SocialHEISTing: Understanding Stolen Facebook Accounts

Jeremiah Onaolapo
University of Vermont
jeremiah.onaolapo@uvm.edu

Nektarios Leontiadis
Facebook
leontiadis@fb.com

Despoina Magka
Facebook
despoinam@fb.com

Gianluca Stringhini
Boston University
gian@bu.edu

Abstract

Online social network (OSN) accounts are often more user-centric than other types of online accounts (e.g., email accounts) because they present a number of demographic attributes such as age, gender, location, and occupation. While these attributes allow for more meaningful online interactions, they can also be used by malicious parties to craft various types of abuse. To understand the effects of demographic attributes on attacker behavior in stolen social accounts, we devised a method to instrument and monitor such accounts. We then created, instrumented, and deployed more than 1000 Facebook accounts, and exposed them to criminals. Our results confirm that victim demographic traits indeed influence the way cybercriminals abuse their accounts. For example, we find that cybercriminals that access teen accounts write messages and posts more than the ones accessing adult accounts, and attackers that compromise male accounts perform disruptive activities such as changing some of their profile information more than the ones that access female accounts. This knowledge could potentially help online services develop new models to characterize benign and malicious activity across various demographic attributes, and thus automatically classify future activity.

1 Introduction

Social accounts are almost indispensable in our daily lives. Discovering old and new friends, consuming news, and securing the next lucrative job are a few of the many activities that social accounts facilitate. Compared to webmail and instant messaging accounts, social accounts provide much more than messaging functionality alone. Social accounts also accumulate personal information over time which unfortunately puts them within the sight of cybercriminals.

In this paper, we aim to understand what happens to social accounts after cybercriminals acquire credentials to them through illicit means. Specifically, we focus on understanding how the demographic attributes of stolen accounts influence the activity of criminals that connect to them. To this end we created, deployed, and monitored 1008 realistic decoy Facebook accounts (for ethical reasons, it is not possible for us to study accounts that belong to real persons, to avoid harming them). We incorporated various age and gender configurations in the accounts. To lure criminals into interacting with the accounts, we leaked credentials to a subset of them on the Surface Web and Dark Web, mimicking the modus operandi of cybercriminals that distribute stolen account credentials. We monitored the accounts for six months, extracted comprehensive activity records of people who visited the accounts, and analyzed those records offline.

Our research questions are as follows. How can we characterize the behavior of criminals in stolen accounts? Do differences in account demographics (age and gender) affect the activity of criminals in compromised social accounts? For how long do criminals stay in social accounts after logging in? What is the nature of content that they search for? What is the nature of content that they post?

In the course of experiments, we observed 322 unique accesses to 284 accounts. We show that the age and gender of an account owner indeed have a relationship with the types of actions that criminals carry out in the account; for example, attackers tend to search the friend list and start chats when interacting with teen accounts more than with adult ones, and perform disruptive activities while interacting with male profiles (e.g., editing their profile), while we never observed this behavior for female accounts. Our findings suggest that profile attributes have an influence on the actions that attackers take when compromising accounts, and open up future interesting research directions in both better understanding the modus operandi of attackers and developing better mitigations against account hijacking.

Key Lesson. Age and gender differences (in victims) influence the way cybercriminals behave when they access stolen Facebook accounts. This is in line with existing research literature which shows that age and gender are significant factors in cybercrime and online abuse victimization [37, 51, 59]. In view of this, we propose that mitigation systems and inter-

ventions should be customized along various demographic groups. In other words, we need to evolve security systems away from defending the mythical "average user" towards developing adaptive defense systems that address the significant differences in cybercrime victimization.

Contributions. First, we present a system to deploy and monitor honeypot accounts on Facebook. Our approach can be ported to other social networks to help understand the use of stolen accounts. Second, we instrument over 1000 Facebook accounts and collect 322 unique accesses over a period of six months. Third, we analyze how different demographic traits influence the way attackers interact with compromised Facebook accounts. Fourth, we put our results in the context of existing research, and discuss the need to develop tailored mitigation systems along the demographic attributes of users of online services.

2 Background

In this section, we first motivate our work in light of previous research. We then discuss related work and introduce Facebook accounts and the tools that we use to build our measurement infrastructure. Finally, we discuss our threat model.

2.1 Motivation

The existing research literature has explored various factors that influence cybercrime victimization. Victims suffer from different harms depending on their age, gender, and personality. Henson et al. [29] surveyed 10K undergraduate college students on their use of OSNs. They show that male and female users utilize OSNs in different ways, especially regarding the type and amount of content they upload, their flirting behavior, and the amount of time they spend on OSNs. Lévesque et al. [37] studied factors in malware victimization: they demonstrate that age and gender influence the likelihood malware victimization. In particular, Lévesque et al. show that men are at more risk of encountering malware than women, across most types of malware. Multiple studies show that women are disproportionately targeted by sexual harassment and stalking online [22, 36, 51], and that younger people are more likely to receive online harassment [51].

Age also plays a significant role in victimization. Näsi et al. [40] reported that younger people are more likely to be victims than the older ones (participants were selected from people between ages 15 and 30). Oliveira et el. [41] demonstrated that older women are more susceptible to phishing attacks than people from other age groups. On the other hand, Sheng et al. [45] showed that younger people (18 to 25 years old) are more likely to be victims of phishing attacks.

On a related note, van de Weijer and Leukfeldt [54] studied the Big Five personality traits as factors related to the likelihood of cybercrime victimization. They reported conscientiousness and emotional stability (lower scores) and openness to new experiences (higher scores) as factors that predict cybercrime victimization. Egelman and Peer [24] dispelled the myth of the "average user," and proposed a targeted approach to nudge individual users towards better security and privacy controls. Although [54] and [24] disagree on the utility of the Big Five personality traits, they both point to the need for individualized interventions for users and victims alike.

Since age, gender, and personality play a significant role in online victimization, it is therefore logical to expect that the behavior of a criminal on breaching a specific online account would depend on those attributes (of the victim). The existing research literature has focused more on victims and their susceptibility to online crimes and abuse; instead, we study how the demographic attributes of a victim account influences the behavior of criminals. To the best of our knowledge, this is the first paper that explores such activity within Facebook accounts. In the following section, we highlight existing literature in account compromise.

2.2 Related Work

Account Takeover. Cybercriminals gain access to online accounts through various means, including information-stealing malware [15, 47], data breaches [27, 55], and manual account hijacking [18]. Redmiles [43], via qualitative interviews, studied how people respond to attacks on their Facebook accounts. Thomas et al. [52] examined suspended accounts on Twitter and thus characterized spam accounts and techniques that spam actors rely on. Social spam and fake accounts have been studied extensively [16, 35, 53, 56, 57, 61]. Work has been conducted on understanding the threat of compromised accounts and developing systems to detect such attacks [23, 49]. Instead, we focus on understanding how the demographic attributes of online accounts influence the activity of criminals when they compromise such accounts; we explored age range and gender variables but this approach could be extended to other demographic attributes as well. In the next section, we highlight a number of papers that employed honeypot approaches related to ours.

Honeypots. DeBlasio et al. [20] studied compromised websites by leveraging honey webmail accounts. Han et al. [28] studied the phishing ecosystem by deploying sandboxed phishing kits and recording live interactions of various parties that accessed those kits. Other papers studied the behavior of criminals in compromised webmail and cloud document accounts by deploying honey accounts and honey documents [18, 34, 42]. In this paper, we focus on the influence that demographic traits have on the activity of malicious actors accessing compromised accounts; elements that the online services studied by previous work did not have available.

Kedrowitsch et al. [32] explored ways to improve Linux sandboxes for evasive malware analysis. Cao et al. [19] deployed an operational network honeypot to automatically

detect and evade SSH attack attempts. Barron and Nikiforakis [14] deployed honeypot machines and observed how the system properties of those machines influenced the behavior of attackers. In this paper, we focus on compromised social network accounts instead of compromised machines.

2.3 Facebook Accounts

A potential Facebook user first creates an account and an associated *profile*. Afterwards, they can send *friend requests* to their peers. They can post updates on their profile *timeline*, for instance, by writing text, uploading a photo, or posting a URL (or a combination of those actions). Facebook also allows users to send private messages to their friends via *Messenger* (Facebook's messaging application). Users can click *like* (and other "reactions") on posts, photos, and other content of interest to them. Facebook usage is not limited to individual users. Informal groups, businesses, and corporate entities can also maintain Facebook presence by creating *pages* and *groups*. Users can search for, and connect to, friends, groups, and pages they are interested in. These features, among others, highlight the social nature of Facebook.

2.4 Test Accounts

In addition to regular accounts, Facebook provides sandboxed accounts that are disconnected from their main social graph. These accounts, known as *test accounts*, are similar to real accounts, but exist in an isolated environment (a sandbox). Hence, they cannot connect to regular Facebook accounts, but can connect to other test accounts (i.e., as "friends"). They are often used for testing purposes, for instance, in security vulnerability testing [6]. The inherent isolation of test accounts makes them particularly suitable for our studies in understanding malicious activity in compromised social accounts, since it ensures that real users will not be harmed in any way during experiments, and this matches our ethics requirements for studies of this nature. We discuss these ethical considerations in Section 3.5. At the same time, we ensure that the accounts look believable. Facebook also provides a dashboard for managing test accounts. The dashboard, which is accessible only from a real Facebook account, allows the account manager to reset passwords of test accounts under their control.

