each class:
  – **Human**: Posts thoughtful content; feelings.
  – **Bot**:
    • Unoriginal content (abundance of retweets, copied tweets)
    • Automated tweets (e.g. RSS) / Duplicate tweets
    • Spam or malicious URLs
    • Links with unrelated content.
  – **Cyborg**:
    • Different types of tweets.
    • Original content + RSS.

• Manually inspect: Tweet history, Profile, All URLs.
Manual Labeling

• Manual Labeling System [10]
Twitter as a Labeling Mechanism

- Three states of a Twitter user:
  - Active
  - Suspended
  - Deleted
- **Idea**: Use these states as a label [4-7].
- Researchers take two snapshots of each user.
Twitter as a Labeling Mechanism

Initial Crawl
- Finds seed set of users.
- Crawls Profile, Network, ...

Account suspended
This account has been suspended. Learn more about why Twitter suspends accounts.

Suspended

Sorry, that page doesn’t exist!
Search for a username, friend or hashtag.

Deleted

Active
Twitter as a Labeling Mechanism

• Advantages
  – Scalable
  – Cheap

• Drawbacks
  – Many suspended users are not suspended due to bot behavior (false positives).
  – *Twitter does not want to suspend accounts.*
  – Many bots are missed by Twitter (false negatives).
Twitter Labeling Example Dataset

• Example: Arab Spring Libya Data

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Geo Boxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>#libya, #gaddafi, #benghazi, #brega, #misrata, #nalut, #nafusa, #rhaibat</td>
<td>&lt;Geo Box around Libya&gt;</td>
</tr>
</tbody>
</table>

– Recrawled in 2014....

<table>
<thead>
<tr>
<th>Property</th>
<th>Libya</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>1,150,192</td>
</tr>
<tr>
<td>Retweets</td>
<td>576,167</td>
</tr>
<tr>
<td>Unique Users</td>
<td>94,535</td>
</tr>
<tr>
<td>Labeling Approach</td>
<td>Suspended Accts</td>
</tr>
<tr>
<td>Bot Ratio</td>
<td>7.5%</td>
</tr>
</tbody>
</table>
Honeypots

- Honeypot is “a trap set to detect ... attempts at unauthorized use of information systems”.
- **Idea**: Create “social” honeypots to identify bots in the wild.
  - Honeypots should imitate the behavior of the bots they are trying to detect.
  - Can find more bots than just those identified by Twitter.
Honeypots

The Social Honeypot Project [8]
Honeypot Goals

• Act as obvious bot accounts
• Attract other bot accounts
• Collect many bot accounts
  – Bots are identified when they follow our account.
• Assumption: Real users do not follow bots.
Honeypots

• Advantages
  – Accurate
  – Automatic

• Drawbacks
  – Time consuming
  – Does not yield regular accounts
Honeypot Ingredients

• **Post “Luring” Content**
  – Post content that will be seen
  – trending topics, hashtags, “famous” tweets...

• **Maintain Network Connections**
  – “Follow back”, Retweets
  – Fame begets fame

• **Promote Other Honeypots**
  – Retweet each other’s tweets
  – Mention each other

<table>
<thead>
<tr>
<th>Paris Trends</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>#QLFCL</td>
<td></td>
</tr>
<tr>
<td>#Capital</td>
<td></td>
</tr>
<tr>
<td>Lorient</td>
<td></td>
</tr>
<tr>
<td>James Bond</td>
<td></td>
</tr>
<tr>
<td>#Missionimpossible3</td>
<td></td>
</tr>
<tr>
<td>Rafael</td>
<td></td>
</tr>
<tr>
<td>Nicole Scherzinger</td>
<td></td>
</tr>
<tr>
<td>Fekir</td>
<td></td>
</tr>
<tr>
<td>Coutinho</td>
<td></td>
</tr>
<tr>
<td>Ohezzal</td>
<td></td>
</tr>
</tbody>
</table>
Case Study: Content Polluters

• Lee et. al [8] created one of first honeypot farms.
  – 60 honeypots deployed.
  – Collected 22,000+ bot users in 8 months.

