Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse

Kurt Thomas[†◇] *Damon McCoy*[‡] *Chris Grier*[†*] *Alek Kolcz*[◇] *Vern Paxson*[†*]
[†]*University of California, Berkeley* [‡]*George Mason University*
[*]*International Computer Science Institute* [◇]*Twitter*
{*kthomas, grier, vern*}*@cs.berkeley.edu* *mccoy@cs.gmu.edu* *ark@twitter.com*

Abstract

As web services such as Twitter, Facebook, Google, and Yahoo now dominate the daily activities of Internet users, cyber criminals have adapted their monetization strategies to engage users within these walled gardens. To facilitate access to these sites, an underground market has emerged where fraudulent accounts – automatically generated credentials used to perpetrate scams, phishing, and malware – are sold in bulk by the thousands. In order to understand this shadowy economy, we investigate the market for fraudulent Twitter accounts to monitor prices, availability, and fraud perpetrated by 27 merchants over the course of a 10-month period. We use our insights to develop a classifier to retroactively detect several million fraudulent accounts sold via this marketplace, 95% of which we disable with Twitter's help. During active months, the 27 merchants we monitor appeared responsible for registering 10–20% of all accounts later flagged for spam by Twitter, generating $127–459K for their efforts.

1 Introduction

As web services such as Twitter, Facebook, Google, and Yahoo now dominate the daily activities of Internet users [1], cyber criminals have adapted their monetization strategies to engage users within these walled gardens. This has lead to a proliferation of *fraudulent accounts* – automatically generated credentials used to disseminate scams, phishing, and malware. Recent studies from 2011 estimate at least 3% of active Twitter accounts are fraudulent [29]. Facebook estimates its own fraudulent account population at 1.5% of its active user base [13], and the problem extends to major web services beyond just social networks [14].

The complexities required to circumvent registration barriers such as CAPTCHAs, email confirmation, and IP blacklists have lead to the emergence of an underground market that specializes in selling fraudulent accounts in bulk. *Account merchants* operating in this space brazenly advertise: a simple search query for "buy twitter accounts" yields a multitude of offers for fraudulent Twitter credentials with prices ranging from $10–200 per thousand. Once purchased, accounts serve as stepping stones to more profitable spam enterprises that degrade the quality of web services, such as pharmaceutical spam [17] or fake anti-virus campaigns [25].

In this paper we describe our investigation of the underground market profiting from Twitter credentials to study how it operates, the impact the market has on Twitter spam levels, and exactly how merchants circumvent automated registration barriers.[1] In total, we identified and monitored 27 account merchants that advertise via web storefronts, blackhat forums, and freelance labor sites. With the express permission of Twitter, we conducted a longitudinal study of these merchants and purchased a total of 121,027 fraudulent Twitter accounts on a bi-weekly basis over ten months from June, 2012 – April, 2013. Throughout this process, we tracked account prices, availability, and fraud in the marketplace. Our findings show that merchants thoroughly understand Twitter's existing defenses against automated registration, and as a result can generate thousands of accounts with little disruption in availability or instability in pricing.

In order to fulfill orders for fraudulent Twitter accounts, we find that merchants rely on CAPTCHA solving services; fraudulent email credentials from Hotmail, Yahoo, and mail.ru; and tens of thousands of hosts located around the globe to provide a diverse pool of IP addresses

[1] Our study is limited to Twitter, as we were unable to acquire permission to conduct our research from other companies we saw being abused.

to evade blacklisting and throttling. In turn, merchants stockpile accounts months in advance of their sale, where "pre-aged" accounts have become a selling point in the underground market. We identify which registration barriers effectively increase the price of accounts and summarize our observations into a set of recommendations for how web services can improve existing automation barriers to increase the cost of fraudulent credentials.

Finally, to estimate the overall impact the underground market has on Twitter spam we leveraged our understanding of how merchants abuse the registration process in order to develop a classifier that retroactively detects fraudulent accounts. We applied our classifier to all accounts registered on Twitter in the last year and identify several million suspected fraudulent accounts generated and sold via the underground market. During active months, the 27 merchants we monitor appeared responsible for registering 10–20% of all accounts later flagged by Twitter as spam. For their efforts, the merchants generated an estimated total revenue between $127,000–$459,000 from the sale of accounts.

With Twitter's cooperation, we disable 95% of all fraudulent accounts registered by the merchants we track, including those previously sold but not yet suspended for spamming. Throughout the suspension process, we simultaneously monitor the underground market for any fallout. While we do not observe an appreciable increase in pricing or delay in merchants delivering new accounts, we find 90% of all purchased accounts immediately after our action are suspended on arrival. We are now actively working with Twitter to integrate our defense into their real-time detection framework to help prevent abusive signups.