Although test accounts look similar to real Facebook accounts, there are limits to their capabilities. Since test accounts are disconnected from the regular Facebook graph, attempts to interact with regular accounts will fail. For instance, attempts to search for a real account or fan page will not succeed. Nevertheless, such search terms will be recorded in the test account's activity records and will be available to the researcher in control of the test accounts. Also, attempts to authenticate to other Facebook-affiliated platforms (e.g., Instagram) using test accounts will fail, while such attempts via real accounts will succeed (for valid account credentials).

Despite these limitations, test accounts provide a level of realism that is close to that of real Facebook accounts, hence are a good fit for this paper. Therefore, we only use test accounts to conduct this research.

2.5 Download Your Information (DYI)

A Facebook user may desire to download and review their own account data and activity. To facilitate this, Facebook accounts present a built-in tool known as Download Your Information (DYI) which allows users to request and download a compressed archive containing their account data and activity over time [1]. The DYI tool is available via the *Settings* menu of Facebook accounts. After requesting and downloading the compressed archive (DYI archive), the user can then uncompress the archive offline and peruse its contents. It is usually structured like an offline web site organized in directories (sections) and web pages that can be viewed offline in a Web browser. Alternatively, DYI data can be downloaded in JavaScript Object Notation format (JSON).

A DYI archive provides information on login times, IP addresses, user-agent strings, messages, group chats, timeline posts, profile edits, and photo uploads, among others. It provides a comprehensive record of activity within a Facebook account. However, it does not provide 100% coverage of all observable phenomena within a Facebook account—for instance, it does not record page scrolling information. Despite this, DYI archives constitute a rich source of information for our experiments. For these reasons, we rely on DYI functionality in Facebook accounts to retrieve activity data from test accounts at the end of experiments (see Section 3.2). Note that we also refer to test accounts as honey accounts.

2.6 Threat Model

Attackers compromise credentials of online accounts through phishing attacks, information-stealing malware, network attacks, and database breaches, among other ways [21, 47, 49]. Afterwards, they connect to the accounts to search for valuable information to monetize. Some criminals also use compromised accounts to send spam messages [26]. In this paper, we focus on attackers that target social accounts and misuse them in various ways, for instance, by sending unsolicited messages to contacts of the victim or stockpiling stolen social credentials for sale. The attackers under study have similar privileges (within the stolen accounts) to owners of the accounts, since those attackers have knowledge of the access credentials that owners possess. Attackers also have the ability to extend the reach of their malicious activity to other entities (i.e., accounts) connected to the victim's social graph, for instance, by abusing inherent trust and sending malicious payloads to them.

3 Methods

We created 1008 Facebook test accounts in total, comprising equal numbers of female adult, male adult, female teen, and male teen accounts. In this section, we describe how we created, instrumented, and deployed them.

3.1 Setting Up Honey Accounts

The process of populating the test accounts with data spanned about 6 months, from November 7, 2017 until May 16, 2018. We discuss those specific steps next.

Demographic Factors. Lévesque et al. [38] examined gender and age, among other demographic factors, as risk factors in malware infections. Inspired by their approach, we designed personas around two demographic attributes, namely age range (teen/adult) and gender (male/female). We wanted to observe differences or similarities in the behavior of criminals to the honey accounts, depending on the demographic attributes of the accounts.

Profile Names and Passwords. We assigned first and last names to the profiles by generating random combinations of names using the API of a *random user generator* [11]. We then assigned passwords to the profiles by randomly selecting passwords from the publicly available *RockYou* password list, comprising 32 million passwords that were exposed during a 2009 data breach [39]. To increase the realism of the accounts, we established friend connections among them to mimic the social nature of real Facebook accounts.

Profile Photos. We sourced profile photos for the accounts by downloading Creative Commons (CC) stock photos from *Pixabay* [12], *Flickr* [4], *Pexels* [2], and *Unsplash* [7]. We chose only CC0-licensed photos from those sources; the photos can be used for any purpose and do not require attribution. We manually matched photos to accounts, taking care to ensure that each profile photo represented the previously designated demographic attributes of its host account. For instance, for a female adult account, we chose a profile photo that shows an adult woman. Finally, we uploaded the curated profile photos to honey accounts using a photo upload automation tool that we built for this purpose. Thus, at a glance the demographic label of any given account can be inferred by anyone that connects to the account.

Timeline Data. To further mimic real Facebook accounts, we posted some content on the timelines of honey accounts. To this end, we collected publicly available tweets containing popular hashtags, using the Twitter Streaming API [3] according to their terms of service. These popular hashtags, identified in previous work [13], include `#sports`, `#music`, and `#news`, among others. We removed personally identifiable information (PII) from the tweets and posted the sanitized text snippets on timelines of honey accounts using an automation tool that we built. Hence, the honey accounts display diverse content on topics that people usually post about on social networks, and are more convincing as a result.

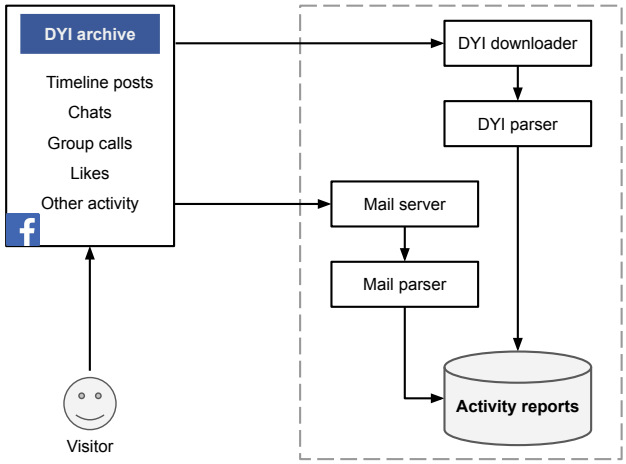


Figure 1: Monitor overview. We observe attempts to change account passwords via notification emails that arrive at our mail server, and collect records of account activity via the DYI downloader.

We ensured that the accounts looked realistic by populating them with real-world data and connecting them with each other (i.e., friend connections). Our accounts stopped posting messages and interacting with each other shortly before the credentials were leaked (we discuss those leaks in Section 3.3). From that point, the only activity we carried out was accepting the friend requests that the attackers made. We did not drive any further activity from the accounts. In particular, we did not interact with any attacker, for example by replying to their private messages. This was primarily done to follow our IRB protocol, which forbade us from interacting with attackers (see Section 3.5 for more details).

We acknowledge that this might have introduced a sense of "staleness" in the accounts, and might have influenced the activity of the attackers. While some attackers might have been goofing around, the fact that we find statistically significant differences in the activity performed on accounts with different demographics gives us confidence that we are capturing real attack trends.

3.2 Data Collection Infrastructure

In this section, we present the data collection infrastructure that we built to retrieve activity data from honey accounts; Figure 1 illustrates its components and how they interact. Next, we explain each key component.

Download Your Information (DYI) Archive. As previously mentioned (Section 2), Facebook accounts, including test accounts, allow account owners to download DYI archives con-

taining comprehensive records of their activity. We rely on this feature to collect activity records of criminals.

DYI Downloader and Parser. DYI archives are composed of Web pages containing activity details for offline viewing. We automatically download them and run them through a parser to extract and categorize their contents. Such contents include login and logout information, device information, and password changes, among others.

Mail Server and Parser. While setting up test accounts, we associated certain email addresses to the honey accounts. Those email addresses (one per Facebook test account) point to a mail server under our control. On that mail server we receive email notifications from honey accounts about password changes, incoming friend requests, and private messages, among others. Unlike DYI records which we download only once, the mail server provides us with real-time information about account activity and allows us to react immediately when necessary (e.g., to revert password changes).

3.3 Leaking Honey Credentials

Stolen credentials are often distributed on paste sites and other outlets by cybercriminals [48]. We mimicked the credential-leaking approach to attract cybercriminals to our honey accounts by leaking credentials (Facebook IDs and passwords) via paste sites on the Surface Web (`Pastebin.com`, `Paste.org.ru`) and the Dark Web (`Stronghold`). These are ideal outlets because they allow public pastes and show recent pastes to all visitors.

We did not leak the entire population of honey accounts. Instead, we leaked two-thirds of them (672 credentials out of the entire set of 1008). We did this to observe if criminals would attempt to compromise the accounts that were not leaked by leveraging existing friend connections among the accounts. For instance, they might send phishing messages or malicious links to accounts whose credentials we did not leak (we set up friend connections among the test accounts).

Given the large number of credentials that we leaked (672 accounts), we divided them into seven chunks, each chunk comprising a maximum of 100 credentials. Note that paste sites allow users to see "recent pastes" on their home pages, but only a small number of submissions appear at a time (e.g., eight in the case of `Pastebin.com`). For this reason, we leaked the credentials on a daily basis. To ensure that our leaks favor paste site visitors from multiple time zones that differed from ours, we leaked credentials twice daily. Finally, to ensure that the credentials were adequately exposed during leaks, we randomized the order of credentials in each chunk prior to leaking them.

