The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks

Oana Goga (MPI-SWS), Giridhari Venkatadri (MPI-SWS), Krishna P. Gummadi (MPI-SWS)

IMC 2015, October 28th
Weak identities

Are unverified identities that do not require users to prove their online identities match their offline person.
Weak identities

Are unverified identities that do not require users to prove their online identities match their offline person.

- Lower sign-on barriers, provide anonymity
- Leave systems vulnerable to Sybil attacks (fake identities)
Identity impersonation attacks

Special class of fake identities attacks: the attacker spoofs the identity of another real-world user.
Identity impersonation attacks

Special class of fake identities attacks: the attacker spoofs the identity of another real-world user.

celebrity impersonation attack
Identity impersonation attacks

Special class of fake identities attacks: the attacker spoofs the identity of another real-world user.
Identity impersonation attacks

Special class of fake identities attacks: the attacker spoofs the identity of another real-world user.

- Damage the online image of victims & affect victims in the offline world!
Identity impersonation attacks

Special class of fake identities attacks: the attacker spoofs the identity of another real-world user.

• Damage the online image of victims & affect victims in the offline world!
• Impersonation attacks are increasingly easy to mount due to the availability of personal information online!
Current situation

• Lack of understanding of impersonation attacks online!
  - No large dataset about real-world impersonation attacks
Current situation

• Lack of understanding of impersonation attacks online!
  ➤ No large dataset about real-world impersonation attacks

• Lack of frameworks to automatically detect impersonation attacks online
  ➤ Detection relies on manual reports
Contributions

First extensive study of real-world impersonation attacks in online social networks.

1. Methodology to gather data about impersonation attacks
2. Characterization of impersonation attacks in Twitter
3. Automatic detection of impersonation attacks
Contributions

First extensive study of real-world impersonation attacks in online social networks.

1. Methodology to gather data about impersonation attacks
2. Characterization of impersonation attacks in Twitter
3. Automatic detection of impersonation attacks
Challenges in data gathering

People results for nick feamster

Nick Feamster  @feamster
Associate Professor of Computer Science, Georgia Tech
Followed by Pablo Rodriguez and 2 others

Nicholas Feamster  @ntfeamster

Nick Feamster  @feamster_
Associate Professor of Computer Science, Georgia Tech
Challenges in data gathering

How to determine which identities try to portray the same user?
Challenges in data gathering

How to determine which identities try to portray the same user?

How similar the profiles of two identities should be to qualify as portraying the same user?
Challenges in data gathering

People results for nick feamster

Nick Feamster @feamster
Associate Professor of Computer Science,
Georgia Tech
Followed by Pablo Rodriguez and 2 others

doppelgänger pair

Nick Feamster @feamster_
Associate Professor of Computer Science,
Georgia Tech
Challenges in data gathering

How to determine if a doppelgänger pair is an impersonation attacks?
Challenges in data gathering

How to determine if a doppelgänger pair is an impersonation attacks?

victim-impersonator pair
Challenges in data gathering

How to determine if a doppelgänger pair is an impersonation attack?

- victim-impersonator pair
- avatar-avatar pair
Challenges in data gathering

How to determine if a doppelgänger pair is an impersonation attack?

Victim-impersonator pair

Avatar-avatar pair

How to determine which identity is legitimate and which is an impersonator?
Challenge 1: Identifying doppelgänger pairs

• Identify pairs of identities that most humans believe they portray the same person
  • Every identity has a name, location, bio and photo
  • Automated rule-based matching scheme (trained on human-annotated data, determines when the profile attributes of two identities matches sufficiently)
Challenge 2 and 3
Challenge 2 and 3

Identify victim-impersonator pairs

- Exploit Twitter suspension signals: when Twitter suspends one but not both identities
Challenge 2 and 3

Identify victim-impersonator pairs

- Exploit Twitter suspension signals: when Twitter suspends one but not both identities

Identify avatar-avatar pairs

- Exploit interactions between identities: clear indication that one identity is aware of the other
Challenge 2 and 3

Identify victim-impersonator pairs

- Exploit Twitter suspension signals: when Twitter suspends one but not both identities

Solves challenge 3 as well!

Impersonating identity = suspended identity

Identify avatar-avatar pairs

- Exploit interactions between identities: clear indication that one identity is aware of the other
# Twitter dataset

<table>
<thead>
<tr>
<th>Type of Pair</th>
<th>Random Dataset</th>
<th>BFS Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
## Twitter dataset

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
## Twitter dataset

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
Contributions

First extensive study of real-world impersonation attacks in online social networks.

1. Methodology to gather data about impersonation attacks
2. Characterization of impersonation attacks in Twitter
3. Automatic detection of impersonation attacks
Contributions

First extensive study of real-world impersonation attacks in online social networks.