In summary, we frame our contributions as follows:

- We perform a 10 month longitudinal study of 27 merchants profiting from the sale of Twitter accounts.
- We develop a classifier based on registration signals that detects several million fraudulent accounts that merchants sold to generate $127,000–$459,000 in revenue.
- We investigate the impact that the underground market has on Twitter spam levels and find 10–20% all spam accounts originate from the merchants we study.
- We investigate the failures of existing automated registration barriers and provide a set of recommendations to increase the cost of generating fraudulent accounts.

2 Background

Fraudulent accounts are just a single facet of the menagerie of digital criminal goods and services for sale in the underground market. We provide an overview of previous investigations into the digital blackmarket, outline the role that account abuse plays in this space, and summarize existing strategies for detecting spam and abuse. Finally, in order to carry out our investigation of the market for fraudulent Twitter accounts, we adhere to a strict set of legal and ethical guidelines set down by our institutions and by Twitter, documented here.

2.1 Underground Market

At the center of the for-profit spam and malware ecosystem is an underground market that connects Internet miscreants with parties selling a range of specialized products and services including spam hosting [2, 11], CAPTCHA solving services [19], pay-per-install hosts [4], and exploit kits [9]. Even simple services such as garnering favorable reviews or writing web page content are for sale [21, 31]. Revenue generated by miscreants participating in this market varies widely based on business strategy, with spam affiliate programs generating $12–$92 million [17] and fake anti-virus scammers $5-116 million [25] over the course of their operations.

Specialization within this ecosystem is the norm. Organized criminal communities include carders that siphon credit card wealth [7]; email spam affiliate programs [16]; and browser exploit developers and traffic generators [9]. The appearance of account merchants is yet another specialization where sellers enable other miscreants to penetrate walled garden services, while at the same time abstracting away the complexities of CAPTCHA solving, acquiring unique emails, and dodging IP blacklisting. These accounts can then be used for a multitude of activities, outlined below, that directly generate a profit for miscreants.

2.2 Impact of Fraudulent Accounts

Miscreants leverage fraudulent social networking accounts to expose legitimate users to scams, phishing, and malware [8, 10]. Spam monetization relies on both grey-market and legitimate affiliate URL programs, ad syndication services, and ad-based URL shortening [29]. Apart from for-profit activities, miscreants have also leveraged fraudulent accounts to launch attacks from within Twitter for the express purposes of censoring political speech [28]. All of these examples serve to illustrate the deleterious effect that fraudulent accounts have on social networks and user safety.

2.3 Spam Detection Strategies

The pervasive nuisance of spam in social networks has lead to a multitude of detection strategies. These include analyzing social graph properties of sybil accounts [6, 33, 34], characterizing the arrival rate and distribution of posts [8], analyzing statistical properties of account profiles [3, 26], detecting spam URLs posted by accounts [27], and identifying common spam redirect paths [15]. While effective, all of these approaches rely on *at-abuse* time metrics that target strong signals such as sending a spam URL or forming hundreds of relationships in a short period. Consequently, at-abuse time classifiers delay detection until an attack is underway, potentially exposing legitimate users to spam activities before enough evidence of nefarious behavior triggers detection. Furthermore, dormant accounts registered by account merchants will go undetected until miscreants purchase the accounts and subsequently send spam. Overcoming these shortcomings requires *at-registration* abuse detection that flags fraudulent accounts during the registration process before any further interaction with a web service can occur.

2.4 Ethical Considerations

Our study hinges on infiltrating the market for fraudulent Twitter credentials where we interact with account merchants and potentially galvanize the abuse of Twitter. We do so with the express intent of understanding how sellers register accounts and to disrupt their future efforts, but that does not allay our legal or ethical obligations. Prior to conducting our study, we worked with Twitter and our institutions to set down guidelines for interacting with merchants. A detailed summary of the restrictions placed on our study is available in Appendix A

3 Marketplace for Twitter Accounts

We infiltrate the market for Twitter accounts to understand its organization, pricing structure, and the availability of accounts over time. Through the course of our study, we identify 27 account merchants (or sellers) whom we purchase from on a bi-weekly basis from June, 2012 – April, 2013. We determine that merchants can provide thousands of accounts within 24 hours at a price of $0.02 – $0.10 per account.

3.1 Identifying Merchants

With no central operation of the underground market, we resort to investigating common haunts: advertisements via search engines, blackhat forums such as *blackhatworld.com*, and freelance labor pages including Fiverr and Freelancer [20, 21]. In total, we identify a disparate group of 27 merchants. Of these, 10 operate their own websites and allow purchases via automated forms, 5 solicit via blackhat forums, and 12 advertise via freelance sites that take a cut from sales. Advertisements for Twitter accounts range in offerings from credentials for accounts with no profile or picture, to "pre-aged" accounts[2] that are months old with unique biographies and profile data. Merchants even offer 48 hours of support, during which miscreants can request replacements for accounts that are dysfunctional. We provide a detailed breakdown of the merchants we identify and their source of solicitation in Table 1. We make no claim our search for merchants is exhaustive; nevertheless, the sellers we identify provide an insightful cross-section of the varying levels of sophistication required to circumvent automated account registration barriers, outlined in detail in Section 4.