3.4 Threats to Validity

We acknowledge the existence of factors that may affect the validity of our findings. First, the content of the honey accounts comprise stock photos and other publicly available data, which might be obvious under close scrutiny. Also, a close look might reveal that the honey accounts were created fairly recently, and that they stopped posting new statuses after we stopped populating them—this can possibly influence the credibility of our accounts. We do not consider these to be major issues since such criminals would have connected, at least once, to the accounts, and we would have recorded their activity already.[1] We also do not have a systematic way to determine what happens if users of paste sites—our leak outlets—realized that the accounts were fake. Note that paste sites do not have direct feedback mechanisms (e.g., comment fields), unlike forums. Finally, we leaked credentials anonymously on paste sites; our leaks were not connected to any single identity. Hence, we replicated an anonymous leak.

Recall that we used sandboxed accounts (test accounts) that are disconnected from regular Facebook accounts. A close observation may reveal the presence of features that differ slightly from real accounts. Note that we leaked credentials through paste sites only. Our findings may not be representative of malicious activity in social accounts stolen via other outlets, for instance, malware or underground forums. Despite these factors, this paper offers insights into malicious activity in stolen social accounts and will help in developing detection and mitigation systems and techniques.

3.5 Ethics

We carefully considered the ethical implications of our work while setting up and running experiments. First, we used accounts that were isolated from the regular Facebook social graph to avoid harming legitimate Facebook users. This sandbox approach is in line with common practices in malware research [44]. Second, we used publicly available stock photos and tweets to populate the accounts. We did this to ensure that no private information was leaked in this study. Third, by leveraging the test dashboard, we ensured that account passwords could be changed easily by us, to lock criminals out, if we observed attempts to harm people via honey accounts. In addition, our monitor system recorded all attempts to change the email addresses associated with the honey accounts. Our initial mitigation plan was to connect to such accounts and restore their original email addresses, which were under our control. We later found that Facebook already had a mitigation mechanism in place: attempts to change email addresses were blocked by Facebook, and access to the affected accounts was temporarily disabled until we reset them via the test dashboard.

[1] We also acknowledge the possibility of rare exceptions in which prospective visitors may perform a reverse image search on an account's publicly-accessible profile picture without logging in, realize that it is a stock photo—thus, likely a fake account—and then discard its credentials without ever connecting to it.

To further strengthen our ethics protocol, we asked our Facebook contacts to keep an eye on the accounts with a view to shutting down any account that violates Facebook's policies. After our analysis, we securely discarded PII that accrued in the accounts during experiments. Finally, since our experiments involved deceiving criminals to interact with decoy accounts, we sought and obtained ethics approval from our institution prior to starting experiments.

4 Data Analysis

In this section, we provide an overview of the activity performed by criminals in honey accounts. We leaked credentials to the accounts in a three-week period (from June 1, 2018 to June 22, 2018), and our observations span six months (from June 1, 2018 to December 1, 2018). Our analysis and the corresponding insights are based entirely on data collected from honey accounts under our control; we did not use any internal Facebook data.

4.1 Discarding Defective Accounts

As described in Section 3.2, our data collection method involves downloading DYI archives from honey accounts. In the process, we discovered that 79 accounts were defective, and we could not download activity information from them. Those defective accounts presented infinitely-spinning GIFs instead of loading page content, possibly due to a configuration issue while setting up the test accounts. We were unable to download activity data from them. In addition, three accounts were blocked by Facebook; we could not retrieve DYI data from them. We excluded those defective and blocked accounts from analysis, and this reduced the effective number of honey accounts under analysis from 1008 accounts to 926 accounts. These functional accounts comprise 472 adult accounts and 454 teen accounts (from the age range perspective), or 469 female accounts and 457 male accounts (from the gender perspective). Finally, the effective number of (functional) leaked accounts reduced from 672 to 619.

4.2 Accesses and Associated Actions

284 (46%) of the functional leaked accounts received unauthorized accesses. We did not leak 307 accounts. Unfortunately, due to the sandboxed nature of these accounts, it was not possible for attackers to find these accounts independently and connect to them. This study cannot therefore estimate the difference in risk of leaked and unleaked accounts. We did however observe that 46 unleaked accounts (15%) received interactions by attackers, in the form of friend requests or private messages. It is possible that some of these were an attempt to further gain access to those unleaked accounts. However, our inability to interact with attackers, because of our IRB protocol, did not allow us to investigate this further.

Facebook accounts record unique accesses to them, and each access is labeled with a unique string identifier known as a *cookie*. Cookies can be found in the login records section of DYI archives. An access is recorded when a criminal connects to a honey account. Note that access identifiers (cookies) can persist across logins into different accounts. For instance, if a criminal connects to account *A* and then connects to another account *B* using the same device and browser within a short time, the same cookie will be recorded in both accounts. After logging in, a criminal performs some *actions*, for instance, sending a private message or writing a status update. We use the terms *cookie* and *access* interchangeably in this paper. We observed various types of accesses in the accounts and named them according to the actions associated with them in the accounts. These types of accesses, codified into a taxonomy of accesses, are described next.

Hijacker. A hijacker access is recorded when the password of a honey account (or its email address) is changed.

Chatty. This type of access happens when a criminal sends private messages, creates group chats, posts an update on the timeline of another account, or posts on their own timeline.

Emotional. An emotional access is recorded during clicks on a Facebook "like" button (or any other reaction) on photos and posts.

Searcher. This type of access occurs when a criminal enters search terms in the Facebook search bar.

Profile Editor. A profile editor access is recorded when a criminal edits an account's profile information (e.g., by changing the profile photo).

Friend Modifier. This type of access occurs when a criminal adds or removes a friend from an account.

Curious. A curious access occurs when a criminal connects to an account but does not perform any of the previously listed actions. In other words, curious accesses comprise accesses that resulted in actions that were not captured by our monitor infrastructure because of limits to its coverage (e.g., clicking on photos to expand them or scrolling through the account profile).[2] To this end, in curious accesses, we record the act of logging in as an action itself, unlike the previously listed access types. Hence, curious accesses encompass a lower bound of actions that our monitor infrastructure was not equipped to capture.

These types of accesses are not mutually exclusive, except the curious type. For instance, an access that is chatty can also be emotional, depending on the various actions associated with it. However, curious accesses can only belong to the curious category.

4.3 Actions

In total, we observed 322 unique accesses to 284 accounts, which resulted in 1159 actions in those accounts. This number

[2] Also note that we do not have fine-grained data on what possibly happened during *curious* accesses.

Table 1: Summary of actions grouped by access type. Curious, searcher, and chatty accesses clearly dominate the table.

Access type	*Number of actions*	*Percentage*
Curious (*cur*)	518	44.7
Searcher (*sea*)	342	29.5
Chatty (*cha*)	127	11.0
Friend modifier (*fri*)	113	9.7
Hijacker (*hij*)	29	2.5
Emotional (*emo*)	18	1.6
Profile editor (*pro*)	12	1.0
Total	**1159**	**100.0**

of accesses is in line with what reported by previous work that followed a similar methodology when leaking online credentials [42]. Table 1 shows a summary of actions grouped by access type. Curious, searcher, and chatty accesses dominate the table of actions, responsible for 45%, 30%, and 11% of all actions respectively. Emotional and profile editor accesses constitute the least active types. This indicates that criminals who carry out actions in Facebook accounts are particularly interested in searching for information via the Facebook search bar, and writing private messages and public posts.

One of Facebook's core functions is connecting people; it provides ways to locate and connect to other Facebook users—to eventually make them *Facebook friends*. Recall that we created friend connections across the entire population of honey accounts, prior to the experiments, as mentioned in Section 3.1. In the course of experiments, we further observed additional friend requests made by cybercriminals to the accounts. In total, 157 accounts received friend requests from other accounts. These comprise 83 teen accounts and 74 adult accounts, from the age perspective, or 31 male accounts and 126 female accounts, from the gender perspective. These margins in received friend requests across age range and gender groups foreshadow further distinctions that we highlight throughout this paper. Finally, it is interesting that 46 unleaked accounts received friend requests (we did not leak 307 accounts, as explained in Section 3), while 111 leaked accounts received friend requests (we leaked 619 functional accounts). This shows that the attempted reach of criminals extended beyond the corpus of credentials they obtained.

IP Addresses. 90% of the IP addresses recorded in the accounts accessed less than 5 accounts each. 50% of them accessed exactly one account, as shown in Figure 2. The most prolific IP address accessed 93 accounts—an outlier, as shown by the long tail in Figure 2. In a general sense, a variety of attackers connected to the accounts—the recorded activity is not simply a reflection of the activity of a handful of attackers. In Section 4.10, we further discuss those IP addresses.

Next, we study the timing of activity in honey accounts, with particular emphasis on how long the recorded accesses lasted.

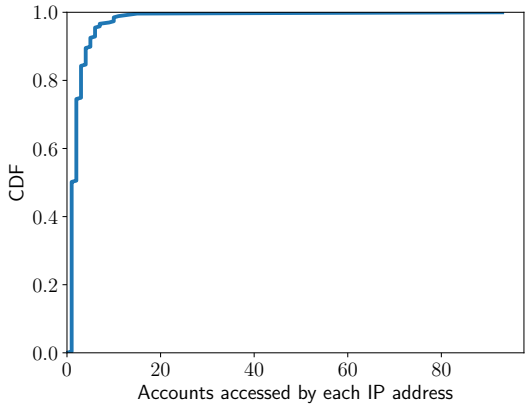


Figure 2: CDF of the number of accounts accessed by each IP address. 90% of IP addresses accessed less than 5 accounts each. 50% accessed exactly one account. A particularly prolific IP address accessed 93 accounts, hence the long tail.