1. Methodology to gather data about impersonation attacks
2. Characterization of impersonation attacks in Twitter
3. Automatic detection of impersonation attacks
# Twitter dataset

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
Types of impersonation attacks
Types of impersonation attacks

• Celebrity impersonation attacks
  • **Purpose**: exploits or maligns the reputation of the victim
  • **Detection**: victim has more than 10,000 followers or is verified
Types of impersonation attacks

- Celebrity impersonation attacks 3% (in the random dataset)
  - Purpose: exploits or maligns the reputation of the victim
  - Detection: victim has more than 10,000 followers or is verified
Types of impersonation attacks

• Celebrity impersonation attacks 3% (in the random dataset)
  • Purpose: exploits or maligns the reputation of the victim
  • Detection: victim has more than 10,000 followers or is verified

• Social engineering attacks
  • Purpose: abuses victim’s friends: reveal sensitive info, send money
  • Detection: attacker contacts victim’s friends
Types of impersonation attacks

- Celebrity impersonation attacks → 3% (in the random dataset)
  - Purpose: exploits or maligns the reputation of the victim
  - Detection: victim has more than 10,000 followers or is verified

- Social engineering attacks → 2% (in the random dataset)
  - Purpose: abuses victim’s friends: reveal sensitive info, send money
  - Detection: attacker contacts victim’s friends
Types of impersonation attacks

- Celebrity impersonation attacks
  - **Purpose**: exploits or maligns the reputation of the victim
  - **Detection**: victim has more than 10,000 followers or is verified

- Social engineering attacks
  - **Purpose**: abuses victim’s friends: reveal sensitive info, send money
  - **Detection**: attacker contacts victim’s friends

**Most impersonation attacks do not target celebrities or try to mount social engineering attacks!**

- Most impersonation attacks (in the random dataset): 3%
- Social engineering attacks: 2% (in the random dataset)
Types of impersonation attacks

- Celebrity impersonation attacks
  - Purpose: exploits or maligns the reputation of the victim
  - Detection: victim has more than 10,000 followers or is verified

- Social engineering attacks
  - Purpose: abuses victim’s friends: reveal sensitive info, send money
  - Detection: attacker contacts victim’s friends

Most impersonation attacks do not target celebrities or try to mount social engineering attacks!

What is possibly motivating the attackers?
Doppelgänger bot attacks hypothesis

H1: The attackers create these identities to abuse Twitter (and not the victims)

H2: The attackers attempt to create real-looking fake identities to evade the Twitter Sybil defense system
H1: The attackers create these identities to abuse Twitter (and not the victims)

H2: The attackers attempt to create real-looking fake identities to evade the Twitter Sybil defense system
Doppelgänger bot attacks hypothesis

H1: The attackers create these identities to abuse Twitter (and not the victims)

H2: The attackers attempt to create real-looking fake identities to evade the Twitter Sybil defense system

doppelgänger bot attacks ≠ doppelgänger pair!
# Twitter dataset

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
<tr>
<td></td>
<td>RANDOM DATASET</td>
<td>BFS DATASET</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
The Twitter dataset consists of two main datasets: RANDOM DATASET and BFS DATASET.

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
## Twitter dataset

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial accounts</td>
<td>1.4 million</td>
<td>142,000</td>
</tr>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
</tbody>
</table>
Doppelgänger bot attacks evidence for hypothesis I

H1: The attackers create these identities to abuse Twitter (and not the victims)
H1: The attackers create these identities to abuse Twitter (and not the victims)

Evidence:

- Large number of impersonators follow the same users
- The users they follow are suspected of having bought fake followers (http://trulyfollowing.app-ns.mpi-sws.org/)
Doppelgänger bot attacks evidence for hypothesis I

H1: The attackers create these identities to abuse Twitter (and not the victims)

Evidence:

- Large number of impersonators follow the same users
- The users they follow are suspected of having bought fake followers (http://trulyfollowing.app-ns.mpi-sws.org/)

follower fraud
Doppelgänger bot attacks evidence for hypothesis 2

H2: Attackers create real-looking fake identities to evade the Twitter Sybil defense system
Doppelgänger bot attacks evidence for hypothesis 2

H2: Attackers create real-looking fake identities to evade the Twitter Sybil defense system

Evidence:

- Twitter took in median 278 days to suspend the impersonating identities
- Other traditional Sybil detection schemes perform badly
Doppelgänger bot attacks
evidence for hypothesis 2

H2: Attackers create real-looking fake identities to evade the Twitter Sybil defense system.

Evidence:
- Twitter took in median 278 days to suspend the impersonating identities.
- Other traditional Sybil detection schemes perform badly.

Can we do something to detect impersonating identities faster?
Contributions

First extensive study of real-world impersonation attacks in online social networks.

1. Methodology to gather data about impersonation attacks
2. Characterization of impersonation attacks in Twitter
3. Automatic detection of impersonation attacks
Contributions

First extensive study of real-world impersonation attacks in online social networks.