• **Problem:** how to find legitimate users?
  – Observed another sample of 19,000 users.
  – Used “Twitter Labeling” technique.
Case Study: Political Extremists

• **Goal:** Collect topic-specific bots.
  – Extremist-specific keywords.

• **Architecture:**
  – One controller.
  – 27 bot accounts.

• Captured 3,178 bots in 2 months.

• **Problem:** how to find legitimate users?
  – Real users do not follow bots.
  – Vet accounts, and collect who they follow.
Honeypot Code


This directory includes python code for a set of honeypots on twitter.

TahoraNazer / HoneypotCode

- [Branch: master](#) / HoneypotCode / +

Merge pull request #1 from wulang211/master  

- TahoraNazer authored 10 minutes ago

- latest commit ee6ea8a14f

<table>
<thead>
<tr>
<th>Folder</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>code and data</td>
<td>13 minutes ago</td>
</tr>
<tr>
<td>src</td>
<td>code and data</td>
<td>13 minutes ago</td>
</tr>
<tr>
<td>README.md</td>
<td>Initial commit</td>
<td>27 minutes ago</td>
</tr>
</tbody>
</table>
Buying Bot Followers:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Provider</th>
<th>Highlights</th>
<th>Quick Review</th>
<th>Pricing</th>
<th>Links</th>
</tr>
</thead>
</table>
| #1   | Devumi   | • Up to 500k Twitter Followers  
      |          | • 100% Money-Back Guarantee  
      |          | • Replacement Guarantee  
      |          | • No Password Required  
      |          | • Excellent Customer Support  
      |          | • Saw Results in 24 Hours  
      |          | Test Account: @DevumiReview  
      |          | Our Experience  
      |          | • Completed within 24 Hours  
      |          | • Over-Delivered by 30%  
      |          | • Quick & Friendly Customer Support  
      |          | • Offers Sponsored Mentions  
      |          | • No Losses in Followers  
      |          | From $12  
      |          | Accepts: Credit Card, PayPal & Bitcoin  
      |          | Visit Site  
      |          | Road Review |
| #2   | FastFollowerz | • Up to 1M Twitter Followers  
      |          | • 100% Money-Back Guarantee  
      |          | • Replacement Guarantee  
      |          | • No Password Required  
      |          | • Good Customer Support  
      |          | • Saw Results in 48 Hours  
      |          | Test Account: @FFzReview  
      |          | Our Experience  
      |          | • Completed within 48 Hours  
      |          | • Over-Delivered by 7%  
      |          | • Quick & Friendly Customer Support  
      |          | X Unstable (Lost Followers Overtime)  
      |          | From $16  
      |          | Accepts: 2Checkout  
      |          | Visit Site  
      |          | Road Review |
| #3   | Twitter Boost | • Up to 400k Twitter Followers  
      |          | • 100% Money-Back Guarantee  
      |          | • Replacement Guarantee  
      |          | • Decent Customer Support  
      |          | • Saw Results in 1 days  
      |          | Test Account: @TwBoostReview  
      |          | Our Experience  
      |          | • Completed within 1 Days  
      |          | • Over-Delivered by 20%  
      |          | • Daily Retweets Service  
      |          | • No Drops in Followers yet.  
      |          | From $7  
      |          | Accepts: Credit Card, PayPal & Bitcoin  
      |          | Visit Site  
      |          | Road Review |

- Average price: $10 for 1,000 followers.
- Advantage: real bot accounts.
- Disadvantage: cost.
Honeypots without Network

- Generating massive amounts of content.
- Publishing many, many tweets with tags.
- Leverages bots’ need for “authentic content”.