3.2 Purchasing from Merchants

Once we identify a merchant, we place an initial test purchase to determine the authenticity of the accounts being sold. If genuine, we then determine whether to repeatedly purchase from the merchant based on the quality of accounts provided (discussed in Section 4) and the overall impact the seller has on Twitter spam (discussed in Section 6). As such, our purchasing is an iterative process where each new set of accounts improves our understanding of the market and subsequently directs our investigation.

Once we vet a merchant, we conduct purchases on a bi-weekly basis beginning in June, 2012 (at the earliest) up to the time of our analysis in April, 2013, detailed in Table 1. We note that purchasing at regular intervals is not always feasible due to logistical issues such as merchants delaying delivery or failing to respond to requests for accounts. In summary, we place 144 orders (140 of which merchants successfully respond to and fulfill) for a total of 120,019 accounts. Purchases typically consist of a bulk order for 1,000 accounts, though sellers on Fiverr operate in far less volume.

Throughout this process, we protect our identity from merchants by using a number of email and Skype pseudonyms. We conduct payments through multiple identities tied to PayPal, WebMoney, and pre-paid credit

[2]Pre-aged accounts allow miscreants to evade heuristics that disable newly minted accounts based upon weak, early signs of misbehavior. In contrast, in order to limit the impact on legitimate users, disabling older accounts only occurs in the face of much stronger signals of maleficence.

Merchant	Period	#	Accts	Price
alexissmalley†	06/12–03/13	14	13,000	$4
naveedakhtar†	01/13–03/13	4	2,044	$5
truepals†	02/13–03/13	3	820	$8
victoryservices†	06/12–03/13	15	15,819	$6
webmentors2009†	10/12–03/13	9	9,006	$3–4
buuman¤	10/12–10/12	1	75	$7
danyelgallu¤	10/12–10/12	1	74	$7
denial93¤	10/12–10/12	1	255	$20
formefor¤	09/12–11/12	3	408	$2–10
ghetumarian¤	09/12–10/12	3	320	$4–5
jackhack08¤	09/12–09/12	2	755	$1
kathlyn¤	10/12–10/12	1	74	$7
smokinbluelady¤	08/12–08/12	1	275	$2
twitfollowers¤	10/12–10/12	1	80	$6
twitter007¤	10/12–10/12	1	75	$7
kamalkishover◇	06/12–03/13	14	12,094	$4–7
shivnagsudhakar◇	06/12–06/12	1	1,002	$4
accs.biz‡	05/12–03/13	15	17,984	$2–3
buyaccountsnow.com‡	06/12–11/12	8	7,999	$5–8
buyaccs.com‡	06/12–03/13	14	13,794	$1–3
buytwitteraccounts.biz‡	09/12–10/12	3	2,875	$5
buytwitteraccounts.info‡	10/12–03/13	9	9,200	$3–4
dataentryassistant.com‡	10/12–03/13	9	5,498	$10
getbulkaccounts.com‡	09/12–09/12	1	1,000	$2
quickaccounts.bigcartel‡	11/12–11/12	2	1,501	$3
spamvilla.com‡	06/12–10/12	3	2,992	$4
xlinternetmarketing.com‡	10/12–10/12	1	1,000	$7
Total	05/12–03/13	140	120,019	$1–20

Table 1: List of the merchants we track, the months monitored, total purchases performed (#), accounts purchased, and the price per 100 accounts. Source of solicitations include blackhat forums†, Fiverr¤, and Freelancer◇ and web storefronts‡.

cards. Finally, we access all web content on a virtual machine through a network proxy.

3.3 Account Pricing & Availability

Prices through the course of our analysis range from $0.01 to $0.20 per Twitter account, with a median cost of $0.04 for all merchants. Despite the large overall span, prices charged by individual merchants remain roughly stable. Table 1 shows the variation in prices for six merchants we tracked over the longest period of time. Price hikes are a rare occurrence and no increase is more than $0.03 per account. So long as miscreants have money on hand, availability of accounts is a non-issue. Of the orders we placed, merchants fulfilled 70% in a day and 90% within 3 days. We believe the stable pricing and ready availability of fraudulent accounts is a direct result

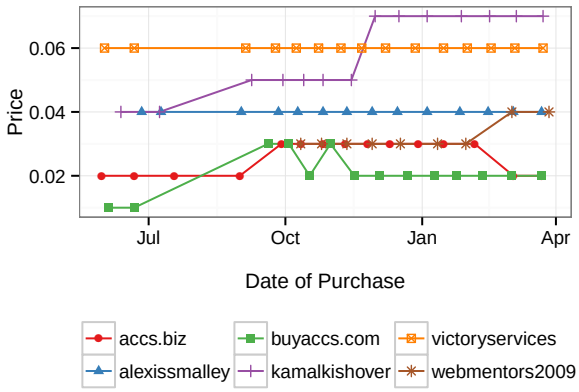

Figure 1: Variation in prices over time for six merchants we track over the longest period of time.

of minimal adversarial pressures on account merchants, a hypothesis we explore further in Section 4.