4.4 Timing of Account Activity

We set out to understand the time patterns of accesses to accounts. To this end, we measured how long it took criminals to connect to the accounts after we leaked account credentials, and how long they stayed connected to the accounts. These measurements were carried out across all accounts, and also on groups of accounts (by age range and gender), to observe differences in activity patterns across different types of accounts. We present detailed measurements next.

Leaks to Logins. Recall that we leaked credentials of honey accounts via paste sites to attract criminals to them. To observe how long it took them to connect to the accounts after the leaks, we computed time lags between the first leak (dated June 1, 2018) and first access to each account. Note that the account credentials were leaked simultaneously at multiple times. As the CDF in Figure 3 shows, the accounts were mostly not accessed instantly. Instead, criminals connected to them gradually over several days. By the 25th day, more than 50% of accounts that were visited had received at least one access.

Spike in Accesses. The spike recorded in logins after the 25th day since first leak (see Figure 3) was caused by the previously mentioned prolific IP address that accessed 93 accounts in a single day. Those accesses all occurred on June 28, 2018, which coincides with the spike in Figure 3. The user-agent string associated with those accesses indicates that the connections were made from an Android device—and the accesses were possibly made in an automated manner. However, this is just an indication, since user-agent strings can be easily changed; they are not reliable.

Access Duration. To understand how long criminals stayed in the honey accounts, we computed the duration of their accesses. To achieve this, we recorded the time that a cookie

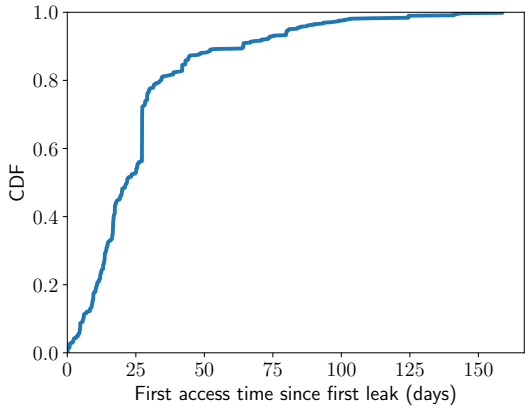


Figure 3: CDF of the time difference between first instance of credential leaks (across all outlets) and first logins.

first appeared in an account as t_0, and the last time it appeared in that account as t_{last}. Given this information, access duration can be computed as $t_{last} - t_0$ for each access. Figure 4 shows CDFs of access duration grouped by access type. Curious accesses are mostly short-lived, with the exception of a long tail of accesses, comprising a tiny fraction that stay in accounts for 80 days or more. It is possible that curious accesses that stayed connected to accounts for extended amounts of time were made by stealthy criminals that perform no action in stolen accounts to avoid being detected. Instead, they possibly monitor the accounts for an extended period to observe new sensitive content that could potentially benefit them. Finally, hijacker accesses mostly connect to accounts for less than one hour in our dataset.

We further computed access duration by age range to see if there were differences in access duration in adult accounts compared to teen accounts. The CDFs in Figure 5 show that criminals spend approximately the same time in teen accounts as adult accounts, but accesses to adult accounts present a longer tail than accesses to teen accounts. Finally, we computed access duration by gender, to see if there were differences in access duration in female accounts compared to male accounts. The CDFs in Figure 6 show that criminals spend slightly more time in female accounts than in male accounts.

Statistical Tests on Access Duration. To test the statistical significance of differences in access duration, we relied on the two-sample Kolmogorov-Smirnov (KS) test [33, 46]. The null hypothesis is that both samples under examination belong to identical statistical distributions. The output of the test is a KS statistic and p-value. A small KS statistic or high p-value shows that we cannot reject the null hypothesis. First, we tested the access durations of each access type against all access durations, to see the access types for which we can reject the null hypothesis. As Table 2a shows, searcher, curious, and profile editor accesses differ most from the distribution of all accesses (i.e., we can clearly reject the null hypothesis), while hijacker accesses differ least (we cannot reject the null hypothesis). Next, we compared adult and teen access durations ($p = .92$). Likewise, we compared female and male access durations ($p = .13$). In both tests, the null hypothesis cannot be rejected.

4.5 Effects of Demographic Attributes

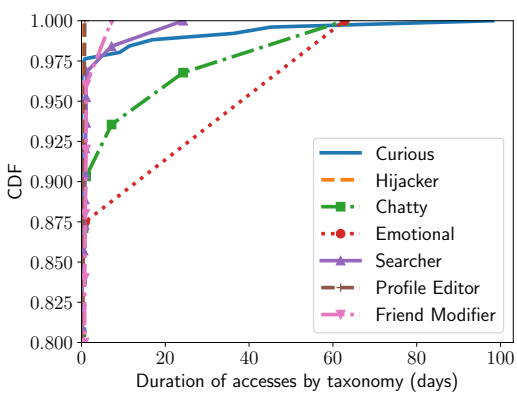
To understand whether the age and gender of an account affect the actions performed by cybercriminals, we calculated the proportions of access types in each age range and gender. From the age perspective, Figure 7a shows that criminals add and remove friends from adult accounts much more than they do in teen accounts. On the other hand, they edit profiles and are chattier in teen accounts than they are in adult accounts. From the gender perspective, Figure 7b shows that female accounts present more friend list modification activity than male accounts. On the other hand, search activity and profile editing occurs more in male accounts than female accounts; no profile edits were recorded in female accounts.

Statistical Tests on Age and Gender. To understand how age and gender differences affect the activity of criminals, we carried out Fisher's exact test [25] to determine if access types were independent of demographic attributes (i.e., age range and gender). The null hypothesis states that there is no association between demographic attributes and access types. Table 2b shows that there is indeed a significant relationship between account age and access type, particularly in chatty and friend modifier accesses, for which we reject the null hypothesis. Similarly, Table 2c shows a significant relationship between account gender and access type, especially in friend modifier, searcher, and profile editor accesses. This shows that the demographic attributes of accounts indeed influence the activity of criminals in those accounts.

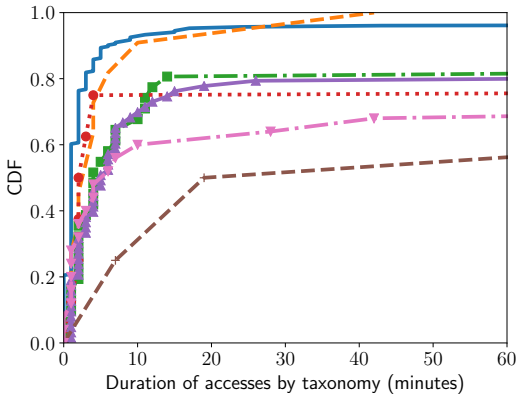
4.6 Action Sequences

A browsing session on social media does not comprise a single action; it is usually a sequence of actions. To further understand the activity of criminals in compromised accounts, we studied transitions among actions in the accounts during accesses. We studied these transitions to observe differences across male and female accounts, and teen and adult accounts. For instance, if a criminal connects to an account, clicks "like" on a photo (*emotional*), sends a private message to another account (*chatty*), and finally changes the password of the original account (*hijacker*), we denote that flow of ordered actions as an $emo \rightarrow cha \rightarrow hij$ chain. Note the use of shorthand labels. Table 1 shows the full list of shorthand labels.

We modeled access types as states and then computed probabilities of state transitions by following the flows we observed in the accounts. This resulted in directed graphs with weighted edges. We present them in Figures 8 and 9 to

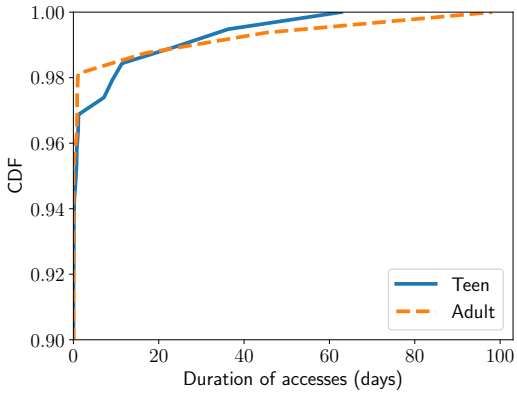


(a) All.

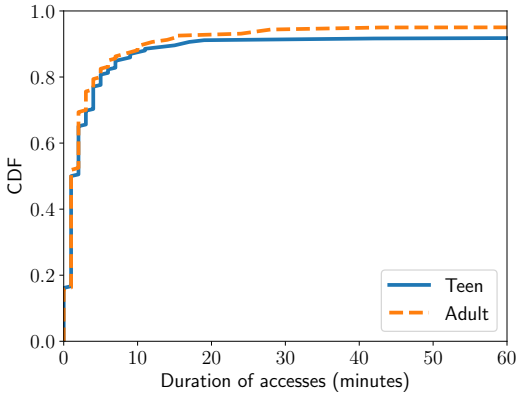


(b) First hour only (in minutes).

Figure 4: CDFs of access duration per access type. 4a shows the entire span of accesses, while 4b shows the first hour only. To enhance the visibility of the curves, the y-axis of 4a shows only the 80th to the 100th percentile ticks, while 4b shows all percentile ticks.