---

Rita Jennings @r  
111:1164 This passion, and the death of a dear friend, would go near to make a man look sad.  
#AMNDBots

Carys @  
109:832 Exit PUCK.--DEMETRIUS, HELENA &c, sleep.  
#AMNDBots
Honeypots without Network


A collection of helper scripts and bots that act out the play "A Midsummer Night's Dream" on Twitter
Bot Datasets

• Bot Datasets:
  – Lee’s Honeypot Dataset [9]: http://infolab.tamu.edu/data/

• Code:
Outline

Finding Ground Truth
- Manual Labeling
- Twitter as a Labeling Mechanism
- Employing Honeypots

Algorithms for Detecting Bots
- Bot Detection Heuristics
- Classification Algorithms

Applications
- Current Systems
- Resources
1. The Fundamentals: Heuristics
   - How to use the gold standard datasets.
   - How to evaluate your approach.

2. State-of-the-art Classification Algorithms

3. Real-World Systems

4. The Future of Bot Detection
Heuristics

• A heuristic is “any approach ... not guaranteed to be optimal or perfect, but sufficient for the immediate goals”
• Used in bot detection [1, 2, 3].
• **Assumption:** there is a “superfeature” which can distinguish bots from regular users.
Common Heuristics

- Friend / Follower Ratio
- Account Age (days)
- Frequency of URLs
- ...

\[ f(\cdot) \rightarrow R \]
Heuristics Intuition

\[ f(\cdot) \quad f(\cdot) \quad f(\cdot) \quad f(\cdot) \quad f(\cdot) \quad f(\cdot) \]
Evaluation

- **Precision:** \[ P = \frac{|B \cap G|}{|B|} \]  
  “How many of those classified were correct?”

- **Recall:** \[ R = \frac{|B \cap G|}{|G|} \]  
  “How many of the known bots were found?”

- **F_1 Score:** \[ F_1 = \frac{2PR}{P + R} \]

- **Fall-Out:** \[ F = 1 - \frac{|U - G - B|}{|U - G|} \]  
  “How many false alarms?”
Evaluation

- **Recall:**
  \[ R = \frac{|B \cap G|}{|G|} \]
  “How many of the known bots were found?”

- **Fall-Out:**
  \[ F = 1 - \frac{|U - G - B|}{|U - G|} \]
  “How many false alarms?”
Heuristics Example

- **Example:** Lee et. al 2010 [8]
- Proposed simple heuristics to separate bots from humans.
Classification
Classification

• Build a classifier that can identify bots.
  – Quality of classifier based upon quality of features.

• Features be based upon:
  – Heuristics -> “feature engineering”.
  – Features derived automatically (e.g., from text).
  – Latent features.
Classification

Training Data
(labeled text corpus)

Feature Extraction

Model Training

Model

Predictions

Test Data
(text corpus to be labeled)

Design Time

Run Time

Bot Detection in Social Media, August 25, 2015
Classification with Heuristics

- User is represented by \( n \) heuristics.

\[
\begin{array}{cccc}
H_1 & H_2 & \ldots & H_n \\
\hline
\hline
\end{array}
\]

Ground Truth

- HUMAN
- BOT
- ...
- HUMAN

Model

- Training

Predictions
Classification with Heuristics

- **Example:** Lee et. al 2011 [9]
- **Authors hand-pick several heuristics**

**Feature:**
Standard deviation of #followers over time.

\[
\sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (f_{i+1} - f_i)}
\]
Classification with Heuristics

- **Example:** Lee et. al 2011 [9]
- Authors hand-pick several heuristics
- Four “categories”:

<table>
<thead>
<tr>
<th>Demographics (UD)</th>
<th>Network (UFN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account age</td>
<td>#following/#followers</td>
</tr>
<tr>
<td>Screen name length</td>
<td>% bidirectional friends</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content (UC)</th>
<th>History (UH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#links/#tweets</td>
<td>Standard Deviation of followers over time</td>
</tr>
<tr>
<td>#unique links/#tweets</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Classification with Heuristics

- **Example:** Lee et. al 2011 [9]
- Authors hand-pick several features
- Four “categories”
- Tested features with 30 classifiers

