Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing

Peter Snyder
University of Illinois at Chicago
Chicago, IL
psnyde2@uic.edu

Periwinkle Doerfler
New York University
Brooklyn, NY
pid207@nyu.edu

Chris Kanich
University of Illinois at Chicago
Chicago, IL
ckanich@uic.edu

Damon McCoy
New York University
Brooklyn, NY
mccoy@nyu.edu

ABSTRACT

Doxing is online abuse where a malicious party harms another by releasing identifying or sensitive information. Motivations for doxing include personal, competitive, and political reasons, and web users of all ages, genders and internet experience have been targeted. Existing research on doxing is primarily qualitative. This work improves our understanding of doxing by being the first to take a quantitative approach. We do so by designing and deploying a tool which can detect dox files and measure the frequency, content, targets, and effects of doxing on popular dox-posting sites.

This work analyzes over 1.7 million text files posted to `pastebin.com`, `4chan.org` and `8ch.net`, sites frequently used to share doxes online, over a combined period of approximately thirteen weeks. Notable findings in this work include that approximately 0.3% of shared files are doxes, that online social networking accounts mentioned in these dox files are more likely to close than typical accounts, that justice and revenge are the most often cited motivations for doxing, and that dox files target males more frequently than females.

We also find that recent anti-abuse efforts by social networks have reduced how frequently these doxing victims closed or restricted their accounts after being attacked. We also propose mitigation steps, such a service that can inform people when their accounts have been shared in a dox file, or law enforcement notification tools to inform authorities when individuals are at heightened risk of abuse.

CCS CONCEPTS

• **Networks → Online social networks**; • **Social and professional topics → Social engineering attacks**; **Identity theft**;

KEYWORDS

Doxing, Identity Theft, Online Abuse

ACM Reference Format:
Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing. In *Proceedings of IMC '17*. ACM, New York, NY, USA, 13 pages.
https://doi.org/10.1145/3131365.3131385

IMC '17, November 1–3, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5118-8/17/11…$15.00
https://doi.org/10.1145/3131365.3131385

1 INTRODUCTION

Doxing is an attack where a victim's private information is released publicly online [33]. While unsophisticated at a technical level, this type of attack can cause substantial harm to its victims. These harms range from harassing online comments to direct physical danger [21]. Doxing is one of a few cyberattacks that can cause direct, serious, and lasting harm to its victims.

Existing studies of doxing have used qualitative approaches [9, 17], or worked from a risk management perspective [23, 27]. While valuable, these previous efforts do not provide a broad, large-scale, quantitative measurement of doxing as an online phenomenon.

Quantitative measurements of the harm caused by doxing are vital, given the limited resources that exist for defending against cyberattacks. Comparing the outcomes of different harassment campaigns, or even different types of cyberattack, allows defenders to focus resources where they will do the most good protecting users.

This work fills a gap in our understanding of doxing by providing the first quantitative, large-scale measurement of this online harassment technique. We provide the following three contributions:

- A software pipeline to automatically detect doxes and extract mentioned online social networking accounts. We then monitored these social networking accounts for signs of harassment and changes to their privacy settings, to understand the impact of doxing at scale.
- A comprehensive analysis of the information shared in dox files, including what types of sensitive information are included, what the motives and targets of doxers are, and what networks can be identified among the attackers.
- Through a bit of serendipity, our measurement spans a period before and after new anti-harassment tools were deployed by a large OSN operator. We provide an analysis of whether

these anti-harassment techniques successfully protect doxing victims.

Finally, we discuss harm-mitigation strategies enabled by our automated dox-detection tools. The creation of automated tools for law enforcement and victim could mitigate doxing related harms, such as identity theft or Swatting.[1]

2 BACKGROUND

The origin of the term "dox" is unclear, but one common explanation for the term is as shortened form of the word "documents", as in "drop documents". The term first came into use in the 1990s [1], to describe humiliating or intimidating someone by linking online personas to sensitive personal information. Since then, doxing has transformed into a harassment tactic employed by high profile groups, such as Anonymous and gamergate. A simple internet search for "doxing tutorial" returns hundreds of results where different sites explain methods to find someone's name, address, email address, phone number, IP address, criminal history, and social network accounts. Additional details about these methods and tools can be found in previous studies [20, 27]. Most of these sources of information are cheap and can be used to quickly assemble a large amount of sensitive information about a victim.

The commercialization of doxing has made it an even easier form of harassment to conduct online. While attackers originally had to gather information about their targets themselves, recent dox-for-hire services have made the process cheap and easy for abusers. Dox-for-hire services compile information, such as the victim's name, address, email, and social networking accounts, for as low as $5 US. These dox files are then often distributed on web sites that allow anonymous posting and do not proactively remove harassing content.

There is a good deal of existing work about online harassment, both quantitative and qualitative. These studies come from the tech and education spheres, and have generally focused on underage victims [18, 19, 36]. The impact of gender in online harassment has also been studied, finding that women experience online harassment at higher rates, particularly in gaming communities [4, 25]. These studies generally survey potential victims to measure how frequently online abuse occurs, and then rely on in-depth interviews with a subset of victims to understand the impacts and mitigation strategies [10, 13].

Current research on doxing consists of anecdotal reports and qualitative studies [9]. A recent study of young women's experiences with online harassment provides a measurement on the frequency with which doxing occurs in general [37]. But, beyond this, very little quantitative information about the phenomenon exists. In the absence of this understanding, previous studies have investigated the tools and techniques published in doxing online tutorials and proposed self doxing as a strategy to understand and limit the potential effects of doxing [20, 27]. While this strategy is illuminating, it is costly and time consuming, and so does not scale to an understanding of doxing as an internet scale phenomina.

[1]Swatting is the deceptive practice of calling emergency services and convincing them to send polices to another person's address based on a false report of an emergency, such as a hostage incident.

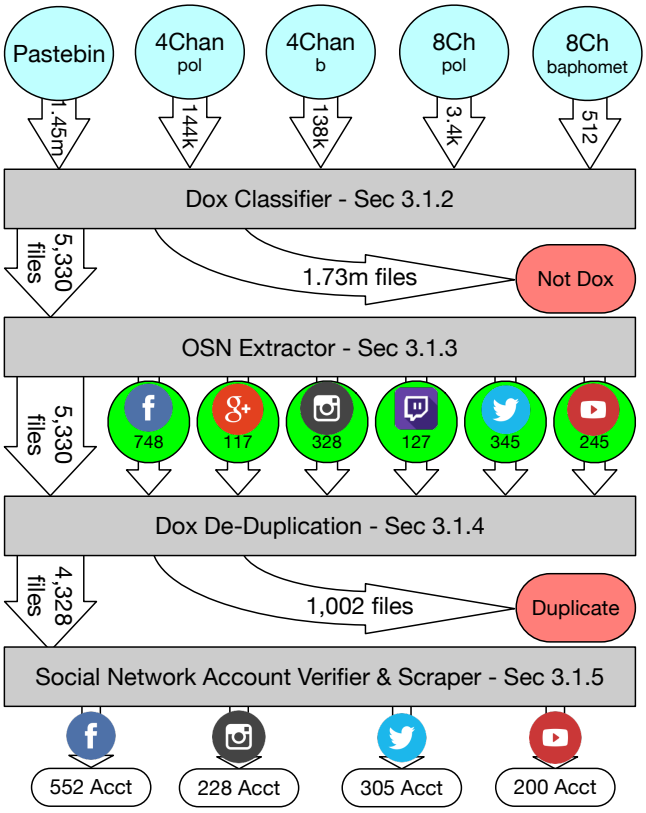

Figure 1: Diagram depicting the full pipeline for collecting documents from online text sharing sites, identifying dox files, identifying duplicates, and measuring online social networking account behavior.