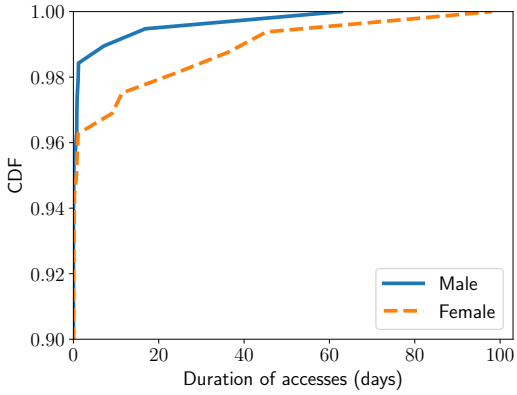


(a) All.

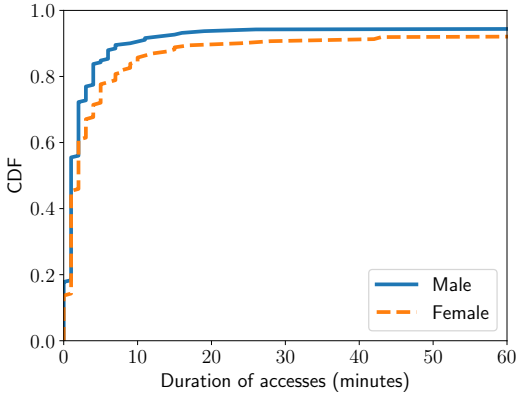


(b) First hour only (in minutes).

Figure 5: CDFs of access duration per age range. 5a shows the entire span of accesses, while 5b shows the first hour only. To enhance the visibility of the curves, the y-axis of 5a displays only the 90th to the 100th percentile ticks.



(a) All.



(b) First hour only (in minutes).

Figure 6: CDFs of access duration per gender. 6a spans all accesses, while 6b shows the first hour only. To enhance the visibility of the curves, the y-axis of 6a shows only the 90th to the 100th percentile ticks.

Table 2: Statistical tests on access durations, age differences, and gender differences (significance level=.01).

Access durations	
Access type	*P-value*
Searcher	$p < .01$
Curious	$p < .01$
Profile editor	$p < .01$
Chatty	$p = .01$
Friend modifier	$p = .034$
Emotional	$p = .289$
Hijacker	$p = .82$

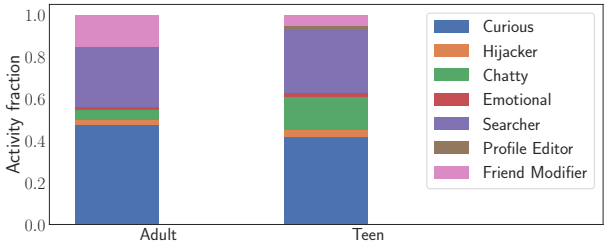
(a) Searcher, curious, and profile editor accesses differed most from the distribution of all accesses to the accounts (KS test).

Age differences	
Access type	*P-value*
Chatty	$p < .01$
Friend modifier	$p < .01$
Profile editor	$p = .045$
Curious	$p = .066$
Emotional	$p = .344$
Hijacker	$p = .452$
Searcher	$p = .518$

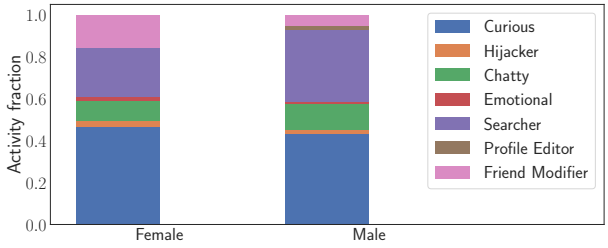
(b) Chatty and friend modifier accesses were influenced by the age ranges portrayed in the accounts (Fisher's exact test).

Gender differences	
Access type	*P-value*
Friend modifier	$p < .01$
Searcher	$p < .01$
Profile editor	$p < .01$
Chatty	$p = .09$
Emotional	$p = .095$
Curious	$p = .26$
Hijacker	$p = .348$

(c) Friend modifier, searcher, and profile editor accesses were influenced by gender (Fisher's exact test).



(a) Accesses per age range.



(b) Accesses per gender.

Figure 7: Distributions of access types across age ranges and genders.

shed light on differences in action transitions across different age ranges and genders respectively. Note that the sum of probabilities (weights) of outgoing edges do not always sum up to 1, but instead to values close to 1, because of rounding errors. Our approach is similar to the one employed by Wang et al. [56] in building clickstream models to detect Sybil accounts. In this section, we explore selected one-hop transitions (e.g., *emo*→*cha*) that are particularly interesting and deserve a closer look. These sequences of activities consider unique accesses. They are therefore depicting the same attacker performing a sequence of actions on a certain account during the same browsing session. To outline their browsing session, we tracked them using cookies (see Section 4.2), sorted their actions in a chronological order, and built activity chains.

Age. As Figure 8 shows, *pro*→*pro* (0.7), *emo*→*hij* (0.17), and *emo*→*emo* (0.083) transitions exist in teen accounts, while they are absent from adult accounts. On the other hand, *emo*→*fri* (0.17), *emo*→*sea* (0.17), and *sea*→*emo* (0.0067) transitions exist in adult accounts, but are absent from teen accounts. In our dataset, criminals remain in the profile editing state in teen accounts only, and they stay in the searcher state in teen and adult accounts at roughly the same rate (approximately 0.7). Also, they remain in the chatty state within teen accounts more than they do in adult accounts. Conversely, criminals stay in the friend modifier state within adult accounts more than they do in teen accounts. These findings corroborate and shed more light on the demographic results presented in Section 4.5. They also indicate that action sequences could possibly be used to distinguish between attacker activity in teen and adult accounts.

Gender. The first striking observation in Figure 9b is the disconnected *pro* node; transitions to or from the *pro* state do not exist on the female graph. This gender difference is further highlighted by the relatively high probability of accesses staying in the *pro* state within male accounts (0.58). It indicates that profile editing constitutes a strong distinguishing activity from the gender perspective. Chatty accesses tend to remain in the chatty state within male accounts (0.62) more than they do in female accounts (0.53), while friend modifier accesses maintain their state in female accounts (0.74) more than they do in male accounts (0.23). Similar to our observations from the age range perspective, criminals stay in the searcher state at roughly the same rate in male (0.67) and female (0.65) accounts. Finally, Figure 9 shows *pro*→*pro* (0.58), *pro*→*sea* (0.33), and *sea*→*pro* (0.023) transitions in male accounts only; they are absent from female accounts. Conversely, it shows *emo*→*emo* (0.083), *emo*→*fri* (0.082), and *emo*→*sea* (0.083) transitions

Table 3: The most common words in search text (left) and chatty text (right).

Searchers	*Count*	*Chatty*	*Count*
atheism	28	wave	14
debat	27	hi	12
bihar	19	[EXPLETIVE]	6
robson	15	hii	5
karla	10	fake	5
religion	10	babi	5
facebook	9	que	4
honest	9	http	4
india	9	password	3
ancud	8	metoo	3

Table 4: Accounts that recorded a specific top search term.

Top search term	*Number of accounts*
atheism	9
debat	9
bihar	7
robson	8
karla	2
religion	8
facebook	6
honest	5
india	4
ancud	2

in female accounts only; they do not exist in male accounts.

These findings indicate that behavioral patterns could potentially help in distinguishing malicious users from benign users in the future. However, that task is not in our scope of work since we do not have access to the action flows of legitimate users (baseline flows); large online services have the capability to compute them.