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>F₁</th>
<th>AUC</th>
<th>FNs</th>
<th>FPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>98.42%</td>
<td>0.984</td>
<td>0.998</td>
<td>301</td>
<td>354</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Accuracy</th>
<th>F₁</th>
<th>AUC</th>
<th>FNs</th>
<th>FPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>UD</td>
<td>76.17%</td>
<td>0.762</td>
<td>0.839</td>
<td>6,007</td>
<td>3,882</td>
</tr>
<tr>
<td>UH</td>
<td>85.34%</td>
<td>0.854</td>
<td>0.899</td>
<td>4,130</td>
<td>1,950</td>
</tr>
<tr>
<td>UC</td>
<td>86.39%</td>
<td>0.864</td>
<td>0.932</td>
<td>2,811</td>
<td>2,837</td>
</tr>
<tr>
<td>UFN</td>
<td>96.46%</td>
<td>0.965</td>
<td>0.992</td>
<td>510</td>
<td>958</td>
</tr>
<tr>
<td>UD+UC</td>
<td>88.61%</td>
<td>0.886</td>
<td>0.953</td>
<td>2,469</td>
<td>2,256</td>
</tr>
<tr>
<td>UD+UH</td>
<td>92.45%</td>
<td>0.925</td>
<td>0.967</td>
<td>1,743</td>
<td>1,389</td>
</tr>
<tr>
<td>UC+UH</td>
<td>94.38%</td>
<td>0.944</td>
<td>0.979</td>
<td>1,111</td>
<td>1,221</td>
</tr>
<tr>
<td>UFN+UH</td>
<td>97.11%</td>
<td>0.971</td>
<td>0.994</td>
<td>496</td>
<td>702</td>
</tr>
<tr>
<td>UD+UFN</td>
<td>97.50%</td>
<td>0.975</td>
<td>0.995</td>
<td>437</td>
<td>597</td>
</tr>
<tr>
<td>UFN+UC</td>
<td>97.92%</td>
<td>0.979</td>
<td>0.996</td>
<td>413</td>
<td>448</td>
</tr>
<tr>
<td>UD+UC+UH</td>
<td>95.42%</td>
<td>0.954</td>
<td>0.985</td>
<td>878</td>
<td>1,022</td>
</tr>
<tr>
<td>UD+UFN+UH</td>
<td>97.78%</td>
<td>0.978</td>
<td>0.996</td>
<td>395</td>
<td>524</td>
</tr>
<tr>
<td>UFN+UC+UH</td>
<td>98.13%</td>
<td>0.981</td>
<td>0.997</td>
<td>361</td>
<td>413</td>
</tr>
<tr>
<td>UD+UFN+UC</td>
<td><strong>98.26%</strong></td>
<td>0.983</td>
<td><strong>0.997</strong></td>
<td>333</td>
<td><strong>388</strong></td>
</tr>
</tbody>
</table>
Classification

Training Data (labeled text corpus)

Test Data (text corpus to be labeled)

Feature Extraction

Model Training

Model

Predictions

- Accuracy
- $F_1$
- AUC
- FNs
- FPs

<table>
<thead>
<tr>
<th>Demographics (UD)</th>
<th>Network (UFN)</th>
<th>Content (UC)</th>
<th>History (UH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account age</td>
<td>#following/#followers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screen name length</td>
<td>% bidirectional friends</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#links/#tweets</td>
<td>Standard Deviation of followers over time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#unique links/#tweets</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Three-Class Classification: Human, Bot, Cyborg

- **Example:** Chu et. al 2010 [1]
- Three Classes: Humans, Bots, Cyborgs
- Cyborgs:
  - Bot-Assisted Humans (e.g., RSS)
  - Human-Assisted Bots

<table>
<thead>
<tr>
<th>Rank</th>
<th>Human</th>
<th>Bot</th>
<th>Cyborg</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Web (50.53%)</td>
<td>API (42.39%)</td>
<td>Twitterfeed (31.29%)</td>
</tr>
<tr>
<td>#2</td>
<td>TweetDeck (9.19%)</td>
<td>Twitterfeed (26.11%)</td>
<td>Web (23.00%)</td>
</tr>
<tr>
<td>#3</td>
<td>Tweetie (6.23%)</td>
<td>twitRobot (13.11%)</td>
<td>API (6.94%)</td>
</tr>
</tbody>
</table>
Three-Class Classification: Human, Bot, Cyborg