1. Methodology to gather data about impersonation attacks
2. Characterization of impersonation attacks in Twitter
3. Automatic detection of impersonation attacks
Detection of impersonation attacks

- doppelgänger pair
  - victim-impersonator pair
    - victim
    - impersonator
  - avatar-avatar pair
Detection of impersonation attacks

- doppelgänger pair
- rule-based matching scheme
- victim-impersonator pair
- avatar-avatar pair
- victim
- impersonator
Detection of impersonation attacks

- **doppelgänger pair**
  - **victim-impersonator pair**
  - **avatar-avatar pair**

Rule-based matching scheme
Detection of impersonation attacks

- **Rule-based matching scheme**
  - doppelgänger pair
  - victim-impersonator pair
  - avatar-avatar pair
  - The impersonating identity is newer and has a lower reputation
Detection of impersonation attacks

- doppelgänger pair
- victim-impersonator pair
- avatar-avatar pair

Rule-based matching scheme

The impersonating identity is newer and has a lower reputation

18
Automated detection of victim-impersonator pairs

SVM classifier to distinguish between victim-impersonator pairs and avatar-avatar pairs

• Training and testing:
  • labeled doppelgänger pairs from our dataset

• Features that characterize pairs of identities:
  • user-names, screen-names, location, profile photos, bios, interest similarity; number of common followers, followings, users mentioned, and retweeted; time difference between creation dates, first and last tweets, outdated account
Automated detection of victim-impersonator pairs

SVM classifier to distinguish between victim-impersonator pairs and avatar-avatar pairs

- Training and testing:
  - labeled doppelgänger pairs from our dataset
- Features that characterize pairs of identities:
  - user-names, screen-names, location, profile photos, bios, interests, number of common followers, followings, users mentioned, and retweeted; *time difference* between creation dates, first and last tweets, outdated account

detects

90% of victim-impersonator pairs

80% of avatar-avatar pairs

at less than 5% false positive rate
## Classifying unlabeled doppelgänger pairs

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
<tr>
<td>detected victim-impersonator pairs</td>
<td>1,863</td>
<td>9,031</td>
</tr>
<tr>
<td>detected avatar-avatar pairs</td>
<td>4,390</td>
<td>4,964</td>
</tr>
</tbody>
</table>
Classifying unlabeled doppelgänger pairs

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
<tr>
<td>detected victim-impersonator pairs</td>
<td>1,863</td>
<td>9,031</td>
</tr>
<tr>
<td>detected avatar-avatar pairs</td>
<td>4,390</td>
<td>4,964</td>
</tr>
</tbody>
</table>
Classifying unlabeled doppelgänger pairs

<table>
<thead>
<tr>
<th></th>
<th>RANDOM DATASET</th>
<th>BFS DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>doppelgänger pairs</td>
<td>18,662</td>
<td>35,642</td>
</tr>
<tr>
<td>victim-impersonator pairs</td>
<td>166</td>
<td>16,408</td>
</tr>
<tr>
<td>avatar-avatar pairs</td>
<td>2,010</td>
<td>1,629</td>
</tr>
<tr>
<td>unlabeled pairs</td>
<td>16,489</td>
<td>17,605</td>
</tr>
<tr>
<td>detected victim-impersonator pairs</td>
<td>1,863</td>
<td>9,031</td>
</tr>
<tr>
<td>detected avatar-avatar pairs</td>
<td>4,390</td>
<td>4,964</td>
</tr>
</tbody>
</table>

*one year later 50% were suspended!*
Summary

• First study to characterize and detect identity impersonation attacks online

• Method to gather real-world large-scale data about impersonation attacks

• Beside celebrity impersonators and social engineering attacks there are doppelgänger bot attacks
  • Attackers target a wide range of users, anyone can be a victim!

• Method to automatically detect impersonation attacks online
Questions?
Backup slides
Features

- Victim-impersonator pairs have more similar profile attributes
- Victim-impersonator pairs have no social neighborhood overlap
- Bigger time difference between accounts creation date in victim-impersonator pairs
## Doppelgänger bot attacks: characterization

<table>
<thead>
<tr>
<th></th>
<th>Who are the victims?</th>
<th>Who are the attackers?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How popular?</strong></td>
<td>73 followers</td>
<td>60 followers*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*lower than victims, higher than random</td>
</tr>
<tr>
<td><strong>How influential?</strong></td>
<td>40% victims appear in lists</td>
<td>0% attacker appear in lists</td>
</tr>
<tr>
<td><strong>How old?</strong></td>
<td>October 2010</td>
<td>June 2013</td>
</tr>
<tr>
<td><strong>How active?</strong></td>
<td>181 tweets*</td>
<td>100 tweets*</td>
</tr>
<tr>
<td></td>
<td>*0 for random users, 20 for</td>
<td>higher numbers of retweets, favorite and followings</td>
</tr>
<tr>
<td></td>
<td>random users with one post</td>
<td>but not excessive</td>
</tr>
</tbody>
</table>