4.7 What Searchers Seek

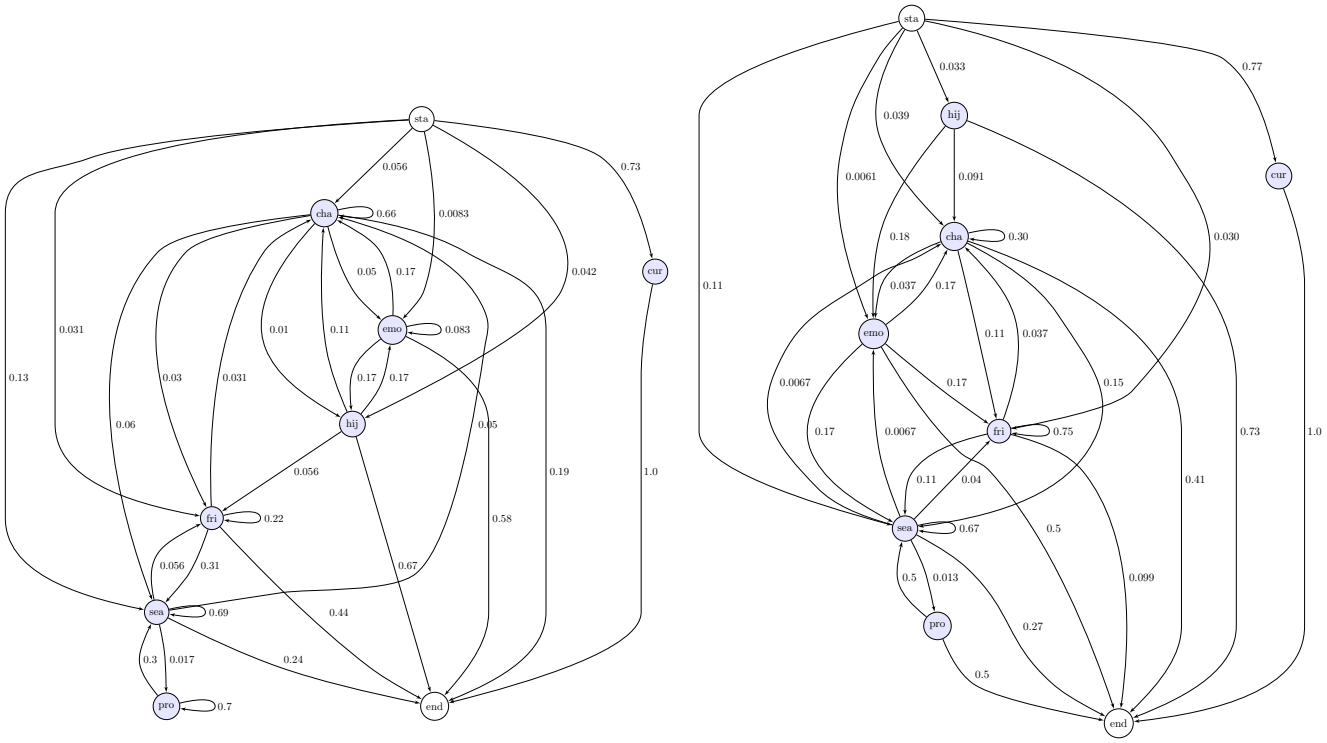
As shown in Table 1, searcher accesses were responsible for a substantial share of actions in honey accounts (30%). Various search terms were recorded in 87 accounts (entered via the Facebook search bar). To understand what the criminals were searching for, as an indication of their intent, we analyzed the search logs present in DYI archives. Table 3 (left-hand side) shows the most common words in the search logs. Those words were extracted and counted using the following steps (implemented in Python). First, we combined all search terms into a single document. Next, we tokenized the document into words and removed all English stop words (e.g., "the") using the *nltk.tokenize* package [10]. We then stemmed the remaining words using the Porter Stemmer function in the *nltk.stem* package [9]. Finally, we counted the resulting words; the top ten words are presented in Table 3. The search terms include religion-related words as a result of numerous searches for debates on atheism and religion. Other interesting search terms that showed up in search logs include "britney spears," "mark zuckerberg," and "bin carding," along with searches for explicit content. We found that the attackers did not limit their search for specific terms to individual accounts—they also searched other accounts.

To understand the "spread" of search terms, we counted the number of accounts that recorded the top search terms. Table 4 shows the number of accounts in whose logs the top searched words appeared. Note that some words showed up multiple times in an individual account and were counted each time. For instance, if we find the search terms "debates: atheism" and "debates: atheism and religions" in the logs of a particular account, we count "atheism" twice and "religion" once. Note that searches fail to return the expected content in Facebook test accounts since they are disconnected from the regular Facebook graph. Table 4 indicates that searchers proceed to try other accounts when their first choice fails to return search results.

4.8 Social Chatter

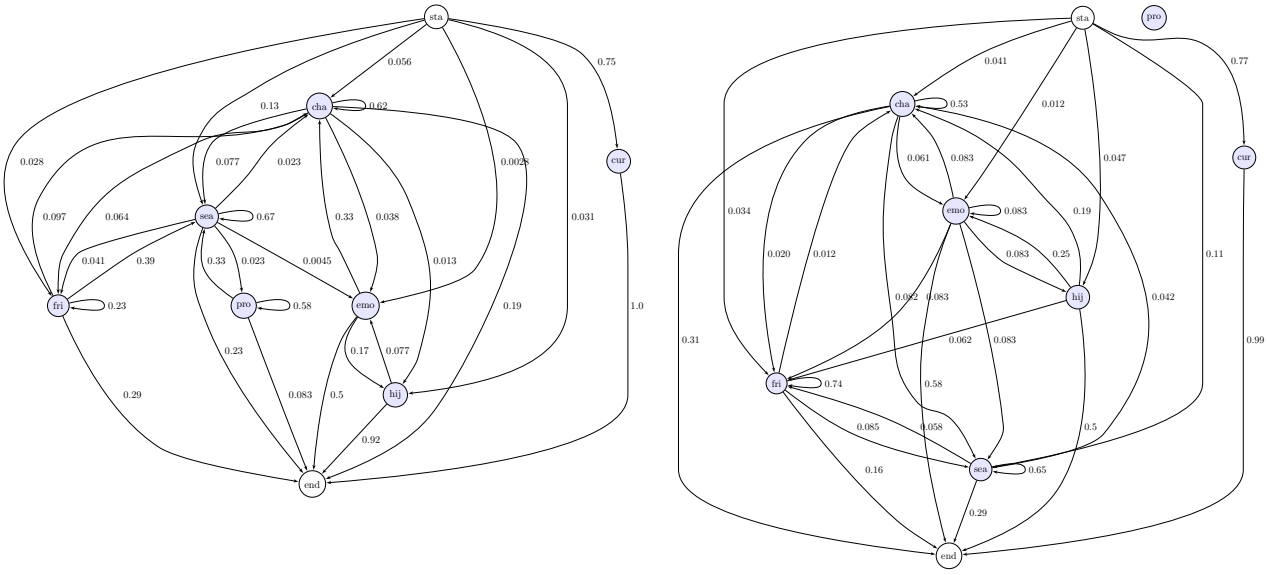
Recall that Table 1 shows that chatty accesses were responsible for 11% of all recorded actions. We observed chatty behavior in 45 accounts. These comprise attempted group calls, "waves," private messages, and posts on own timeline and other timelines. We found some posts warning account owners about leaked credentials (unknown to the posters, we leaked honey credentials intentionally). We did not observe any post containing phishing or malware-laden links; Facebook actively blocks such activity or retroactively hides previously-posted malicious content. To observe the top words in the chatty text corpus, we once again applied the word-counting technique outlined in Section 4.7. The top ten chatty words are shown in Table 3 (right-hand side).

Note the presence of the word "fake" in Table 3; some comments stated that the accounts were fake (only within 4 accounts). This shows that a handful of criminals were not fooled. Despite this, we still collected useful information about them, at least, about their authentication actions and subsequent activity. Note that we designed the accounts to appear realistic. Hence, we succeeded in collecting activity data anyway. Since we leaked credentials repeatedly on paste sites (see Section 3.3), which do not have comment fields or other direct feedback mechanisms, it is unlikely that those who detected the fakeness of the accounts disclosed this to other criminals, aside from the comments they posted in some accounts (which we could delete if we wanted to).



(a) Teen. (b) Adult.

Figure 8: Activity sequences per age range. Node *sta* means "start" and indicates the entry point to the graph, not an access type. Similarly, node *end* indicates the exit point from the graph, not an access type.



(a) Male. (b) Female.

Figure 9: Activity sequences per gender. Node *sta* means "start" and indicates the entry point to the graph, not an access type. Similarly, node *end* indicates the exit point from the graph, not an access type. Note the disconnected profile editor node (*pro*) in the graph of female accounts.

Table 5: Browsers in accesses. A small fraction of accesses were apparently made using PhantomJS.

Browser	*Instances*	*Percentage*
Chrome	134	41.6
Firefox	119	37.0
Android Browser	25	7.8
Unknown Browser	20	6.2
Edge	10	3.1
Safari	7	2.2
Opera	4	1.2
PhantomJS	2	0.6
Internet Explorer	1	0.3
Total	**322**	**100.0**

Table 6: Operating systems in accesses.

OS	*Instances*	*Percentage*
Windows	210	65.2
Android	60	18.6
Unknown OS	22	6.8
MacOS	14	4.3
Linux	10	3.1
iPhone iOS	6	1.9
Total	**322**	**100.0**

Note that we used an automatic language translation tool, the *Googletrans* API [5], to translate non-English textual data to English prior to processing (in Sections 4.7 and 4.8).

4.9 System Configuration of Accesses

Leveraging the user-agent string information available in DYI archives, we extracted browser and operating system information from the observed accesses. A wide range of browsers and operating systems were used to access the honey accounts. Table 5 shows a summary of those browsers. Chrome and Firefox dominate the table of browsers, at 42% and 37% respectively. A small fraction of accesses (less than 1%) were apparently made using PhantomJS,[3] a browser automation tool. This suggests that some connections may have been made automatically.

Table 6 shows an overview of the operating systems on the devices that connected to honey accounts. Windows and Android dominate the list (65% and 19% respectively). A small fraction of accesses were also made with iPhones. Note that these are merely indicators: user-agent strings can be changed, and as such are not reliable.

[3] https://phantomjs.org/

4.10 Origin of Accesses

In total, we observed 415 IP addresses (IPv4 and IPv6 addresses) from 53 countries. Of these IP addresses, 39 were TOR exit nodes. It is possible that some of the remaining IP addresses were proxies or VPN nodes. To understand the geographical locations that accesses originated from, we extracted all IP addresses associated with accesses from the DYI archives. We then carried out IP geolocation using *IP-API* [8], an IP geolocation service that provides timezone and location information, given one or more IP addresses. Figure 10 shows a world map with markers showing the locations that accesses originated from. As the map indicates, connections originated from many locations around the world. Interesting patterns include activity along the coasts of the Americas, a dense cluster in Europe, and activity in India. No access originated from China—note that Facebook is banned in China. It is possible that criminals connected to some accounts via proxies or VPNs. However, we did not observe any evidence that confirms or refutes this.