• Account Properties Component
  – # tweets with URLs / # tweets
  – # followers / # friends
  – Client makeup (web / mobile / automated)
  – Compare URLs against Google’s Blacklists
Bot Detection in Social Media, August 25, 2015
Three-Class Classification: Human, Bot, Cyborg

- Machine Learning Component
  - Takes words in message, compares them with other messages.
  - Calculates probability that message is spam by comparing with other messages.
Three-Class Classification: Human, Bot, Cyborg

Raw Text
User 1: the brown fox
User 2: the white tiger

Word Counts
User 1: {“the”: 1, “brown”: 1, “fox”: 1}
User 2: {“the”: 1, “white”: 1, “tiger”: 1}

Feature Vectors

<table>
<thead>
<tr>
<th></th>
<th>“the”</th>
<th>“brown”</th>
<th>“fox”</th>
<th>“white”</th>
<th>“tiger”</th>
<th>&lt;label&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>BOT</td>
</tr>
<tr>
<td>User 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>HUMAN</td>
</tr>
</tbody>
</table>
• Entropy Component
  – Entropy of time intervals between posts.
Three-Class Classification: Human, Bot, Cyborg

- Decision Maker
  - Learns a scoring for each class $i$, $S_i$.

$$S_i = w_{i0} + \sum_{i=1}^{n} w_{i1}v_1 + w_{i2}v_2 + \ldots + w_{im}v_m$$

- User is labeled with highest-scoring class.
### Three-Class Classification: Human, Bot, Cyborg

- Humans are never confused with bots.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Classified</th>
<th>Human</th>
<th>Cyborg</th>
<th>Bot</th>
<th>Total</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Human</td>
<td>949</td>
<td>51</td>
<td>0</td>
<td>1000</td>
<td>94.90%</td>
</tr>
<tr>
<td></td>
<td>Cyborg</td>
<td>98</td>
<td>828</td>
<td>74</td>
<td>1000</td>
<td>82.80%</td>
</tr>
<tr>
<td></td>
<td>Bot</td>
<td>0</td>
<td>63</td>
<td>937</td>
<td>1000</td>
<td>93.70%</td>
</tr>
</tbody>
</table>
Classification: Early Detection

- **Example**: Zafarani and Liu 2015 [17]
- Classify username with minimum information.
- Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.

---

Example: Zafarani and Liu 2015 [17]

Classify username with minimum information.

Minimum Information: the username.
Early Detection

• Features:
  – Information Surprise: $p(\text{username} \mid \text{site})$
  – Gender: Compare names with census data
  – Language: What language was it written in?
  – Keyboard:
    • Percentage of keys typed using the same hand in succession.
    • Percentage of keys on each row.
    • Travel (cm) of fingers on the keyboard.
Early Detection

• Results:

<table>
<thead>
<tr>
<th>Technique</th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>0.9932</td>
<td>0.9644</td>
</tr>
<tr>
<td>Baseline $b_1$: Keyword Detection</td>
<td>0.51</td>
<td>0.66</td>
</tr>
<tr>
<td>Baseline $b_2$: Username Randomness</td>
<td>0.70</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>Baseline $b_3$: Letter Repetition</td>
<td>0.61</td>
<td>$\approx 0$</td>
</tr>
</tbody>
</table>

• Rank features by importance:
  – Information surprise of username
  – Number of digits used in username.
  – Keyboard-related features.
Clustering
Clustering

- **Traditional:** Separate users into \( n \) clusters.