3.4 Other Credentials For Sale

Our permission to purchase accounts is limited to Twitter credentials, but many of the merchants we interact with also sell accounts for Facebook, Google, Hotmail, and Yahoo. We compare prices between web services, but note that as we cannot vet non-Twitter credentials, some prices may represent scams.

Facebook Prices for Facebook accounts range from $0.45–1.50 per *phone verified account* (PVA) and $0.10 for non-PVA accounts. Phone verification requires that miscreants tie a SIM card to a newly minted Facebook account and verify the receipt of a text message, the complexities of which vastly increase the price of an account.[3] For those sellers that advertise their registration process, SIM cards originate from Estonia or Ukraine.

Google Prices for Google PVA accounts range from $0.03–0.50 per account.

Hotmail Prices for Hotmail accounts cost $0.004 – 0.03 per account, a steep reduction over social networking or PVA credentials. We see similar prices for a multitude of web mail providers, indicating that email accounts are in demand and cheaper to create.

Yahoo Yahoo accounts, like Hotmail, are widely available, with prices ranging from $0.006 – 0.015 per account.

[3] Advertisements that we encountered for phone verification services ranged in price from $.10 – $.15 per verification for bulk orders of 100,000 verifications, and $.25 per verification for smaller orders.

Merchant	Reaccessed	Resold
getbulkaccounts.com	100%	100%
formefor	100%	99%
denial93	100%	97%
shivnagsudhakar	98%	98%
quickaccounts.bigcartel.com	67%	64%
buytwitteraccounts.info	39%	31%
ghetumarian	30%	28%
buytwitteraccounts.biz	20%	18%
jackhack08	12%	11%
buyaccountsnow.com	10%	1%
kamalkishover	8%	0%
buyaccs.com	7%	4%
alexissmalley	6%	0%
victoryservices	3%	2%
Total	10%	6%

Table 2: List of dishonest merchants that reaccessed and resold credentials we purchased to other parties.

3.5 Merchant Fraud

Operating in the underground market is not without risk of fraud and dishonesty on the part of account merchants. For instance, eight of the merchants we contacted attempted to sell us a total of 3,317 duplicate accounts. One merchant even schemed to resell us the same 1,000 accounts three times. For those merchants willing to honor their "48 hours of support", we requested replacement accounts for duplicates, bringing our account total up to 121,027 unique credentials.

Apart from duplicate credentials, some merchants were quick to resell accounts we purchased to third parties. In order to detect resales, we coordinate with Twitter to monitor all successful logins to accounts we purchase after they come under our control. We denote these accounts *reaccessed*. We repeat this same process to detect new tweets or the formation of relationships. Such behaviors should only occur when an account changes hands to a spammer, so we denote these accounts as *resold*. Such surreptitious behavior is possible because we make a decision not to change the passwords of accounts we purchase.

Table 2 shows the fraction of purchased accounts per seller that merchants reaccessed and resold. A total of 10% of accounts in our dataset were logged into (either by the seller or a third party; it is not possible to distinguish the two) within a median of 3 days from our purchase. We find that 6% of all accounts go on to be resold in a median of 5 days from our purchase. This serves to highlight that some merchants are by no means shy about scamming potential customers.

4 Fraudulent Registration Analysis

Account merchants readily evade existing abuse safeguards to register thousands of accounts on a recurring basis. To understand these failings, we delve into the tools and techniques required to operate in the account marketplace. We find that merchants leverage thousands of compromised hosts, CAPTCHA solvers, and access to fraudulent email accounts. We identify what registration barriers increase the price of accounts and summarize our observations into a set of recommendations for how web services can improve existing automation barriers to increase the cost of fraudulent credentials in the future.

4.1 Dataset Summary

To carry out our analysis, we combine intelligence gathered from the underground market with private data provided through a collaboration with Twitter. Due to the sensitivity of this data, we strictly adhere to a data policy set down by Twitter, documented in Appendix A. In total, we have the credentials for 121,027 purchased accounts, each of which we annotate with the seller and source of solicitation. Furthermore, we obtain access to each account's associated email address; login history going back one year including IP addresses and timestamps; signup information including the IP and user agent used to register the account; the history of each account's activities including tweeting or the formation of social connections, if any; and finally whether Twitter has flagged the account as spam (independent of our analysis).

4.2 Circumventing IP Defenses

Unique IP addresses are a fundamental resource for registering accounts in bulk. Without a diverse IP pool, fraudulent accounts would fall easy prey to network-based blacklisting and throttling [12, 18, 35]. Our analysis leads us to believe that account merchants either own or rent access to thousands of compromised hosts to evade IP defenses.

