Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing

Peter Snyder* – Periwinkle Doerfler+ – Chris Kanich* – Damon McCoy+
Overview

- Doxing is a target form of online abuse
- Prior work is qualitative or on defensive
- We don't understand scope and targets of problem
- This work is the first qualitative, large scale measurement of doxing
Outline

• Problem area
• Measurement methodology
• Results and findings
• Discussion and conclusions
Outline

• Problem area
• Measurement methodology
• Results and findings
• Discussion and conclusions
What is Doxing? (1/2)

- Method of targeted online abuse
- Attackers compile sensitive information about the target
  - **Personal**: Name, addresses, age, photographs, SSN
  - **Relationships**: Family members, partners, friends
  - **Financial**: Work history, investments, criminal history
  - **Online**: Email, social network accounts, passwords, IPs
What is Doxing? (2/2)

- Information is compiled into plain text files
- Released "anonymously"
  - Text sharing sites (e.x. pastebin.com, skidpaste.com)
  - Online forums (e.x. 4chan, 8chan)
  - Torrents
  - IRC, Twitch, social networks, etc.
Full Name: █████  █████
Aliases: █████
Age: □
DOB: □/□/□
Address: □  □□□□□□□□□□□□ □□□□□ □□□□□□□□□□□□, □□□□□ □□□□□ // Confirmed
Mobile Number: +□ (□□□) □□□-□□□□ // Confirmed
Email: █████████@██████.██ // Confirmed

Illness: Asthma

ISP Records:
ISP: Rogers Cable // Previous
IP Address: □□.□□.□□.□□ // Previous

Parental Information:
Father: □  □  □□□□□
Age: □
Aliases) [Redacted]
Name) [Redacted]
DOB □/□/□
Address) [Redacted]
Cell Phone) [Redacted] – Sprint, Mobile
Caller ID) [Redacted]
Old Home Phone) [Redacted] – CenturyLink, Landline
Last 4 of Mastercard) [Redacted]
Emails) [Redacted], [Redacted]
Snapchat) [Redacted]
Twitter) @[Redacted]
Facebook) https://facebook.com/[Redacted], [Redacted]
Skype) [Redacted], [Redacted]
Doxing Harms
Frequency, Targets and Effects

• Prior work is based in qualitative or preventative / risk management approaches

• Research Questions:

1. How frequently does doxing happen?

2. What information is shared in doxes? Who is targeted?

3. What is knowable about the large scale effects and harms?

4. Are anti-abuse tools effective?
Outline

• Problem area
• Measurement methodology
• Results and findings
• Discussion and conclusions
General Measurement Strategy

• Find places online where doxes are frequently shared
• Train a classifier to determine how much activity is doxing
• Measure extracted doxes to determine contained information
• Watch the OSN accounts of doxing victims for abuse
Steps to Protect Victims

- Worked closely with IRBs; multiple rounds of study design
- Only recorded publicly available data
- Careful data storage / analysis methods: only recorded high level summary data
- Data protection best practices (key based encryption, single data store, strict access controls)
Dox Collection Pipeline

- Fully automated
- Single IP at the University of Illinois at Chicago
- Two recording periods:
  - Summer of 2016
  - Winter of 2016
Text File Collection

- Data recorded from
  - pastebin.com
  - 4chan.org (pol, b)
  - 8ch.net (pol, baphometh)
- API and scrapers
- Selected because:
  - "Original" sources of doxes
  - Anecdotal reputation for toxic behavior / doxing activity
Text File Classification

- Scikit-Learn, TfidfVectorizer, SGDClassifier

- Training Data:
  - Manual labeling of Pastebin crawl
  - "proof-of-work" sets
# Text File Classification

<table>
<thead>
<tr>
<th>Label</th>
<th>Precision</th>
<th>Recall</th>
<th># Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dox</td>
<td>0.81</td>
<td>0.89</td>
<td>258</td>
</tr>
<tr>
<td>Not</td>
<td>0.99</td>
<td>0.98</td>
<td>3,546</td>
</tr>
<tr>
<td>Avg / Total</td>
<td>0.98</td>
<td>0.98</td>
<td>3,804</td>
</tr>
</tbody>
</table>
Social Networking Account Extractor