- **Bot Detection:** Use \( n \) clusters as features.
Clustering

• **Example:** Lee et. al 2014 [16]
• **Goal:** detect bots at account creation time.
• **Bots often attack quickly.**
**Clustering**

- **Idea:** Cluster by username.
- **Two features:**
  - Length of username (characters)
  - Cluster usernames

<table>
<thead>
<tr>
<th>Username</th>
<th>Length</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>3</td>
<td>1/2</td>
</tr>
<tr>
<td>atom</td>
<td>4</td>
<td>1/6</td>
</tr>
<tr>
<td>melon</td>
<td>5</td>
<td>1/6</td>
</tr>
</tbody>
</table>

![Diagram showing clustering process](image)
Classifying Clusters

- Each *cluster* is classified
- Classified using engineered features:
  - Average edit distance between members.
  - Character unigram distribution distance from verified names.
  - Length distribution distance from verified names.
Results

- URL – Just separating by banned URLs.
- Can achieve better results in 30 mins.
Clustering: Topic Modeling

• Topic models discover abstract “topics” within a corpus.

• Latent Dirichlet Allocation (LDA): clusters tokens into $K$ topics [18].

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>...</th>
<th>Topic $K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>0.2</td>
<td>0.1</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Document 2</td>
<td>0.7</td>
<td>0.02</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document $n$</td>
<td>0.1</td>
<td>0.3</td>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>cat, dog, horse, ...</td>
</tr>
<tr>
<td>Topic 2</td>
<td>ball, field, player, ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Topic $K$</td>
<td>red, green, blue, ...</td>
</tr>
</tbody>
</table>
Clustering: Topic Modeling

- Intuition: bots focus on different topics than real users.
- This approach automatically learns these topic preferences.

### Table

<table>
<thead>
<tr>
<th>User</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>...</th>
<th>Topic K</th>
</tr>
</thead>
<tbody>
<tr>
<td>User_1</td>
<td>0.2</td>
<td>0.1</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>User_2</td>
<td>0.7</td>
<td>0.02</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User_n</td>
<td>0.1</td>
<td>0.3</td>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>

Classifier
Clustering: Matrix Factorization

• Factor a matrix into (usually) two factors.

\[
V \approx WH
\]

User x Feature Matrix

User x Latent Dimension

Latent Dimension x Feature
Clustering: Matrix Factorization

- **Example:** Hu et. al 2014 [19]
- Can sentiment be used to find bots?

Twitter bot replies to offer prizes related to events such as NFL or Miley Cyrus
• Find difference in the sentiment of regular users and bots.
• Leverage this difference to classify users...
Clustering: Matrix Factorization

- Content Information
  - $X = U V^T$
  - Encoding Matrix

- Sentiment Extraction and Modeling
  - $s$
  - $L$
  - $C$

- Social Network Information
  - $\Delta$
  - $G$
# Social Spammer Detection Results on TUSH Dataset

<table>
<thead>
<tr>
<th></th>
<th>Training Data One (50%)</th>
<th></th>
<th>Training Data Two (100%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F₁-measure (gain)</td>
<td>Precision</td>
</tr>
<tr>
<td>Content_Net</td>
<td>0.893</td>
<td>0.924</td>
<td>0.908 (N.A.)</td>
<td>0.919</td>
</tr>
<tr>
<td>Content_Lap</td>
<td>0.926</td>
<td>0.939</td>
<td>0.932 (+2.67%)</td>
<td>0.931</td>
</tr>
<tr>
<td>SMFSR</td>
<td>0.935</td>
<td>0.939</td>
<td>0.937 (+3.12%)</td>
<td>0.948</td>
</tr>
<tr>
<td>SparseSD</td>
<td>0.951</td>
<td>0.955</td>
<td>0.953 (+4.93%)</td>
<td>0.959</td>
</tr>
<tr>
<td>SDS</td>
<td>0.969</td>
<td>0.965</td>
<td>0.967 (+6.47%)</td>
<td>0.975</td>
</tr>
</tbody>
</table>

### Bar Chart

- **Precision**
- **Recall**
- **F₁-measure**

Legend:
- Content
- Network
- Sentiment
- Cont_Lap
- Cont_Sent
- SDS
Final Thoughts on Classification

Labeled Data Acquisition:
1. Manual
2. Twitter
3. Honeypots
4. Purchasing
5. Making Noise