Additionally, prior work has characterized the abuse ecosystems that exist specifically on 4chan's "politically incorrect" subforum [14], which we scrape for doxing detection purposes. We build on previous work not only by characterizing the abuse that accompanies doxing, but also the associated negative effects like closed or protected accounts.

There has been a large effort to design abusive speech filters [3, 5, 8, 15, 26, 30–32, 38, 40, 41] through traditional NLP methods, data mining methods, and deep learning based approaches. Some OSN platforms, such as Facebook, Instagram, and Twitter, have deployed algorithm changes and tools focused on filtering harassing posts [7]. Despite advertising the public availability of training data [8, 26], prior studies have not publicly released a corpus of messages. Wulczyn et al. [39] released a labeled corpus as of February 2017, but in initial evaluation we found that their classifier, which was trained on abusive messages from Wikipedia talk pages, was not effective for detecting abusive messages on Instagram or Twitter. Thus, we focus on using account privacy setting changes as a proxy for both the existence and effect of doxing related abuse.

Label	Precision	Recall	F1	# Samples
Dox	0.81	0.89	0.85	258
Not	0.99	0.98	0.99	3,546
Avg / Total	0.98	0.98	0.98	3,804

Table 1: Precision and recall figures of the TF-IDF classifier trained to identify dox files.

Label	% Doxes Including	Extractor Accuracy
Instagram	11.2	95.2
Twitch	9.6	95.2
Google+	18.4	90.4
Twitter	34.4	86.4
Facebook	48.0	84.8
YouTube	40.0	80.0
Skype	55.2	83.2
First Name	82.4	77.6
Last Name	82.4	62.4
Age	44.8	81.6
Phone	65.6	58.4

Table 2: Measures of the accuracy of the social network extractor, as compared with the success of applying the same extraction technique to other types of data.

3 METHODOLOGY

This section describes the methodology for our quantitative, large scale measurement of doxing. The section first describes how we collected and identified dox files, continues with how we determined the content of these dox files, and ends with how we measured the effect of doxing on online social networking accounts.

3.1 Dox Measurement Pipeline

We break the task of measuring doxes up into five steps: collecting text documents from popular text sharing websites, extracting likely doxes from the collected text files, labeling and extracting online social networking accounts from the dox files, identifying and removing duplicate dox files from further consideration, and finally repeatedly visiting referenced online social networking accounts to measure differences in their availability and privacy. Each step is described in the following subsections, and a diagram of the overall architecture is shown in Figure 1.

3.1.1 Text File Collection. The first step in our pipeline is to capture a large set of text files, distributed on the internet, that might be dox files. Doxers use a wide variety of methods to distribute and publicize the dox files they generate, ranging from onion sites, torrents, IRC and anonymous text sharing websites like pastebin.com.

This work considers text posted on the sites `pastebin.com`, `4chan.org` and `8ch.net` for two reasons. First, these sites host a large amount of dox files, giving us a large amount of material and study. Second, there appears to be few dox files that appear elsewhere that do not *also* appear on `pastebin.com`, `4chan.org` and `8ch.net`. We manually investigated other locations where doxes are shared, such as torrents of previous dox files shared on hacking oriented forums, onion sites designed to share doxes, or other websites where text files are anonymously shared[2], and found that these other venues generally host copies of doxes already shared on `pastebin.com`, `4chan.org` and `8ch.net`.

This work covers data collected during two periods. First, we collected all text files posted to pastebin.com for the six week period from 7/20/2016 to 8/31/2016. All files were collected from a single IP located at a university, using a paid API provided by pastebin.com. Second, we collected all text files posted to pastebin.com, all postings to the "pol" and "b" forums of 4chan.org, and all postings to the "pol" and "baphomet" forums of 8ch.net during the seven week period of 12/19/2016 to 2/6/2017.

3.1.2 Dox Classifier. The second step in our pipeline was to filter down the hundreds of thousands of text files collected from `pastebin.com`, `4chan.org` and `8ch.net` and extract only the dox files for further consideration. We built a classifier based on the open-source scikit-learn package [28]. Using this system, we first transformed each labeled training example into a TF-IDF vector (using the system's `TfidfVectorizer` class), and then built a stochastic gradient descent-based model using the system's `SGDClassifier` class, with 20 iterations to train our model. With the single exception of specifying that the SGD classifier use 20 training passes, we used the default arguments and parameters for both classes, as defined in version 0.17.1 of scikit-learn.

This default configuration does not remove stop words from the texts. As a single pre-processing step, we transformed HTML version of postings left on `4chan.org` and `8ch.net` into plan text versions using `html2text` [34], which replaces HTML markup with semantically equivalent plain-text representations (e.x. changing `<ul>`, `<ol>` and `<li>` tags in an HTML fragment to indented, newline separated text strings in a plain-text document).

Our labeled data is from two sources. Our negative labeled examples are from a random crawl of pastebin.com. We pulled several thousand text files from pastebin.com, manually examined each one to make sure it was not a dox file, and used the remaining 4,220 files as our negative labeled set.

The positive (dox) labeled examples came from two sources. First, we pulled examples of doxes from "proof-of-work" sets released by dox-for-hire services, who release archives of their previous work as an example of what they can do for future clients. These "proof-of-work" sets are released through torrents and archive sites. Our second set of positive training examples came from the small number of doxes found in the previously mentioned random crawl of pastebin.com. In total, our positive (dox) set consisted of 749 files.

To evaluate the effectiveness of our classifier, we split our labeled data into a randomly selected two-thirds training set, one-third evaluation set. Table 1 gives the results of our evaluation. As the table shows, our classifier performs well, and is slightly more likely to make false positive errors than false negative ones.

3.1.3 Online Social Network Account Extractor. The third step in our dox measurement pipeline was to programmatically extract references to online social networking accounts from the dox files. We then built a realtime, automated method of monitoring these

[2]We have omitted links to these pages to avoid publicizing these doxing sites.

accounts for signs of possible abuse. We also used the extracted online social networking account references as unique identifiers to de-duplicate dox files.

Dox files are semi-structured. While it is generally easy for humans to identify the types of information contained in a dox file, it is not trivial to do so programmatically. Consider the following examples of a Facebook account **example** being included in a dox file:

(1) Facebook: `https://facebook.com/example`
(2) FB example
(3) fbs: example - example2 - example3
(4) facebooks; example and example2

We programmatically extracted the online social networking accounts from this semi-structured data by first randomly selecting 125 dox files from the positive-label set described above. We then hand labeled each file, noting the location and value of each online social networking account.

We then built a text extractor that attempted to match this hand labeling, using a mixture of statistical and heuristic approaches. Table 2 shows the accuracy of our extractor. As the table shows, we were able to programmatically extract online social networking account references with a high degree of accuracy.

3.1.4 Dox De-Duplication. The fourth step in our dox measurement pipeline was to identify doxes targeting victims already targeted by previous dox files. This was done to avoid double counting or otherwise affecting the results of this work.

In some cases, finding identical doxes was simple. We removed 214 (3.9%) dox files from further consideration by comparing new doxes against the bodies of previously seen dox files.