IP Address Diversity & Geolocation As a whole, miscreants registered 79% of the accounts we purchase from unique IP addresses located across the globe. No single subnet captures the majority of abused IPs; the top ten $/24$ subnets account for only 3% of signup IPs, while the top ten $/16$ subnets account for only 8% of registrations. We provide a breakdown of geolocations tied to addresses under the control of merchants in Table 3. India is the most popular origin of registration, accounting for 8.5% of all fraudulent accounts in our dataset.

Registration Origin	Unique IPs	Popularity
India	6,029	8.50%
Ukraine	6,671	7.23%
Turkey	5,984	5.93%
Thailand	5,836	5.40%
Mexico	4,547	4.61%
Viet Nam	4,470	4.20%
Indonesia	4,014	4.10%
Pakistan	4,476	4.05%
Japan	3,185	3.73%
Belarus	3,901	3.72%
Other	46,850	48.52%

Table 3: Top 10 most popular geolocations of IP addresses used to register fraudulent accounts.

Other 'low-quality' IP addresses (e.g. inexpensive hosts from the perspective of the underground market [4]) follow in popularity. In summary, registrations come from 164 countries, the majority of which serve as the origin of fewer than 1% of accounts in our dataset. However, in aggregate, these small contributors account for 48.5% of all registered accounts.

Merchants that advertise on blackhat forums or operate their own web storefronts have the most resources at their disposal, registering all but 15% of their accounts via unique IPs from hundreds of countries. Conversely, merchants operating on Fiverr and Freelancer tend to operate solely out of the United States or India and reuse IPs for at least 30% of the accounts they register.

Long-term IP Abuse To understand the long-term abuse of IP addresses, we analyze data provided by Twitter that includes *all* registered accounts (not just our purchases) from June, 2012 – April, 2013. From this, we select a random sample of 100,000 unique IPs belonging to accounts that Twitter has disabled for spamming (e.g. suspended) and an equally sized sample of IPs used to register legitimate Twitter accounts. We add a third category to our sample that includes all the unique IP addresses used by merchants to register the accounts we purchased. For each of these IPs, we calculate the total number of Twitter accounts registered from the same IP.

A CDF of our results, shown in Figure 2, indicates merchants use the IP addresses under their control to register an abnormal number of accounts. Furthermore, the merchants we track are more cautious than other Twitter spammers who register a larger volume of accounts from a single IP address, making the merchants harder to detect. In total, merchants use 50% of the IP addresses under their control to register fewer than 10 accounts,

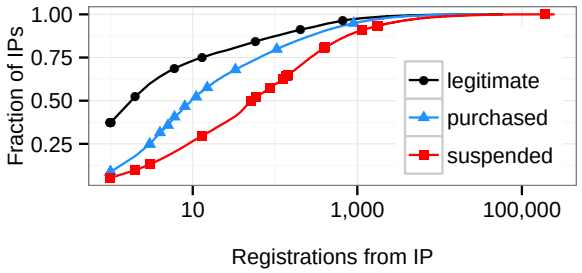

Figure 2: CDF of registrations per IP tied to purchased accounts, legitimate accounts, and suspended (spam) accounts.

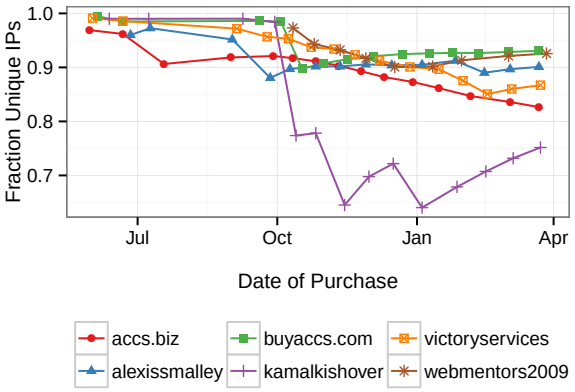

Figure 3: Availability of unique IPs over time for the six merchants we track over the longest period. All but one seller we repeatedly purchase from are able to acquire new IP addresses to register accounts from over time.

compared to 73% of IPs tied to legitimate users and only 26% for other spammers. We note that the small fraction of legitimate IP addresses used to register thousands of accounts likely belong to mobile providers or other middleboxes.

IP Churn & Pool Size In order to sustain demand for new accounts without overextending the abuse of a single IP address, merchants obtain access to tens of thousands of IP addresses that change over time. Figure 3 shows the fraction of accounts we purchase that appear from a unique IP address[4] as a function of time. We restrict our analysis to the six merchants we track over the longest period. Despite successive purchases of 1,000 accounts, all but one seller maintains IP uniqueness above roughly 80% of registered accounts, indicating that the IPs available to merchants change over time.

[4]We calculate uniqueness over the IP addresses in our dataset, not over all IPs used to register accounts on Twitter.