- Extract social networking accounts
- Custom, heuristic-based identifier
- Example:
  - Facebook: https://facebook.com/example
  - FB example
  - fbs: example - example2 - example3
  - facebooks; example and example2
- Evaluated on 125 labeled doxes
<table>
<thead>
<tr>
<th>Platform</th>
<th>% Doxes Including</th>
<th>Extractor Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instagram</td>
<td>11.2</td>
<td>95.2</td>
</tr>
<tr>
<td>Twitch</td>
<td>9.7</td>
<td>95.2</td>
</tr>
<tr>
<td>Google+</td>
<td>18.4</td>
<td>90.4</td>
</tr>
<tr>
<td>Twitter</td>
<td>34.4</td>
<td>86.4</td>
</tr>
<tr>
<td>Facebook</td>
<td>48.0</td>
<td>84.8</td>
</tr>
<tr>
<td>YouTube</td>
<td>40.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>
Dox De-duplication

- Similar doxes, identical target
- Hash based comparison fragile to marginal updates
- Compare referenced OSN accounts
- ~14.2% of doxes were duplicates
Dox De-duplication

- Repeatedly visit referenced OSN accounts
- After 1, 2, 3, 7, 14… days
- Only record the status of the account:
  - public
  - private
  - inactive
- Single IP @ UIC
Manual Dox Labeling

• Randomly selected 464 doxes

• Manually label each dox to understand the contents.
  
  • Did it include name, address, phone #, email, etc.?
  
  • Age and gender of the target (if included)
  
  • Categorization of the victim
  
  • Categorization of the motive of attacker
    ("why I doxed this person…")
## Collection Statistics

<table>
<thead>
<tr>
<th>Study Period</th>
<th>Summer 2016</th>
<th>Winter 2016-17</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text Files Recorded</strong></td>
<td>484,185</td>
<td>1,253,702</td>
<td>1,737,887</td>
</tr>
<tr>
<td><strong>Classified as Dox</strong></td>
<td>2,976</td>
<td>2,554</td>
<td>5,530</td>
</tr>
<tr>
<td><strong>Doxes w/o Duplicates</strong></td>
<td>2,326</td>
<td>2,202</td>
<td>4,528</td>
</tr>
<tr>
<td><strong>Manually Labeled</strong></td>
<td>270</td>
<td>194</td>
<td>464</td>
</tr>
</tbody>
</table>
Outline

• Problem area
• Measurement methodology
• Results and findings
• Discussion and conclusions
Outline

• Results and findings
  • Doxing targets
  • Doxing perpetuators
• Effects on social networks
Doxing Targets
Victim Demographics

- Taken from the 464 manually labeled doxes
- Only based on data in doxes
- Harm prevention steps (e.g. not taking demographic data from OSN accounts)

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Age</td>
<td>10 years old</td>
<td>16.3%</td>
</tr>
<tr>
<td>Max Age</td>
<td>74 years old</td>
<td>82.2%</td>
</tr>
<tr>
<td>Mean Age</td>
<td>21.7 years old</td>
<td>64.5%</td>
</tr>
<tr>
<td>Gender, Female</td>
<td></td>
<td>16.3%</td>
</tr>
<tr>
<td>Gender, Male</td>
<td></td>
<td>82.2%</td>
</tr>
<tr>
<td>Gender, Other</td>
<td></td>
<td>0.4%</td>
</tr>
<tr>
<td>Located in USA</td>
<td></td>
<td>64.5%</td>
</tr>
</tbody>
</table>

(64.5% of 300 files that included address)
# Types of Data in Doxes

## Frequently Occurring Data

<table>
<thead>
<tr>
<th>Category</th>
<th># of Doxes</th>
<th>% of Doxes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>422</td>
<td>90.1%</td>
</tr>
<tr>
<td>Phone #</td>
<td>284</td>
<td>61.2%</td>
</tr>
<tr>
<td>Family Info</td>
<td>235</td>
<td>50.6%</td>
</tr>
<tr>
<td>Email</td>
<td>249</td>
<td>53.7%</td>
</tr>
<tr>
<td>Zip Code</td>
<td>227</td>
<td>48.9%</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>155</td>
<td>33.4%</td>
</tr>
</tbody>
</table>