Training Data (labeled text corpus)

Model Training

Feature Extraction

Model

Predictions

Test Data (text corpus to be labeled)

Design Time

Run Time
Final Thoughts on Classification

Feature Extraction:
1. Feature Engineering
2. Automatically Finding Features
3. Clustering to obtain features
Outline

Finding Ground Truth
- Manual Labeling
- Twitter as a Labeling Mechanism
- Employing Honeypots

Algorithms for Detecting Bots
- Bot Detection Heuristics
- Classification Algorithms

Applications
- Current Systems
- Resources
• Bot detection systems “in the wild”.
• Most are proprietary (e.g. Twitter, Facebook).
• Research community has developed their own.
Bot Classification

- Example: *Bot or Not?* [11,12]
- Public system to classify Twitter users.
- Relies on discriminative features.

http://truthy.indiana.edu
Fake Follower Detection

- **TwitterAudit***
- Heuristics-based
  - Has the user tweeted before?
  - Does the user have a profile picture?
  - When did the user join?

*https://www.twitteraudit.com/
Prof. Carley’s Real World Example
Outline

Finding Ground Truth
- Manual Labeling
- Twitter as a Labeling Mechanism
- Employing Honeypots

Algorithms for Detecting Bots
- Bot Detection Heuristics
- Classification Algorithms

Applications
- Current Systems
- Resources
Afterthoughts

• Many aspects to explore

• Trafficking Fraudulent Accounts [13]
  – How bot masters sell/trade accounts
  – What factors influence the price of accounts.
    • Phone-verified accounts.
    • Profile completeness.
  – How bot master’s “season” accounts
Future Directions and Conclusion
Future Directions

- Bot detection with malicious post analysis

- Existing frameworks simply model all content
  - All posts are equally analyzed to represent bots

- Bot detection with malicious post analysis
  - Selecting suspicious information and weighting malicious posts

Features
Future Directions

• Link Farming of Bots

• Bots are “social” with real users
  – Humans connect with bots due to reflexive reciprocity

• Bots with real users as friends are difficult to detect.
Future Directions

- Bot detection with community structure analysis
  - Existing frameworks consider links from pair-wise perspective
    - Linked nodes are assumed to have similar identity
  - Social network structure helps detect outliers
    - Bots randomly connect to humans without joining any social community
Future Directions

• Overcoming limited ground truth
  – Evaluating without ground truth? (CACM 2015.6)
• Bot detection with labeled data is reduced to a binary classification.
• Only labels of some positive instances are available (e.g., honeypots).

• Automatic approaches (suspension list, honeypot) to obtaining ground truth can only provide positive labels.
Future Directions

• Bot detection with PU-Learning
  – PU-Learning: Learning with Positive and Unlabeled Data [20]

• Iteratively select bots and regular users during learning
Useful Bot Detection Resources

• Bot Datasets:
  – Lee’s Honeypot Dataset [9]: http://infolab.tamu.edu/data/

• Code:

• “Crap” Detection Resources:
Twitter Data Analytics

- Common tasks in mining Twitter Data.
  - Free Download with Code & Data
  - Collection
  - Analysis
  - Visualization

[tweettracker.fulton.asu.edu/tda/](tweettracker.fulton.asu.edu/tda/)
• Social Media Mining: An Introduction – a textbook
• A comprehensive coverage of social media mining techniques
  – Free Download
  – Network Measures and Analysis
  – Influence and Diffusion
  – Community Detection
  – Classification and Clustering
  – Behavior Analytics
Acknowledgements

• DMML Lab Members

• MINERVA initiative through the ONR N000141310835 on Multi-Source Assessment of State Stability
References


References


Clustering Example: http://pypr.sourceforge.net/kmeans.html
Matrix Factorization: https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

Twitter Bot Icon: http://icons.iconarchive.com/icons/lboi/tweetscotty/256/twitter-bot-icon.png
Keyboard Tower: http://forum.xcitefun.net/the-keyboard-tower-t71170.html