Other duplicate doxes were more difficult to identify. Many doxers posted the same information several times, but made non-substantive changes between versions. For example, some dox authors would include a posting timestamp, others would re-paste the dox file with minor formatting changes (such as to an ASCII-art insignia), while still others would add small "update" sections describing how the victim has reacted to being doxed so far.

In all of these cases, the same dox target is being described, with the same significant information. To remove these near-duplicate dox files, we compared the online social networking account identifiers extracted in step three of our dox measurement pipeline. If a dox file contained all of the same online social networking accounts as a previously seen dox, we treated it as a duplicate. We saw no instances of dox files which had overlapping but non-identical sets of online social networking accounts.

We removed 788 (14.2%) more dox files from further consideration using this technique. In total, we identified 1,002 dox files, or 18.1% of dox files, as duplicates, or targeting a victim already target in our dataset.

3.1.5 Online Social Network Account Scraper. The final step in our dox measurement pipeline was to monitor the online social networking accounts that were referenced in the dox files for changes in their openness, or status.

We visited each referenced online social networking account several times over the study period. Each time we checked to see if the account was in a public, private, or deleted/disabled state. For accounts that were public (i.e. had content that we could visit without any social ties to the measured account), we also recorded the text of the public posts the account owner had made, and the text of any comments that had been left on those posts. To avoid further harming the privacy of doxing victims, we did not record any further information listed on each online social networking account, such as a date of birth, post address, or a email address provided by the account holder. We only recorded the status of the account, and the text of public posts and comments.

We measured each online social networking account several times during the study period; immediately when the dox was observed on a text sharing site, and then again one, two, three and seven days after the initial observation, and then every seven days after that. Measurement points varied slightly from this schedule because of the load-balancing and queuing steps in our pipeline, but rarely deviated more than a day.

All recordings of online social networking accounts were made from a single IP address located at the University of Illinois at Chicago.

3.2 Labeling Dox Content

In order to understand the types of data included in dox files, we manually labeled 464 doxes randomly selected from the 5,530 text files our classifier identified as being a dox file. We noted the types of demographic information included about each dox target, along with any information we could glean about the party performing the dox (e.g. a given online alias).

Where possible, we categorized each doxing victim into one of several broad categories, such as gamer, hacker, or celebrity, based on the types of social accounts associated with victim[3]. A dox file stating that target maintained a large number of accounts on video-gamed based websites would be classified as a gamer, while a dox file that indicated the target maintained sites on many programming and hacking sites would be categorized as a hacker.

Finally, we noted the motivation for the party releasing the dox when possible. Many dox files included a "why I doxed this person" pre-or-postscript, giving a usually brief[4] description of why the person was targeted. Motivations ranged from political to competitive, and are described in more detail in section 5.3.1.

3.3 Ethical Considerations

We took several steps to protect the privacy of the doxing targets included in this study. It was a top priority to avoid causing additional harm, given the sensitive nature of the data collected and analyzed in this study.

First, we only collected and analyzed data that was already publicly released, and were careful to not combine data sources in a way that would further identify the doxing targets. For example, while we could have better understood the demographics of doxing targets by collecting demographic data from linked online social

[3]To avoid further harming each doxing victims' privacy, beyond the harm caused by the doxing itself, we did not visit any links observed in the dox files, with the exception of the popular online social networking accounts discussed in Sections 3.1.5 and 6, where we only recorded the status of the account, and the text of publicly shared comments and postings.

[4]Sometimes, this explanation was extremely verbose: one dox contained a multi-paragraph essay, detailing the doxer's multi-year history with the dox target, along with a litany of supposed wrongs.

networking accounts, or further measured the accuracy of the data in dox files by comparing it against other publicly available data source, we did not feel doing so would be ethical. These actions would have amounted to continuing the work of the doxers, and would have run the risk of further harming the victims.

Similarly, we carefully designed our data collection and storage systems to avoid further identifying the doxing victims. We did so to avoid turning our database into a new, highly sensitive, centralized source of highly identifying information. With the exception of the referenced online social networking accounts, we did not extract or store any information taken from the doxes; we only stored non-identifying, aggregate counts of the data. For example, instead of creating a "zipcode" column in our database, we only recorded (in a different datastore) whether a dox file contained a zip code. The end result is that our dataset would not aid an attacker in learning about the doxing-targets anymore than if they downloaded the already publicly available doxes themselves.

Institutional Review. When beginning this project, we believed that this analysis of publicly available information was not human subjects research. Upon later discussion with our institution's IRB, we discovered that this collection and analysis of public data did qualify as human subjects research. Thus, we applied for and received IRB approval and restarted data collection and analysis. Due to the unique property of the original data (i.e. that it was collected before Instagram and Facebook deployed new anti-harassment tools), we further requested an allowance from the IRB to use that data in this analysis. They approved our request to perform this analysis because the data was collected with the approved protocol (before said protocol was approved), but stipulated that we must explicitly state that said data was not collected with approval of the IRB.

4 VALIDATION

An underlying assumption of this work is that the information included in the measured dox files is correct, and that a person's information is actually being disclosed. In order to attempt to verify that this assumption is correct, we attempted to validate the information included in the dox files in three ways, described in the following subsections.

Each of these methods are an attempt, as guided by our university's IRB, to check the accuracy of the information included in each dox file without further harming the privacy of the dox target.

4.1 Validation by IP Address

One way we validated the accuracy of the collected dox files was to see if they were internally consistent. We did this by sampling from the dox files that included both the target's IP address *and* the target's postal address. We then geo-located the IP address, to see if it was close to the given postal address. The more frequently these two pieces of information matched, the more we treated it as a signal that the data was internally consistent.

We found that in most cases, the IP and postal addresses were located near each other, suggesting that both pieces of data were accurately describing the same person.

We performed this validation by randomly selecting 50 doxes that included an IP address. We then took the subset of those 50 dox files and removed those that did not also include a postal address. This left 36 doxes that had both IP and US postal addresses. Of these 36 remaining doxes, 32 had a close match between the IP address's geo-location (i.e. the postal address and the IP's address were in same state, province or region). Three records were significantly different (i.e. the IP address resolved to an ASN in a far away state or country than the listed postal address) and one remaining was ambitious (the IP's ASN was in a different, but adjacent, state).

We note here a limitation of this approach. It is possible that a doxer could use one piece of information to derive the latter (i.e. they could select a random IP address, geo-locate it themselves, and then put the resulting postal address in the dox as well). If this were a frequent occurrence, it would destroy the significance of this internal validation measure.

We believe this is not the case though. Of the 36 cases where the geo-located IP address was close to the given postal address, in only 4 cases did the two match exactly. In the other 32 cases, the resolution of the postal address was either greater than the geo-located address (i.e. the postal address included a street address, or some other detail that was not available from geolocation), or the two addresses were in different, but near-by cities.

Type	# of Files	# Deleted	% Deleted
Dox	1,122	144	12.8
Other	483,063	20,501	4.2

Table 3: Comparison of the number of dox versus non-dox pastebin.com posts that were deleted one month after being posted, for all pastebin.com posts made from 7/20/2016 to 8/31/2016.

4.2 Validation by Post Deletion

A second way we validated the dox files was by measuring how many dox files had been deleted by pastebin.com, either through abuse reports or other means. We found that the files our system labeled as doxes were more than three times more likely to be removed from pastebin.com than the average file one month after being pasted. This suggests that the files our system labeled as dox files were frequently, if not overwhelmingly, troublesome to at least one other internet user, who requested their deletion.