We calculate the number of IP addresses under each merchant's control by treating IP reuse as a *closed capture-recapture* problem. Closed capture-recapture measurements – used to estimate an unknown population size – require (1) the availability of independent samples and (2) that the population size under study remains fixed. To begin, we assume each purchase we make is an independent sample of the IP addresses under a merchant's control, satisfying the first requirement. The second requirement is more restrictive. If we assume that merchants use IP addresses tied to compromised hosts, then there is an inherent instability in the population size of IPs due to hosts becoming uninfected, new hosts becoming infected, and ISPs reallocating dynamic IPs. As such, comparisons over long periods are not possible. Nevertheless, if we restrict our analysis to batches of accounts from a single seller that were all registered within 24 hours, we can minimize the imprecision introduced by IP churn.

To this end, we select clusters of over 300 accounts registered by merchants within a 24 hour window. We split each cluster in half by time, with the first half m acting as the set of marked IPs and the second set c as the captured IPs, where there are r overlapping, or recaptured, IPs between both sets. We can then estimate the entire population size $\hat{N}$ (e.g. the number of unique IPs available to a merchant) according to the Chapman-Petersen method [24]:

$$\hat{N} = \frac{(m+1)(c+1)}{(r+1)} - 1$$

And standard error according to:

$$SE = \sqrt{\frac{\hat{N}^2(c-r)}{(c+1)(r+2)}}$$

For 95% confidence intervals, we calculate the error of $\hat{N}$ as $\pm 1.96 \times SE$. We detail our results in Table 4. We find that sellers like *accs.biz* and *victoryservices* have tens of thousands of IPs at their disposal on any given day, while even the smallest web storefront merchants have thousands of IPs on hand to avoid network-based blacklisting and throttling.

4.3 CAPTCHAs & Email Confirmation

Web services frequently inhibit automated account creation by requiring new users to solve a CAPTCHA or confirm an email address. Unsurprisingly, we find neither of these barriers are insurmountable, but they *do* impact the pricing and rate of generation of accounts, warranting their continued use.

Merchant	$\hat{N}$ Estimate	$\pm$ Error
accs.biz	21,798	4,783
victoryservices	17,029	2,264
dataentryassistant.com	16,887	4,508
alexissmalley	16,568	3,749
webmentors2009	10,019	2,052
buyaccs.com	9,770	3,344
buytwitteraccounts.info	6,082	1,661
buyaccountsnow.com	5,438	1,843
spamvilla.com	4,646	1,337
kamalkishover	4,416	1,170

Table 4: Top 10 merchants with the largest estimated pool of IP addresses under their control on a single day.

Email Confirmation All but 5 of the merchants we purchase from readily comply with requirements to confirm email addresses through the receipt of a secret token. In total, merchants email confirm 77% of accounts we acquire, all of which they seeded with a unique email. The failure of email confirmation as a barrier directly stems from pervasive account abuse tied to web mail providers. Table 5 details a list of the email services frequently tied to fraudulent Twitter accounts. Merchants abuse Hotmail addresses to confirm 60% of Twitter accounts, followed in popularity by Yahoo and mail.ru. This highlights the interconnected nature of account abuse, where credentials from one service can serve as keys to abusing yet another.

While the ability of merchants to verify email addresses may raise questions of the processes validity, we find that email confirmation positively impacts the price of accounts. Anecdotally, Hotmail and Yahoo accounts are available on *blackhatworld.com* for $6 per thousand, while Twitter accounts from the same forum are $40 per thousand. This is also true of web storefront such as *buyaccs.com* where mail.ru and Hotmail accounts are $5 per thousand, compared to $20 per thousand for Twitter accounts. Within our own dataset, we find that Twitter accounts purchased without email confirmation cost on average $30 per thousand compared to $47 per thousand for accounts with a confirmed email address. This difference likely includes the base cost of an email address and any related overhead due to the complexity of responding to a confirmation email.

CAPTCHA Solving Twitter throttles multiple registrations originating from a single IP address by requiring a CAPTCHA solution. Merchants solved a CAPTCHA for 35% of the accounts we purchase; the remaining accounts were registered from fresh IPs that did not trigger throttling. While there are a variety of CAPTCHA solving

Email Provider	Accounts	Popularity
hotmail.com	64,050	60.08%
yahoo.com	12,339	11.57%
mail.ru	12,189	11.43%
gmail.com	2,013	1.89%
nokiamail.com	996	0.93%
Other	2,157	0.14%

Table 5: Top 5 email providers used to confirm fraudulent Twitter accounts.

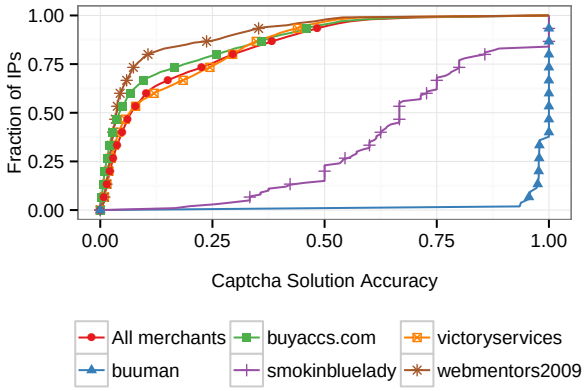

Figure 4: CAPTCHA solution rates per each IP address abused by a variety of merchants as well as the rates for all merchants combined.

services available in the underground market [19], none are free and thus requiring a CAPTCHA slightly increases the cost of creating fraudulent accounts.