5 Discussion

In this section, we first discuss the implications of our results, in particular putting them in the context of previous research on how age and gender affect cybercrime victimization. We then discuss the limitations of our study and propose some ideas for future work.

5.1 Characterizing Attacker Activity

According to our results, search activity, chatty activity, and modification of friend lists (adding or removing friends) constitute the top three types of actions that were observed in the accounts (apart from logging in). Given the social nature of Facebook accounts, the manipulation of friend lists could potentially be an approach to extend the reach of malicious activity beyond the affected accounts. In other words, when the attacker adds new contacts to an existing friend list, they could eventually send phishing messages or scam messages to new or existing contacts.

When criminals connected to our test Facebook accounts, they mostly wrote private messages, public posts, and attempted to search for information. Messages and posts were exchanged across the accounts. We did not find any bulk spam or malware links in them. However, we observed the occurence of racist and abusive content. This matches what was reported by prior research on compromised accounts, which found that sending spam and malicious messages in general was not the main reason why miscreants breached email accounts, but that instead the most common activity was to search for sensitive information in those accounts [18, 42]. This makes even more sense for Facebook accounts, because beyond messaging capabilities these accounts present many

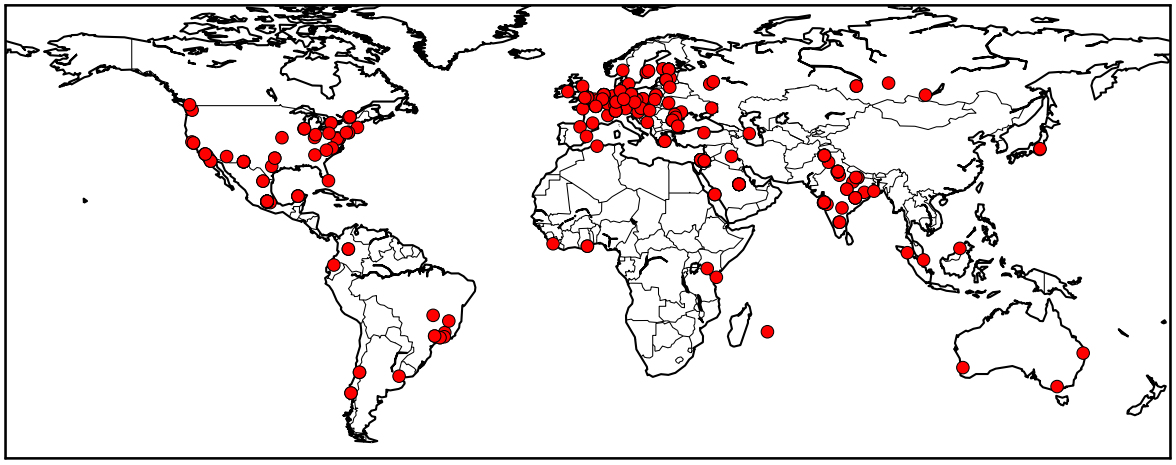


Figure 10: Markers indicate the origins of connections to test accounts.

other features. Hence, it is logical to expect a wider variety of actions, as our findings reveal, than spamming. However, it is important to note that some attackers may have intended to send malicious content later in the conversations, had the target account responded, as seen in fraud cases [31, 50, 58]. Hence, in future experiments, it may be helpful to incorporate chatbots in honey accounts to automatically respond to messages sent by attackers.

Finally, the search terms that were recorded in the test accounts (Section 4.7) reveal a wide variety of themes of interest in the accounts. Modeling benign and malicious search activity (i.e., legitimate users versus criminals) could possibly help to distinguish and mitigate malicious activity in compromised accounts. We leave that to future work since we do not have baseline search data for benign users, and would need such baseline data to develop robust automated mitigation systems.

5.2 Demographic Factors

We show that demographic attributes of accounts (age range and gender) influence the activity of criminals in compromised accounts. In other words, we show a significant relationship between account demographics and the actions that criminals carry out in the accounts. Similarly, we show that sequences of actions differ in the accounts per age range and gender, with the exception of search activity sequences. This indicates that the demographic attributes of accounts should be taken into consideration when building tools to automatically detect malicious activity in stolen social accounts. The modeling of differences in action sequences across account demographics led to interesting findings in itself, and could potentially be extended into techniques to distinguish malicious activity from benign activity (for instance, by a large online service). However, caution must be exercised to avoid user profiling while exploring this potential solution to malicious activity.

In addition to the differences in activity sequences, we observed other distinctions across account demographics in the types of actions that attackers carried out. For instance, the attackers of teen accounts were chattier than those of adult accounts, while the attackers of adult accounts were more interested in adding or removing friends than those of teen accounts. We also observed differences in male and female accounts, especially in profile editing and friend list modification activity. Again, these show that account demographics play an important role in determining the actions that criminals carry out in stolen social accounts. This knowledge could potentially be helpful for large online services seeking to improve their detection systems.

Next, we put our results in the context of prior research literature. Although our work is the first one studying criminal activity in compromised Facebook accounts, it is helpful to understand how our results compare to previous research in cybercrime and online abuse victimization. Note that a significant amount of work was conducted in understanding demographics factors that influence people's likelihood of falling for phishing [41, 45], malware [17, 37], or fraud [60]. In our work, we are interested in understanding what attackers do once they compromise a Facebook account, and therefore instead look at research that studied the type of malicious activity that different demographics are likely to experience online.

Age. The teen accounts in our dataset recorded more profile editing and chatty behavior than adult accounts. This is in line with previous work showing that younger people are more likely to receive online abuse and harassment [51], as well as previous work showing that younger people have a higher chance of being victimized by cybercrime [40]. In

our dataset, the adult accounts suffered much more from the addition or removal of friends than teen accounts. A possible explanation for this is that previous research reported that older people are disproportionately affected by online fraud, for example romance scams [30, 50, 58]. It is possible that the attackers were trying to reach potential victims by making friends requests. Unfortunately, since our IRB protocol did not allow us to interact with criminals, we could not reply to any conversation and understand the purpose of the connection.

Gender. In our dataset, female accounts received more friend requests than male accounts (126 vs 31). A potential reason is that multiple studies reported that women are more likely to receive online abuse like sexual harassment [22, 36, 51]. It is possible that these malicious actions had the goal of harassing the victim, whether sexually or otherwise. Another possible explanation lies in the fact that previous research observed that fraudsters engaging in romance scams were often posing as older men and targeting women [30, 50, 58]. It is possible that cybercriminals were aiming to contact women's accounts to potentially defraud them. Since our IRB protocol did not allow us to interact with criminals, we could not reply to the messages received by our accounts to better understand the intentions of the attacker.

In our dataset, male accounts encountered more search activity than female accounts. Previous research showed that cybercriminals often search stolen accounts for sensitive information that might enable them to mount additional attacks (e.g., financial information) [18, 42]. If this was the intention of cybercriminals, the predilection for male accounts can be explained by previous work that showed that men are more likely to be victimized by scams [59].

At the same time, we observe instances of male accounts for which attackers modified their profile, while female accounts recorded no profile edits. The reason for this could be that the attackers did not find a profitable way of monetizing these accounts, and decided to vandalize them instead. This is in line with previous research that showed that attackers disrupt online resources (e.g., online accounts and online documents) when they cannot find a better way to exploit them [34, 42].

Key Lesson. Cybercriminals orchestrate attack activity differently in online accounts that belong to men, women, adults, and teens, as shown in our work. This observation is further reinforced by the existing research literature which shows that age, gender, and personality traits are factors that influence cybercrime victimization, as previously discussed in Section 2. In view of this, mitigation systems and interventions should be customized along these different groups. In addition, there is a need to evolve security systems away from defending the "average user," who does not really exist [24], towards adaptive mitigation systems that address the demographic-based nature of groups of users.

5.3 Limitations and Future Work

Here, we highlight some limitations of our work and suggest potential future directions. Our study articulates a number of research hypotheses and uses statistical tests to back them up. However, we acknowledge that our experiment only covers the threat of account credentials leaked on paste sites, and might not be representative of all compromises. We discussed threats to the validity of our work in Section 3.4.

We acknowledge that our data is no longer as fresh as it could possibly be (it was collected in 2018). To the best of our knowledge, however, ours is the first study exploring demographic risk factors in Facebook accounts. While the campaigns carried out by attackers might have since changed, we argue that their motives are still the same and that these demographic risk factors still hold.

Prior to the experiments, we wrote some publicly-available data to the timelines of the test accounts and wrote no private messages. On the other hand, real-world Facebook accounts often contain private messages. We acknowledge that this may affect the perception of criminals on visiting the test accounts. In future work, we plan to incorporate private messages to further approximate real accounts.

In the course of experiments, private messages and timeline posts were written to some honey accounts by criminals. We did not respond to any of them as dictated by our IRB protocol. This may have affected the perception of the criminals: such activity in real accounts could elicit responses from account owners. Additionally, this limited our visibility on the attackers' intentions, since we did not observe anything beyond the initial messages. In the future, it would be interesting to incorporate chatbots that will autorespond to messages; this will further deepen the impression of "lived-in" accounts (realism), but also has ethical implications.

We studied only two demographic attributes: age range and gender. In the future, we propose investigating more attributes, for instance, occupation, political leanings, and religious beliefs, among others. In addition to understanding criminal activity in stolen accounts, such attributes may also help the research community to investigate other problems—especially cyberbullying and targeted attacks. Finally, to understand chain attacks, we will store authentication tokens to other services in honey accounts, within private messages, to observe how criminals would misuse them.