Table 3 presents the details of this comparison. As the table shows, text files pasted to pastebin.com and identified by our system as being dox files were over three times more likely to be deleted from pastebin.com within one month of being pasted. pastebin.com provides three methods for removing a file: 1. Files can be deleted by the party posting the file; 2. files can be given an deletion-date when pasted; and 3. files that are reported to pastebin.com as being abusive are deleted by pastebin.com after review.

We expect that it is unlikely that our dox-trained classifier is identifying files that fall under the first two categories above. We expect that the difference in deletion rates is evidence that our dox-identifier is identifying files that are *also* being reported as abusive by other parties. This, in turn, leads us to believe the data in the identified dox files is accurate data (otherwise parties would not have an incentive to report it as abusive and harmful).

4.3 Validation by online social networking accounts

Finally, we believe the difference between the Instagram accounts found in dox files and Instagram accounts in general (discussed in section 6.2.2) is further evidence that the text files our system identifies as doxes contain information at least accurate enough to be damaging to the target. If the data mentioned in the identified dox files was meaningless or random, and did not accurately describe someone, it seems less likely that there would be as large an effect in the referenced online social networking accounts (either in terms of account closure or account privacy setting modification).

We recognize as a limitation the possibility that observers of dox files attack any online social networking accounts listed in a text file. Were that the case, the observed status changes on doxed online social networking accounts would not be incompatible with the dox files containing inaccurate information. However, we believe that the magnitude of the observed affect on online social networking accounts, as well as the information discussed in the previous two subsections, makes it more likely that the identified dox files contain accurate information.

5 DOXERS & DOXING VICTIMS

This section presents measurements and information about what our data collection pipeline reveals about doxing victims, what types of information is shared about the victims by doxers, and what we can determine about the motives and connections between the parties carrying out the doxing attacks.

5.1 Collection Statistics

Data for this work was collected during two periods, a six week period during the summer of 2016, and a seven week period during the winter of 2016-17. These combined collection periods yielded 1,737,887 posts from `pastebin.com`, `4chan.org` and `8ch.net`, 5,530 of which our classifier identified as being dox files. Table 4 gives the exact numbers of files recorded from each source, during each recording period.

5.2 Doxing Victims

5.2.1 Victim Demographics. Using the manual labeling methods described in section 3.2, we categorized the types of information in dox files. This allowed us to learn the general demographic traits of the targets of doxing.

For these measurements, we only recorded the information included in each dox file. To avoid further harming the privacy of the doxing targets, and to comply with the ethical guidelines provided by our institutional review board, we did not attempt to collect any further demographic information about the doxing targets (for instance, by recording demographic information included on linked online social networking accounts).

Table 5 includes general demographic details on the doxing targets. As the table shows, we observed doxing against a wide age range, with more doxes targeting males than other genders. The majority of the doxing targets lived in the United States, though this could partially be because pastebin.com, 4chan, and 8ch are based in the United States and primarily in English.

5.2.2 Disclosed Sensitive Details. We also measured how often other less general demographic details appeared in dox files. Again, in order to avoid further harming the targets of the doxes, and to follow the ethical guidance of our institutional review board, we are not including the specifics of these demographic details, only the frequency with which dox files included each category of demographic data.

As Table 6 shows, dox files frequently included age (often with the precision of a specific date of birth), the real-world name of the dox target (as opposed to an online alias), and information about where the target was located geographically (often with zip-code level precision). Doxes often included information about the target's family members.

There were also types of information in dox files that appeared less frequently, but which had the potential to be very harmful or identifying when they did appear. A representative sample of these items are included in the bottom half of Table 6.

5.2.3 Doxing Victims by Community. Where possible, we classified the types of internet users that are targeted by doxing attacks, based on other types of accounts listed in the dox file. This classification provides some information about which communities are targeted by doxing.

We were able to classify 24.7% of our manually labeled doxes into one of three categories, using the methodology discussed in Section 3.2. The results of this classification are presented in Table 7.

One recurring category of dox targets we noticed were **hackers**, or individuals who maintained accounts on websites, forums, and other web-communities associated with hacking and cybercrime. For example, if a dox file included a link to someone's `hackforums.net` account, or someone's handle on a hacking-related IRC channel, we treated it as an indication that this user spent time in such internet communities. Users with more than two such hacking or cybercrime-related accounts were labeled as a **hacker**.

A second recurring category of dox targets were **gamers**, or web users who maintained multiple accounts on video game enthusiast and streaming communities. Some examples of such communities include `twitch.tv` or `minecraftforum.net` (a website popular with people who play the game Minecraft). If a dox included more than two such gaming-related accounts, we labeled the target as a **gamer**.

A third recurring category of dox-target we observed were **celebrities**, or people who are well known independent of doxing. Examples of such dox-targets include presidential candidates, movie stars, and heads of large companies. If a dox-target was known to any of our dox-labelers, we labeled the dox target as a **celebrity**.

5.3 Doxing Perpertators

This subsection describes what our data set reveals about the people who commit doxing attacks. The first subsection describes the revealed motivations of the doxing attacks in our data set. The second subsection describes what we were able to learn about the networks and connections between parties carrying out doxing attacks.

5.3.1 Motivation of Doxers. Using the methodology discussed in Section 3.2, we categorized doxes by the stated motivations of

Study Period	7/20–8/31/2016	12/19–2/6/2017	Total
Text files recorded	484,185	1,253,702	1,737,887
Classified as a dox	2,976	2,554	5,530
Doxes without duplicates	2,326	2,202	4,528
Doxes manually labeled	270	194	464

Table 4: Statistics regarding text files collected from `pastebin.com`, `4chan.org` and `8ch.net`.

Min Age	10
Max Age	74
Mean Age	21.7
Gender (Female)	16.3%
Gender (Male)	82.2%
Gender (Other)	0.4%
Located in USA*	64.5%

Table 5: Demographic details about the targets of doxing, based on the 464 doxes posted to `pastebin.com`, `4chan.org` and `8ch.net` that were manually labelled. *"Located in USA" is given as a percentage of the 300 dox files that included an address.

Demographic Category	# of Doxes	% of Doxes
Address (any)	422	90.1
Phone Number	284	61.2
Family Info	235	50.6
Email	249	53.7
Address (zip)	227	48.9
Date of Birth	155	33.4
School	48	10.3
Usernames	186	40.1
ISP	100	21.6
IP Address	187	40.3
Passwords	40	8.6
Physical Traits	12	2.6
Criminal Records	6	1.3
Social Security #	10	2.6
Credit Card #	20	4.3
Other Financial Info	41	8.8

Table 6: Counts of the number and percentage of dox files that contain different categories of demographic information (of the 464 manually labeled).

Category	# of Doxes	% of Labeled Doxes
Hacker	17	3.7
Gamer	53	11.4
Celebrity	5	1.1
Total	75	16.2

Table 7: Counts of the number of doxees that could be assigned to a category, based on the other information included in the dox file. A more detailed description of each label is provided in section 5.2.3.

Motivation	# of Doxes	% of Labeled Doxes
Competitive	7	1.5
Revenge	52	11.2
Justice	68	14.7
Political	5	1.1
Total	132	28.4

Table 8: Counts of the number of doxes where a motivation for the doxing could be inferred from the dox file. A description of the meaning for each label is provided in section 5.3.1.

the doxer in the 19.2% of cases where a motivation could be inferred from the text of the dox file. Table 8 includes these categorizations.

We identified four general motivations for doxing. Some doxers gave a **competitive** motivation for attacking their victim, such as wanting to demonstrate their "superior" abilities, or demonstrating that a target claiming to be "un-doxable" was vulnerable.