A second aspect of CAPTCHAs is the success rate of automated or human solvers. By virtue of only buying successfully registered accounts, we cannot exactly measure CAPTCHA failure rates (unless account sellers fail and re-try a CAPTCHA during the same registration session, something we find rare in practice). However, we can examine registration attempts that occur from the same IPs as the accounts we purchase to estimate the rate of failure. To carry out this analysis, we examine all registrations within the previous year, calculating the fraction of registrations that fail due to incorrect CAPTCHA solutions per IP address.

We show a CDF of CAPTCHA solution rates for a sample of merchants in Figure 4. The median CAPTCHA solution rate for all sellers is 7%, well below estimates for automated CAPTCHA solving software of 18–30% [19], a discrepancy we currently have no explanation for. For two of the Fiverr sellers, *buuman* and *smokinbluelady*, the median CAPTCHA solution rate per IP is 100% and 67% respectively, which would indicate a human solver. In total, 92% of all throttled registration attempts from merchants fail. Despite this fact, account sellers are still able to register thousands accounts over the course of time, simply playing a game of odds.

4.4 Stockpiling & Suspension

Without effective defenses against fraudulent account registration, merchants are free to stockpile accounts and sell them at a whim. For many solicitations, merchants consider "pre-aged" accounts a selling point, not a detraction. To highlight this problem, we examine the failure of at-abuse time metrics for detecting dormant accounts and the resulting account stockpiles that occur.

Account Suspension Twitter suspends (e.g. disables) spam accounts due to at-abuse time metrics such as sending spam URLs or generating too many relationships, as outlined in Twitter's rules [30]. In our case, we are interested in whether fraudulent accounts that do *not* perform visible spam actions (e.g. are dormant) nevertheless become suspended. While for miscreants this should ideally be impossible, there are multiple avenues for guilt by association, such as clustering accounts based on registration IP addresses or other features. As such, when Twitter suspends a large volume of active fraudulent accounts for spamming, it is possible for Twitter to catch dormant accounts in the same net.

Of the dormant accounts we purchase, only 8% are eventually detected and suspended. We exclude accounts that were resold and used to send spam (outlined in Section 3.5) from this metric in order to not skew our results. Of the merchants we track, Fiverr sellers take the least caution in registering unlinkable accounts, resulting in 57% of our purchases becoming suspended by the time of our analysis. In contrast, web storefronts leverage the vast resources at their disposal to create unlinkable accounts, where only 5% of our purchased accounts are eventually detected as fraudulent. These poor detection rates highlight the limitation of at-abuse time metrics against automated account registration. Without more sophisticated at-registration abuse signals, merchants are free to create thousands of accounts with minimal risk of Twitter suspending back stock.

Account Aging & Stockpiling We examine the age of accounts, measured as the time between their registration and subsequent date of purchase, and find that accounts are commonly stockpiled for a median of 31 days. While most merchants deal exclusively in back stock, some merchants operate in an on-demand fashion. At the far end of this spectrum is a merchant *spamvilla.com* that sold us accounts registered a median of 323 days ago

– nearly a year in advance of our purchase. In contrast, webstores such as *buyaccs.com* and Fiverr merchants including *smokinbluelady* sell accounts less than a day old. Even though these merchants operate purely on-demand, they are still able to fulfill large requests in short order (within a day in our experience). Both modes of operation illustrate the ease that merchants circumvent existing defenses and the need for at-registration time abuse detection.

4.5 Recommendations

Web services that rely on automation barriers must strike a tenuous balance between promoting user growth and preventing the proliferation of fraudulent accounts and spam behavior. We summarize our findings in this section with a number of potential improvements to existing barriers that should not impede legitimate users. While we draw many of our observations from the Twitter account abuse problem, we believe our recommendations should generalize across web services.

Email Confirmation While account merchants have cheap, disposable emails on hand to perform email confirmation, confirmation helps to increase the cost of fraudulent accounts. In the case of Twitter, email confirmation raises the cost of accounts by 56%. Furthermore, in the absence of clear abuse signals, services can use email *reconfirmation* as a *soft action* against automation, similar to requiring a CAPTCHA before sending an email or tweet. Of the Twitter accounts we purchased, only 47% included the email address and password used to confirm the account. Merchants will sometimes re-appropriate these email addresses and sell them as "second-hand" at a discount of 20%. Without the original credentials, miscreants will be unable to perform email reconfirmation. Even if merchants adapt and begin to provide email credentials as part of their sale, the possibility of reselling email addresses disappears, cutting into a merchant's revenue.

CAPTCHAs CAPTCHAs serve to both increase the cost of accounts due to the requirement of a CAPTCHA solving service as well as to throttle the rate of account creation. In our experience, when required, CAPTCHAs prevent merchants from registering 92% of fraudulent accounts. Services could also leverage this failure rate as a signal for blacklisting an IP address in real-time, cutting into the number of accounts merchants can register from a single IP.