## Highly Sensitive Data

<table>
<thead>
<tr>
<th>Category</th>
<th># of Doxes</th>
<th>% of Doxes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>48</td>
<td>10.3%</td>
</tr>
<tr>
<td>ISP</td>
<td>100</td>
<td>21.6%</td>
</tr>
<tr>
<td>Passwords</td>
<td>40</td>
<td>8.6%</td>
</tr>
<tr>
<td>Criminal Record</td>
<td>6</td>
<td>1.3%</td>
</tr>
<tr>
<td>CCN</td>
<td>20</td>
<td>4.3%</td>
</tr>
<tr>
<td>SSN</td>
<td>10</td>
<td>2.6%</td>
</tr>
</tbody>
</table>

*All numbers from 464 manually labeled doxes*
# Doxing Victims by Community

- Categorization of victim based on listed OSN accounts
- 16.2% of victims categorizable into 3 categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Criteria</th>
<th># of Labeled</th>
<th>% of Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hacker</td>
<td>2 or more OSN accounts on hacking sites (e.g. hackforums.net)</td>
<td>17</td>
<td>3.7%</td>
</tr>
<tr>
<td>Gamer</td>
<td>2 or more OSN accounts on gaming sites (e.g. twitch.tv, minecraftforum.net)</td>
<td>53</td>
<td>11.4%</td>
</tr>
<tr>
<td>Celebrity</td>
<td>Labelers recognized target independent of doxing (e.g. Donald Trump, Hillary Clinton)</td>
<td>5</td>
<td>1.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>75</strong></td>
<td><strong>16.2%</strong></td>
</tr>
</tbody>
</table>
Doxing Perpetrators
Doxer Motivations

- Categorization of doxers based on "why I did it" suffixes
- 28.4% of dox motivations categorizable into 4 categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Criteria</th>
<th># of Labeled</th>
<th>% of Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competitive</td>
<td>Demonstrating attacker's capabilities / victim's weaknesses</td>
<td>7</td>
<td>1.5%</td>
</tr>
<tr>
<td>Revenge</td>
<td>Because of doxee's actions against doxer (e.g. &quot;you cheated in counterstrike.&quot;)</td>
<td>52</td>
<td>11.2%</td>
</tr>
<tr>
<td>Justice</td>
<td>Because of doxee's actions against third party (e.g. &quot;you ripped off my friend&quot;)</td>
<td>68</td>
<td>14.7%</td>
</tr>
<tr>
<td>Political</td>
<td>Because of larger political goal (attacking KKK members or child pornographers)</td>
<td>5</td>
<td>1.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>132</strong></td>
<td><strong>28.4%</strong></td>
</tr>
</tbody>
</table>
Doxer Networks

- Looked for doxer networks based on "credit lines"
- ex: "by Alice and @Bob, thx to Charlie (@Charlie for SSN)"
- 251 aliases given, 213 twitter handles
- Undirected graph from doxes and twitter network
Doxer Networks

- 61 (of 251) aliases appear in cliques of 4 or more
- 34 Twitter accounts were private
Harms from Doxing
Effects on OSN Accounts

1. Are OSN accounts that are doxed more likely to become more private?

2. Does abuse filtering reduce the impact of doxing on OSN accounts?

Summer 2016
First recording period

Fall 2016
Facebook and Instagram add abuse filtering

Winter 2016
Second recording period
Effects on OSN Accounts

• Measure changes in status of OSN accounts after appearing in dox
• Public, private, inactive
• Compared against 13,392 randomly selected Instagram accounts

<table>
<thead>
<tr>
<th>Social Network</th>
<th># Doxes</th>
<th>% Doxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>983</td>
<td>17.8%</td>
</tr>
<tr>
<td>Google+</td>
<td>405</td>
<td>7.3%</td>
</tr>
<tr>
<td>Twitter</td>
<td>449</td>
<td>8.1%</td>
</tr>
<tr>
<td>Instagram</td>
<td>418</td>
<td>7.5%</td>
</tr>
<tr>
<td>YouTube</td>
<td>316</td>
<td>5.7%</td>
</tr>
<tr>
<td>Twitch</td>
<td>185</td>
<td>3.4%</td>
</tr>
</tbody>
</table>
## Doxed vs. Non-Doxed Accounts