Another common motivation was **revenge**, or the doxer attacking because of something the target had done to the doxer. Examples of **revenge** motivations included the doxee "stealing" a significant other from the doxer, or the doxee being an "attention whore" in an online forum or chat.

A third recurring motivation was **justice**, or the doxer attacking the doxee because the doxee had previously done something immoral or unfair to a third party. This is different from a **revenge** motivation, where the harm being "avenged" is committed against the doxer. Examples of **justice** motivated doxings include targets who were alleged to have scammed other people in an online forum, or who worked with law enforcement.

A fourth motivation we observed was **political**, or doxing in support of a larger goal than simply targeting individuals. Examples of **political** doxes included de-anonymizing KKK members, suspected child-pornography trading groups, or people working in industries that the doxers considered to be abusive to animals.

5.3.2 Doxer Networks. We also used our dataset to learn about the relationships between doxers. We first looked into whether we could find connections between the parties that compiled and "dropped" the doxes, and second whether we could identify connections between the parties abusing the online social networking accounts of the doxing victims.

Cliques in Doxing Credits: We first attempted to identify networks of doxers based on the "credits" included in dox files. These "credits" are included *by the doxing party* and mention the aliases of the doxers or collaborating parties for bragging, reputation or other reasons. These "credits" also sometimes include the Twitter handles of the doxers. A "credit" might look like the following: "dropped by

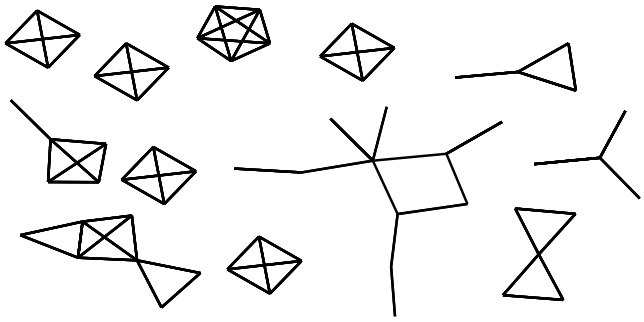

Figure 2: Graph depicting all cliques of doxers in our dataset, consiting of at least four doxers (61 of 251 doxers).

DoxerAlice and @DoxerBob, thanks to Charlie (@DoxerCharlie) for the SSN info".

We built our graph of doxer networks in three steps. First, we created nodes for the 251 doxers mentioned in "credits" throughout our collection of doxes. Second, we created edges between each pair of doxes that appeared in a "credit" together. Third, for the 213 doxers that were associated with Twitter handles in doxes, we also created edges for doxers who followed each other on Twitter (our edges are un-directed).

Figure 2 shows the results of this graph. This figure includes all cliques of at least 4 nodes, or doxers. This graph represents 61 of the 251 identified in dox "credits", with the largest observed clique consisting of 11 doxers. We expect that this graph is only a subset of the actual graph. For example, doxers might have multiple aliases, and many measured Twitter accounts (34) were private at the time of measurement.

Despite these limitations, we include this analysis both for completeness (as a hopefully helpful reference point for other researchers interested in the approach), and because we believe that this kind of network analysis, even if incomplete, could be a useful input to classifier systems and other defenses aimed at flagging potentially harmful content online.

Online Social Network Comments: We also looked for connections between doxers by examining the comments left on online social networking accounts mentioned in dox files. We recorded 33,570 comments left on the accounts of doxing victims, and looked for users who left comments on multiple doxed accounts. Such commenters might be evidence of doxers commenting on their victims' accounts.

We did not find any evidence of such doxing-related commenters. Of the 9,792 commenters who left a comment on a doxing-associated online social networking account, we did not find any commenters that left comments on multiple accounts.

6 EFFECTS ON SOCIAL NETWORKS

6.1 Collection Statistics

We identified which social networks most frequently appear in dox files. We generated these counts using the account extractor described in section 3.1.3.

Of the six measured online social networks, Facebook accounts were included most frequently in dox files, followed by Google+. We suspect that this is because these two social networks often display more personal information about account holders (e.g., Facebook asks users for their job, date of birth, home town) and allow users to explicitly identify social relationships between users. For example, Facebook allows users to identify "friends" as parents, or siblings, or friends, while Twitter has only the "follower" relationship. Facebook also has a real names policy which states that users who provide fake names or pseudonyms will have their accounts deactivated. [5]

Social Network	# Doxes	% Doxes
Facebook	983	17.8
Google+	405	7.3
Twitter	449	8.1
Instagram	418	7.5
YouTube	316	5.7
Twitch	185	3.3

Table 9: Counts of the number of dox files that included a reference to each major online social network.

Google+ had an even more restrictive real name policy that they rescinded in 2011 in large part due to issue of personal safety [11]. We expect that the larger amounts of personal information displayed by Facebook and Google+ means that these sites are used as data sources for doxers when compiling dox files, while other social networks (like Twitter or Instagram) are simply additional details to include in dox files.

6.2 Differences Between Doxed and Non-Doxed Social Networking Accounts

We also measured how being doxed affected the online social networking accounts of the dox targets. We found that online social networking accounts referenced in dox files are more likely to be closed or made more private than the average online social networking account.

6.2.1 Measured Networks. The following subsections present our measurements of Instagram, Facebook, Twitter and YouTube accounts found in dox files. We chose these online social networks for different reasons.

We measured Instagram accounts because Instagram's underlying user id field is monotonically increasing, thus allowing us to generate a random sample of all registered Instagram users. We use this sample as a control group to infer the amount of account closure that happens in the absence of detected doxing. Because Instagram claims over 600 million active users and doxing is a relatively uncommon occurrence, we believe that this sample will be sufficiently likely to be free of doxed accounts.

We recognize that this is an imperfect proxy for "typical" Instagram accounts, since some of these accounts could be abandoned or mostly inactive. We nevertheless used this random sampling technique because we believe it to be the best available proxy for a "typical" Instagram account, given our vantage point.

We initially considered trying to only compare "active" Instagram accounts against the Instagram accounts found in the dox files. However, we ultimately decided against this approach for two reasons. First, many of the Instagram accounts referenced in the

[5] https://www.facebook.com/help/112146705538576

dox files appeared to have low-to-no activity, so comparing the dox-referenced Instagram accounts against only "active", "typical" Instagram accounts would not have been a like comparison. Second, we briefly considered developing an "activity" metric that we could use to compare only active doxed accounts against active non-doxed accounts. However, we decided that developing this metric was beyond the scope of our goal in this work. We recognize this as a limitation in the following analysis, and suggest this as a possible area for future work.

Despite this limitation, we believe the dramatic difference in behavior between the randomly sampled accounts and the doxed accounts provides an imperfect but still useful measure of how accounts change when they are doxed.

We measured Facebook accounts because they were the type of online social networking account most frequently included in dox files. Though we were not able to build a model of the "typical" Facebook account to compare against, we measured Facebook accounts because of the network's popularity among doxers (and on the web in general).

Finally, we measured YouTube and Twitter accounts because of their popularity online, and because of the frequency with which they appeared in dox files.

6.2.2 Increased Account Closure and Privacy. We found a substantial decrease in the openness [6] of all online social networking accounts that were referenced in dox files, with the exception of YouTube accounts (as shown in Table 10). We attribute this to doxed account holders attempting to limit the harm the attackers can inflict. Unfortunately, by closing their account or making it more private doxing victims are becoming more socially isolated, which is another form of harm.