6 Conclusion

We presented the first large-scale honeypot system for monitoring compromised Facebook accounts. We created more than 1000 realistic Facebook accounts, incorporated demographic attributes in them, and observed attacker behavior in them, for six months. We showed that those demographic attributes influenced the actions of attackers in the accounts and characterized the activity of attackers in stolen social ac-

counts. These findings will help the research community to gain a deeper understanding of compromised online accounts towards the development of better security systems.

Acknowledgments

We would like to thank the anonymous reviewers for their comments. This work received support from a Facebook Secure-the-Internet research gift. We would like to thank Mark Atherton for his help during the early stages of this work. We were partially supported by the National Science Foundation (NSF) under Grant 1942610. Most parts of this work were completed while Jeremiah Onaolapo was at University College London (UCL) with the support of the Petroleum Technology Development Fund (PTDF) of Nigeria.

References

[1] Accessing & downloading your information. https://www.facebook.com/help/1701730696756992. Accessed: 2020-09-18.

[2] The best free stock photos & videos shared by talented creators. https://www.pexels.com/. Accessed: 2020-09-18.

[3] Developer docs. https://developer.twitter.com/en/docs. Accessed: 2020-09-18.

[4] Find your inspiration. https://www.flickr.com/. Accessed: 2020-09-18.

[5] Googletrans: Free and unlimited Google translate API for Python. https://py-googletrans.readthedocs.io/en/latest. Accessed: 2020-09-18.

[6] Information (on Facebook test accounts). https://www.facebook.com/whitehat/info/. Accessed: 2020-09-18.

[7] The Internet's source of freely-usable images. https://unsplash.com/. Accessed: 2020-09-18.

[8] IP geolocation API. https://ip-api.com. Accessed: 2020-09-18.

[9] nltk.stem package. https://www.nltk.org/api/nltk.stem.html. Accessed: 2020-09-18.

[10] nltk.tokenize package. https://www.nltk.org/api/nltk.tokenize.html. Accessed: 2020-09-18.

[11] Random user generator. https://randomuser.me/. Accessed: 2020-09-18.

[12] Stunning free images & royalty free stock. https://pixabay.com/. Accessed: 2020-09-18.

[13] D. Alvarez-Melis and M. Saveski. Topic modeling in Twitter: Aggregating tweets by conversations. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, 2016.

[14] T. Barron and N. Nikiforakis. Picky attackers: Quantifying the role of system properties on intruder behavior. In *Annual Computer Security Applications Conference (ACSAC)*, 2017.

[15] H. Binsalleeh, T. Ormerod, A. Boukhtouta, P. Sinha, A. Youssef, M. Debbabi, and L. Wang. On the analysis of the Zeus botnet crimeware toolkit. In *Privacy, Security and Trust (PST)*, 2010.

[16] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: When bots socialize for fame and money. In *Annual Computer Security Applications Conference (ACSAC)*, 2011.

[17] A. M. Bossler and T. J. Holt. On-line activities, guardianship, and malware infection: An examination of routine activities theory. *International Journal of Cyber Criminology*, 3(1), 2009.

[18] E. Bursztein, B. Benko, D. Margolis, T. Pietraszek, A. Archer, A. Aquino, A. Pitsillidis, and S. Savage. Handcrafted fraud and extortion: Manual account hijacking in the wild. In *ACM Internet Measurement Conference (IMC)*, 2014.

[19] P. Cao, Y. Wu, S. S. Banerjee, J. Azoff, A. Withers, Z. T. Kalbarczyk, and R. K. Iyer. CAUDIT: continuous auditing of SSH servers to mitigate brute-force attacks. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2019.

[20] J. DeBlasio, S. Savage, G. M. Voelker, and A. C. Snoeren. Tripwire: Inferring Internet site compromise. In *ACM Internet Measurement Conference (IMC)*, 2017.

[21] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2006.

[22] M. Duggan. Online harassment 2017. 2017.

[23] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. COMPA: Detecting compromised accounts on social networks. In *Symposium on Network and Distributed System Security (NDSS)*, 2013.

[24] S. Egelman and E. Peer. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, pages 16–28, 2015.

[25] R. A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[26] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *ACM Internet Measurement Conference (IMC)*, 2010.

[27] W. G. Halfond, J. Viegas, and A. Orso. A classification of SQL-injection attacks and countermeasures. In *IEEE International Symposium on Secure Software Engineering*, 2006.

[28] X. Han, N. Kheir, and D. Balzarotti. PhishEye: Live monitoring of sandboxed phishing kits. In *ACM Conference on Computer and Communications Security (CCS)*, 2016.

[29] B. Henson, B. W. Reyns, and B. S. Fisher. Does gender matter in the virtual world? Examining the effect of gender on the link between online social network activity, security and interpersonal victimization. *Security Journal*, 26(4):315–330, 2013.

[30] J. Huang, G. Stringhini, and P. Yong. Quit playing games with my heart: Understanding online dating scams. In *Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*. 2015.

[31] J. Isacenkova, O. Thonnard, A. Costin, D. Balzarotti, and A. Francillon. Inside the scam jungle: A closer look at 419 scam email operations. In *Security and Privacy Workshops (SPW)*, 2013.

[32] A. Kedrowitsch, D. D. Yao, G. Wang, and K. Cameron. A first look: Using Linux containers for deceptive honeypots. In *Workshop on Automated Decision Making for Active Cyber Defense*, 2017.

[33] A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[34] M. Lazarov, J. Onaolapo, and G. Stringhini. Honey sheets: What happens to leaked Google spreadsheets? In *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, 2016.

[35] K. Lee, J. Caverlee, and S. Webb. The social honeypot project: Protecting online communities from spammers. In *World Wide Web Conference (WWW)*, 2010.

[36] A. Lenhart, M. Ybarra, K. Zickuhr, and M. Price-Feeney. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute, 2016.

[37] F. L. Lévesque, J. M. Fernandez, and D. Batchelder. Age and gender as independent risk factors for malware victimisation. *Electronic Visualisation and the Arts (EVA 2017)*, pages 1–14, 2017.

[38] F. L. Lévesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *ACM Conference on Computer and Communications Security (CCS)*, 2013.

[39] J. Leyden. Rockyou hack reveals easy-to-crack passwords. https://www.theregister.co.uk/2010/01/21/lame_passwords_exposed_by_rockyou_hack/. Accessed: 2020-09-18.

[40] M. Näsi, A. Oksanen, T. Keipi, and P. Räsänen. Cybercrime victimization among young people: a multi-nation study. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 16(2):203–210, 2015.

[41] D. Oliveira, H. Rocha, H. Yang, D. Ellis, S. Dommaraju, M. Muradoglu, D. Weir, A. Soliman, T. Lin, and N. Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2017.

[42] J. Onaolapo, E. Mariconti, and G. Stringhini. What happens after you are pwnd: Understanding the use of leaked webmail credentials in the wild. In *ACM Internet Measurement Conference (IMC)*, 2016.

[43] E. M. Redmiles. "Should I Worry?" A Cross-Cultural Examination of Account Security Incident Response. In *IEEE Symposium on Security and Privacy*, 2019.

[44] C. Rossow, C. J. Dietrich, C. Grier, C. Kreibich, V. Paxson, N. Pohlmann, H. Bos, and M. van Steen. Prudent practices for designing malware experiments: Status quo and outlook. In *IEEE Symposium on Security and Privacy*, 2012.

[45] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2010.

[46] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

[47] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna. Your botnet is my botnet: Analysis of a botnet takeover. In *ACM Conference on Computer and Communications Security (CCS)*, 2009.

[48] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2011.

[49] G. Stringhini and O. Thonnard. That ain't you: Blocking spearphishing through behavioral modelling. In *Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2015.

[50] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty. Automatically dismantling online dating fraud. *arXiv preprint arXiv:1905.12593*, 2019.

[51] K. Thomas, D. Akhave, M. Bailey, D. Boneh, E. Burztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. SoK: Hate, harassment, and the changing landscape of online abuse. In *IEEE Symposium on Security and Privacy*, 2021.

[52] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of Twitter spam. In *ACM Internet Measurement Conference (IMC)*, 2011.

[53] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson. Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse. In *USENIX Security Symposium*, 2013.

[54] S. G. A. van de Weijer and E. R. Leukfeldt. Big five personality traits of cybercrime victims. *Cyberpsychology Behav. Soc. Netw.*, 20(7):407–412, 2017.

[55] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang. Targeted online password guessing: An underestimated threat. In *ACM Conference on Computer and Communications Security (CCS)*, 2016.

[56] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao. You are how you click: Clickstream analysis for Sybil detection. In *USENIX Security Symposium*, 2013.

[57] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In *Conference on Email and Anti-Spam (CEAS)*, 2008.

[58] M. T. Whitty. Anatomy of the online dating romance scam. *Security Journal*, 28(4):443–455, 2015.

[59] M. T. Whitty. Is there a scam for everyone? Psychologically profiling cyberscam victims. *European Journal on Criminal Policy and Research*, pages 1–11, 2020.

[60] M. T. Whitty and T. Buchanan. The online romance scam: A serious cybercrime. *CyberPsychology, Behavior, and Social Networking*, 15(3):181–183, 2012.

[61] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. Uncovering social network Sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1):2:1–2:29, 2014.