<table>
<thead>
<tr>
<th>Account</th>
<th>Condition</th>
<th>% More Private</th>
<th>% More Public</th>
<th>% Any Change</th>
<th>Total #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instagram</td>
<td>default</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>13,392</td>
</tr>
<tr>
<td>Instagram</td>
<td>doxed, pre-filtering</td>
<td>17.2</td>
<td>8.1</td>
<td>32.2</td>
<td>87</td>
</tr>
<tr>
<td>Instagram</td>
<td>doxed, post-filtering</td>
<td>5.7</td>
<td>1.4</td>
<td>9.9</td>
<td>141</td>
</tr>
<tr>
<td>Facebook</td>
<td>doxed, pre-filtering</td>
<td>22.0</td>
<td>2.0</td>
<td>24.6</td>
<td>191</td>
</tr>
<tr>
<td>Facebook</td>
<td>doxed, post-filtering</td>
<td>3.0</td>
<td>&lt;0.1</td>
<td>3.3</td>
<td>361</td>
</tr>
</tbody>
</table>
## Doxed vs. Non-Doxed Accounts

<table>
<thead>
<tr>
<th>Account</th>
<th>Condition</th>
<th>% More Private</th>
<th>% More Public</th>
<th>% Any Change</th>
<th>Total #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instagram</td>
<td>default</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>13,392</td>
</tr>
<tr>
<td>Instagram</td>
<td>doxed, pre-filtering</td>
<td>17.2</td>
<td>8.1</td>
<td>32.2</td>
<td>87</td>
</tr>
<tr>
<td>Instagram</td>
<td>doxed, post-filtering</td>
<td>5.7</td>
<td>1.4</td>
<td>9.9</td>
<td>141</td>
</tr>
<tr>
<td>Facebook</td>
<td>doxed, pre-filtering</td>
<td>22.0</td>
<td>2.0</td>
<td>24.6</td>
<td>191</td>
</tr>
<tr>
<td>Facebook</td>
<td>doxed, post-filtering</td>
<td>3.0</td>
<td>&lt;0.1</td>
<td>3.3</td>
<td>361</td>
</tr>
</tbody>
</table>
# Doxed vs. Non-Doxed Accounts

<table>
<thead>
<tr>
<th>Account</th>
<th>Condition</th>
<th>% More Private</th>
<th>% More Public</th>
<th>% Any Change</th>
<th>Total #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instagram</td>
<td>default</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>13,392</td>
</tr>
<tr>
<td>Instagram</td>
<td>doxed, pre-filtering</td>
<td>17.2</td>
<td>8.1</td>
<td>32.2</td>
<td>87</td>
</tr>
<tr>
<td>Instagram</td>
<td>doxed, post-filtering</td>
<td>5.7</td>
<td>1.4</td>
<td>9.9</td>
<td>141</td>
</tr>
<tr>
<td>Facebook</td>
<td>doxed, pre-filtering</td>
<td>22.0</td>
<td>2.0</td>
<td>24.6</td>
<td>191</td>
</tr>
<tr>
<td>Facebook</td>
<td>doxed, post-filtering</td>
<td>3.0</td>
<td>&lt;0.1</td>
<td>3.3</td>
<td>361</td>
</tr>
</tbody>
</table>
Facebook Statues after Doxing
Facebook Statues after Doxing
Facebook Statues after Doxing
Facebook Statues after Doxing
Facebook Statues after Doxing

Facebook accounts that changed status, Pre-filtering (22.5%)

Facebook accounts that changed status, Post-filtering (1.7%)
Instagram Statues after Doxing

Instagram accounts that changed status, Pre-filtering (13.8%)

Instagram accounts that changed status, Post-filtering (5.0%)
Outline

• Problem area
• Measurement methodology
• Results and findings
• Discussion and conclusions
Using Data to Help Victims

- **Notification of doxing victims**
  "Have I Been Pwned" style service

- **OSN Account protection**
  Notify social networks of doxing, for defenses

- **Anti-SWAT-ing List**
  Additional information for law enforcement to evaluates

- **Anti-Abuse Policies From Dox Distributing Sites**
  Working with Pastebin to increase automated takedowns
Take Aways

• Automatic dox measurement and classification pipeline

• 1.7m text files, 4,328 doxes, manual labeling of 464

• First quantitative analysis of frequency, targets and contents of doxing online

• Measurement of harm of doxing, via OSN account change