Instagram accounts referenced in dox files were much more likely to become private or to be closed than non-doxed Instagram accounts both before and after harassment filtering. We use changes in account privacy settings as a measure for the social isolation consequences of doxing. Instagram accounts referenced in dox files, across both measurement periods, were 920% more likely to change their privacy settings (in a more private or a more public direction) at least once during the measurement period than random accounts. Doxed accounts were 11,700% more likely to be more private at the end of the measurement period than random accounts. Doxed accounts change privacy settings, and particularly make their accounts more private at rates which are much higher than non-doxed accounts. P-values on both comparisons are asymptotically zero.

Though we do not have background numbers for Facebook or Twitter accounts to compare against, 10% becoming more private over the relatively short duration of our measurement periods seems unlikely to be typical of Facebook or Twitter accounts in general.

Unexpectedly, we also observed that doxed online social networking accounts are more likely to become *more* public than typical online social networking accounts. Our vantage point (only measuring social network activity that is publicly available) did not allow us to determine what specifically caused this increased openness. One possibility is that the increased account openness is a result of accounts being taken over by attackers[7]. Another possible explanation is that some accounts were closed or made more private before our first measurement, and we only captured the victim reopening the account later on. This is possible since many doxes are reposted, and we cannot know when a dox was originally publicly posted.

6.3 Effectiveness of Abuse Filters

Several social networks recently deployed anti-abuse tools, to protect their users from harassment online. We were able to measure the effectiveness of these filters in protecting doxing victims by comparing the status (public, private or inactive) of online social networking accounts attacked by doxes before and after networks deployed these anti-abuse efforts (e.g. [35]). Comparing the changes in "open-ness" in accounts before and after these filtering efforts allowed us to measure the effectiveness of these anti-abuse efforts.

Our initial effort to measure the effectiveness of abuse filtering was to measure the changes in the number of abusive comments left on social networking accounts that had appeared in dox files, using a variety of different abuse detection methods [2, 12]. We ended up not pursuing this approach for multiple reasons, including the difficulty of determining what constitutes abuse from community to community (and current models' inability to account for such differences), and the difficulty of hand labeling the abusiveness of comments without experts in each communities' social norms.

We instead decided to use the more objective measurement of changes in account status between our pre-and-post filtering datasets. We treat an account moving from being public to private (or closed all together) as indicating that the account may be receiving comments that the account holder judges as being abusive. Decreasing numbers of such accounts is partial evidence that networks' anti-abuse methods are better protecting their users from abuse.

For this measurement, we measured the account status of online social networking accounts that were included in a dox file, using the method described in Section 3.1.3, for two weeks. We measured whether the account was **public** (i.e., visible to users who had authenticated with the site, but who did not have any network connection to the account), **private** (i.e., visible to only to the subset of users who had some network relationship with the account) or **inactive** (i.e., the account was closed, deleted, or otherwise inaccessible).

Across all social networks, we observe that users change their account status quickly after their information is include in a dox file. 90.6% observed "more-private" status changes occurred within the first seven days of the dox file being shared online; 35.8% of these "more-private" account changes happen within 24 hours of the account appearing in a dox file.

6.3.1 Facebook. In August of 2016, Facebook altered their algorithms to give greater weight to content that users positively interacted with on their platform [29]. The primary goal of this change was to decrease the visibility of "clickbait" news articles. A side effect of these changes is that harassing and abusive content

[6] We define a decrease in the openness of an online social networking account as privacy settings being updated to make less information public or deleting the account.

[7] Anecdotally we manually found two victims' accounts that had clearly been compromised and defaced. However, we could not create an automated compromised account detector.

Account	Condition	% More Private	% More Public	% Any Change	Total #
Instagram	Default	0.1	0.1	0.2	13,392
Instagram	Doxed (pre filter)	17.2	8.1	32.2	87
Instagram	Doxed (post filter)	5.7	1.4	9.9	141
Facebook	Doxed (pre filter)	22.0	2.0	24.6	191
Facebook	Doxed (post filter)	3.0	<0.1	3.3	361
Twitter	Doxed	6.9	2.6	10.5	305
YouTube	Doxed	0.5	0.0	1.0	200

Table 10: Comparison of the statuses of online social networking accounts that were observed in dox files compared to typical accounts. (This table presents status changes that occurred at any point during the measurement period, in contrast to Figure 3, which depicts only status changes that occurred within 14 days of the doxing attack).

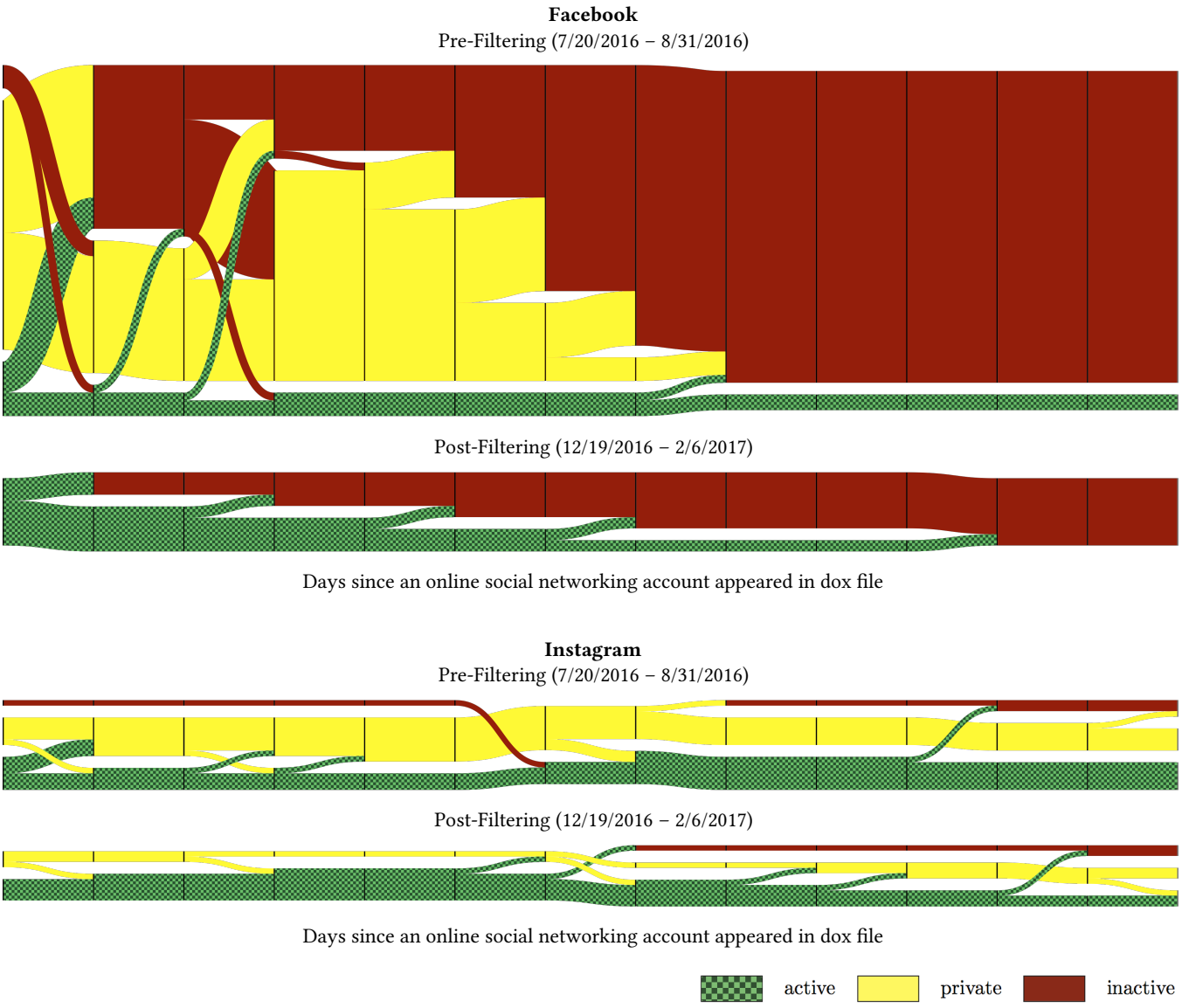

Figure 3: Changes in the status of online social network accounts, starting from when the account was mentioned in a dox file (x = 0) and ending after two weeks (x = 14).

will likely be hidden more often to users, since this kind of content is unlikely to generate positive interactions with Facebook users.

We measured the effectiveness of this approach by comparing how frequently doxed Facebook users made their accounts more private before and after Facebook deployed this change to their algorithms.

As shown in Table 2, Facebook was the most popular online social network in dox files. It was also the online social network that had the most (in absolute numbers) changes in the status of users' accounts after those accounts appeared in a dox file. Figure 3 shows the changes in status for the subset of Facebook accounts that changed their status *within two weeks of being doxed*. Note that this measurement window differs from that in Table 10, which depicts status that occurred at any point during the measurement period, not just within the two weeks following the doxing attack.

Each point along the x-axis depicts the number of days after being doxed. The left most point is the status when the dox was recorded, the second tick is the status of the account one day after the account appeared in a dox file, etc. Green, yellow and red stripes in the graph represent accounts that were public, private or inactive (respectively) at the relevant number of days after the doxing attack. Transitions from one color to the other depict a quantity of accounts changing status.

The top graph in Figure 3 depicts the 43 (22.5%) Facebook accounts in our dataset that changed their status at least once in the two weeks after being doxed during the earlier data collection (before Facebook's public anti-abuse filtering began). The bottom graph depicts the 6 (1.7%) Facebook accounts in dox files that changed their status after the abuse filtering methods were deployed. The figures have been normalized so that heights are comparable across graphs (i.e. bars of the same height represent the same number of people in all graphs).

These difference in size between the pre–and–post graphs show that a far greater number of accounts changed their status during the pre-filtering period than after filtering occurred. The far smaller red portion of the graph, in absolute terms, depicts that far fewer Facebook users made their account more private after being doxed after filtering was deployed. Both of these measures suggest that Facebook's abuse filtering techniques are successfully protecting users from at least some of the online social network related harms from doxing.

6.3.2 Instagram. Instagram publicly announced that they deployed several features in early September 2016 to mitigate harassment by filtering abusive comments [6]. Since our measurement periods are before and after their abuse filtering deployment, we are able to quantify the effectiveness of their filters in mitigating harm experienced by Instagram's doxed users.

Figure 3 depicts the changes in account status of Instagram accounts during our pre-and-post filtering collection periods, in the same manner as described in Section 6.3.1. The top graph shows the 12 (13.8%) accounts that changed their status during the earlier collection period, and the bottom shows the 7 (5.0%) accounts that changed their status during the latter collection period. The graph shows that Instagram's publicized anti-filtering techniques also appear to have reduced how frequently users change their account status after being doxed.

6.3.3 Twitter and YouTube. We also measured the status changes of Twitter and YouTube accounts referenced in dox files. We observed 4% of Twitter accounts (5 of 122), and 1% of YouTube accounts (1 of 71) changing their status in the earlier collection period, and another 4% of Twitter accounts (8 of 183) changing their status in the later collection period. We did not observe any dox-mentioned YouTube accounts changing their status in the latter collection period.

Users of both of these networks appear to change their statuses less frequently in response to appearing in dox files than users of Facebook or Instagram. One possible reason for this is that these accounts may be more frequently monetized and less likely to be personal accounts, than Facebook or Instagram accounts. The monetary costs Twitter and YouTube users face when closing their account might be higher, leading to less account status changing. We can only speculate, given our measurement vantage point. Building a better understanding of why accounts of different online social networks change their statuses at different rates as a response to being doxed would be a valuable area for further research.

7 DISCUSSION

Our quantitative study of doxing might be biased and follow up investigation is needed to better understand doxing. Even though we cannot draw strong conclusions from our study, we can still use it as a starting point to discuss how it improves our understanding of doxing and mention some potential steps to mitigate the problem.

7.1 Notification

As part of this ongoing study, we will continue to operate our dox file detection and OSN extraction pipeline. We are also in communication with Facebook to provide a feed of pastebin.com URLs and OSN accounts we extract through the Facebook Threat Exchange [8]. Our hope is that they can assist in mitigating some of doxing's harmful effects. Some of our ideas are to have Facebook notify the doxing victim and also notify pastebin.com to have the dox file removed. In addition, Facebook might be able to enable stricter harassment filtering and monitor the account for potential account compromise. If this is successful, we can reach out to other OSN platforms, such as Google+ and Twitter, that are also commonly included in doxes so that they can also assist in warning the victim and protecting targets on their platforms.

Additionally, we have been working with pastebin.com to automatically identify and notify the text sharing service when dox files have been shared on their site. This process is ongoing, with the possible goal of an automatic dox file filter that could remove abuse content from their service in a more timely manner than the current, email–based request system.

Finally, we hope to create a public service that would notify internet users when their information has been shared in a dox file. Similar to the "Have I Been Pwned" [16] service, our dox-notification service would allow internet users to register to be notified if a online social network or other unique identifier of theirs has appeared in a dox file. Just as with "Have I Been Pwned", our service would not share *what* information has been shared in a dox file, only *if* information has been shared and possibly where.

[8] https://www.facebook.com/threatexchange/

7.2 Anti-SWAT-ing Watchlist

Finally, we hope that this work can be used to reduce the frequency and harm of SWAT-ing attacks. SWAT-ing is a type of attack where someone calls a police department and reports violence at an address, in the hopes that the police will respond to the report by sending a SWAT-team to investigate the situation. In the context of doxing, it can be conceptualized as a form of harm-amplification; a way for a malicious party with a little bit of information about a target (i.e. their address) to cause a large amount of harm (i.e. a SWAT team kicking in their door).

The relationship between doxing and SWAT-ing is plain (the latter is a way of increasing the harm caused by the former), and well documented by internet security press [22, 24]. The main issue is that police forces lack a way of distinguishing between sincere reports of violence and fraudulent reports of violence by those wishing to further harm doxing victims.

Our work could be useful to police departments seeking better information in how to evaluate and respond to reports of violence. We plan to create a watchlist of addresses and phone numbers that have appeared in dox files, that could be shared with police departments and similar trusted sources. Equipped with such resources, a police department could check to see if an address has been associated with a dox file recently before deploying a SWAT team. While likely not determinative, such information could be useful in preventing doxing victims from also becoming SWAT-ing victims.

7.3 Followup Studies

Additionally, we plan to work with our IRB to create safe protocols for performing a follow up study in which we directly contact doxing victims. The goal of this study will be to better contextualize and quantify other harms doxing targets experience by performing one-on-one personal interviews and create surveys. Finally, we plan to improve the coverage of the doxes we detect by understanding how to identify most subtle instances of doxing that occur in addition to blatant doxes posted to pastebin.com and other doxing sites.

8 CONCLUSION

In this paper, we have presented the first quantitative study of doxing. We created an automated framework to detect doxes online, extracted online social networking accounts referenced in each dox file, and monitored these accounts for signs of abuse. Based on our manual analysis of these doxes, we were able to measure and understand the kinds of highly sensitive information shared in dox files.

Through these techniques, we were able to measure how many people are targeted by doxing attacks on popular text sharing services. We found that doxing happens frequently on these sites, and that some demographics and communities are disproportionately targeted. We find that doxing victims in our data set are overwhelmingly male, have an average age in their 20s, and a significant number are part of gamer communities (or maintain accounts with multiple video-game related websites). We also find that most doxes include highly identifying information of the victim and family members, such as full legal names, phone numbers and online social networking accounts. We found that doxing victims were dramatically more likely to close or make social networking accounts private after being doxed, and that abuse filters deployed by Instagram and Facebook successfully reduced how frequently doxing victims had to closed or increased the privacy of their accounts.

We hope that our quantitative approach helps researchers, internet users and web-services providers understand the scope and seriousness of online doxing. We also hope that our work can complement existing, qualitative work on doxing in order to provide a fuller understanding of doxing abuse online. Finally, we hope that maintainers of online social networks, law enforcement, and other parties with an interest in keeping internet users safe from the effects of doxing can use our techniques to mitigate the harm doxing causes.

9 ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation grants CNS-1717062, CNS-1237265 and CNS-1619620, AWS Cloud Credits for Research, and gifts from Google

REFERENCES

[1] Amanda B. 2012. Doxing Meme. http://knowyourmeme.com/memes/doxing. (2012).
[2] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. *arXiv preprint arXiv:1702.06877* (2017).
[3] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT '12)*. 71–80.
[4] Amanda C. Cote. 2017. "I Can Defend Myself": Women's Strategies for Coping With Harassment While Gaming Online. *Games and Culture* 12, 2 (2017), 136–155.
[5] Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 121–126.
[6] Nicholas Deleon. 2016. Everyone Can Now Use Instagram's Tool That Combats Offensive Language. https://motherboard.vice.com/en_us/article/instagram-tool-to-fight-harassment-offensive-language. (2016). [Online; accessed 5/16/2017].
[7] Caitlin Dewey. 2016. Twitter has a really good anti-harassment tool — and it's finally available to everyone. https://www.washingtonpost.com/news/the-intersect/wp/2016/08/18/twitter-has-a-really-good-anti-harassment-tool-and-its-finally-available-to-everyone/. (2016).
[8] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. 29–30.
[9] David M. Douglas. 2016. Doxing: a conceptual analysis. *Ethics and Information Technology* 18, 3 (2016), 199–210.
[10] Maeve Duggan. 2014. Online Harassment. Pew report. (2014).
[11] EVA GALPERIN and JILLIAN YORK. 2011. Victory! Google Surrenders in the Nymwars. https://www.eff.org/deeplinks/2011/10/victory-google-surrenders-nymwars. (2011).
[12] Google. 2017. Conversation AI. https://jigsaw.google.com/projects/#conversation-ai. (2017). [Online; accessed 5/16/2017].
[13] Dreßing H, Bailer J, Andres A, Wagner H, and Gallas C. 2013. Cyberstalking in a large sample of social network users: prevalence, characteristics, and impact upon victims. *Cyberpsychology, Behavior, and Social Networking* 17, 2 (2013), 61–67.
[14] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2016. A Longitudinal Measurement Study of 4chan's Politically Incorrect Forum and its Effect on the Web. *arXiv preprint arXiv:1610.03452* (2016).
[15] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* (2015).
[16] Troy Hunt. 2017. have i been pwned? https://haveibeenpwned.com/. (2017).
[17] Michelle Ashley Hunzaker. 2016. *Intent or Misinterpretation? Disruptive Behaviors within Ingress.* Master's thesis. North Carolina State University.

[18] Sameer Hinduja Justin W. Patchin. 2010. Cyberbullying and Self-Esteem. *Journal of School Health* 80 (2010).

[19] Dinakar K, Jones B, Havasi C, Lieberman H, and Picard R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems* 2, 3 (2012).

[20] Parul Khanna, Pavol Zavarsky, and Dale Lindskog. 2016. Experimental Analysis of Tools Used for Doxing and Proposed New Transforms to Help Organizations Protect against Doxing Attacks. *Procedia Computer Science* 94 (2016), 459 – 464.

[21] Liz Klimas. 2015. Swatting Prank Ends Horribly for Victim — and He Has the Injury to Prove It. http://www.theblaze.com/stories/2015/07/16/swatting-prank-ends-horribly-for-victim-and-he-has-the-injury-to-prove-it/. (2015).

[22] Brian Krebs. 2016. Serial Swatter, Stalker and Doxer Mir Islam Gets Just 1 Year in Jail. https://krebsonsecurity.com/2016/07/serial-swatter-stalker-and-doxer-mir-islam-gets-just-1-year-in-jail/. (2016). [Online; accessed 10/24/2016].

[23] Roney Simon Mathews, Shaun Aghili, and Dale Lindskog. 2013. A study of doxing, its security implications and mitigation strategies for organizations. http://infosec.concordia.ab.ca/files/2013/02/Roney$_{M}$athews.pdf. (2013).

[24] Nathan Mattise. 2015. 8chan user offers to "swat" GamerGate critic, cops sent to an old address. http://arstechnica.com/tech-policy/2015/01/8chan-tries-swatting-gamergate-critic-sends-cops-to-an-old-address/. (2015). [Online; accessed 10/24/2016].

[25] Torill Elvira Mortensen. 2016. Anger, Fear, and Games. *Games and Culture* (2016). https://doi.org/10.1177/1555412016640408 arXiv:http://dx.doi.org/10.1177/1555412016640408

[26] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. 145–153.

[27] Ingrid N Norris. 2012. *Mitigating the effects of doxing*. Master's thesis. Utica College.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[29] Alex Peysakhovich and Kristin Hendrix. 2016. News Feed FYI: Further Reducing Clickbait in Feed. https://newsroom.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/. (2016). [Online; accessed 5/16/2017].

[30] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: detecting cyberbullying instances in vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 617–622.

[31] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, Vol. 2. IEEE, 241–244.

[32] Sasha Sax. 2016. Flame Wars: Automatic Insult Detection. http://cs224d.stanford.edu/reports/Sax.pdf. (2016).

[33] Bruce Schneier. 2015. Doxing as an Attack. https://www.schneier.com/blog/archives/2015/01/doxing$_{a}$s$_{a}$n$_{a}$t.html. (2015).

[34] Aaron Swartz. 2002. html2text. http://www.aaronsw.com/2002/html2text/. (2002). [Online; accessed 9/20/2017].

[35] Kevin Systrom. 2016. Keeping Comments Safe on Instagram. http://blog.instagram.com/post/150312324357/160912-news. (2016).

[36] Robert S. Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior* 26, 3 (2010), 277 – 287.

[37] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, 1231–1245.

[38] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media (LSM '12)*. 19–26.

[39] Ellery Wulczyn, Dario Taraborelli, Nithum Thain, and Lucas Dixon. 2017. Algorithms And Insults: Scaling Up Our Understanding Of Harassment On Wikipedia. https://medium.com/jigsaw/algorithms-and-insults-scaling-up-our-understanding-of-harassment-on-wikipedia-6cc417b9f7ff. (2017).

[40] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*.

[41] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Workshop on Content Analysis on the WEB 2.0*.