IP Blacklisting While miscreants have thousands of IP addresses at their disposal that rapidly change, IP blacklisting is not without merit. Our results show that merchants use a small fraction of IPs to register tens of thousands of accounts, which services could curb with real-time blacklisting. While public and commercial IP blacklists exist such as CBL [5], previous work has shown these generate too many false positives in the case of social spam [28], requiring service providers to generate and maintain their own blacklists.

Phone Verification While Twitter does not require phone verification, we observe the positive impact phone verification has on increasing the cost of fraudulent accounts for other services. Facebook and GMail accounts that are phone verified cost up to 150x more than their Twitter, non-PVA counterpart. As with CAPTCHAs or email reconfirmation, phone verification can serve as a soft action against spammers who do not clearly fall into the set of accounts that should be automatically disabled.

5 Detecting Fraudulent Registrations

To understand the impact account merchants have on Twitter spam, we develop a classifier trained on purchased accounts to *retroactively* identify abusive registrations. Our technique relies on identifying patterns in the naming conventions and registration process used by merchants to automatically generate accounts. We apply our classifier to *all* Twitter accounts registered in the last year (overlapping with our investigation) and identify several million accounts which appear to be fraudulent. We note this approach is *not* meant to sustain accuracy in an adversarial setting; we only apply it to historical registrations where adaptation to our signals is impossible.

5.1 Automatic Pattern Recognition

Our detection framework begins by leveraging the limited variability in naming patterns used by account generation algorithms which enables us to automatically construct regular expressions that fingerprint fraudulent accounts. Our approach for generating these expressions is similar to previous techniques for identifying spam emails based on URL patterns [32] or spam text templates [22, 23]. However, these previous approaches fail on small text corpuses (e.g. screennames), especially when samples cannot be linked by repeating substrings. For this reason, we develop a technique explicitly for account naming patterns. Algorithm 1 shows a sketch of our approach which we use to guide our discussion.

Common Character Classes To capture accounts that

Algorithm 1 Generate Merchant Pattern

```
Input: List of accounts for a single merchant
Parameters: τ (minimum cluster size)
clusters ← GROUP accounts BY
        (Σ-Seq, repeatedNames, emailDomain)
for all cluster ∈ clusters do
    if cluster.size() > τ then
        patterns ← MINMAXΣ-SEQ (cluster)
        OUTPUTREGEX(patterns, repeatedNames)
    end if
end for
```

all share the same naming structure, we begin by defining a set of character classes:

$$\mathbf{\Sigma} = \{p\{Lu\}, p\{Ll\}, p\{Lo\}, d, \dots\}$$

composed of disjoint sets of characters including uppercase Unicode letters, lowercase Unicode letters, noncased Unicode letters (e.g., Arabic). and digits.[5] We treat all other characters as distinct classes (e.g., +, -, _). We chose these character classes based on the naming patterns of accounts we purchase, a sample of which we show in Table 6. We must support Unicode as registration algorithms draw account names from English, Cyrillic, and Arabic.

From these classes we define a function Σ-*Seq* that captures transitions between character classes and produces an ordered set $\sigma_1\sigma_2\dots\sigma_n$ of arbitrary length, where σ_i represents the i-th character class in a string. For example, we interpret the account Wendy Hunt from *accs.biz* as a sequence $p\{Lu\}p\{Ll\}\ p\{Lu\}p\{Ll\}$. We repeat this process for the name, screenname, and email of each account. We note that for emails, we strip the email domain (e.g. @hotmail.com) prior to processing and use this as a separate feature in the process for pattern generation.

Repeated Substrings While repeated text stems between multiple accounts are uncommon due to randomly selected dictionary names, we find the algorithms used to generate accounts often reuse portions of text for names, screennames, and emails. For instance, all of the accounts in Table 6 from *victoryservices* have repeated substrings between an account's first name and screenname.

To codify these patterns, we define a function *repeatedNames* that canonicalizes text from an account's fields, brute forces a search of repeated substrings, and then codifies the resulting patterns as invariants. Canonicalization entails segmenting a string into multiple substrings based on Σ-Seq transitions. We preserve full names by ignoring transitions between upper and lowercase letters; spaces are also omitted from canonicalization. We then convert all substrings to their lowercase equivalent, when applicable. To illustrate this process, consider the screenname WendyHunt5. Canonicalization produces an ordered list [wendy,hunt,5], while the name Wendy Hunt is converted to [wendy,hunt].

The function repeatedNames proceeds by performing a brute force search for repeated substrings between all canonicalized fields of an account. For our previous example of WendyHunt5, one successful match exists between name[1] and screenname[1], where $[i]$ indicates the i-th position of a fields substring list; this same pattern also holds for the name and screenname for Kristina Levy. We use this positional search to construct invariants that hold across accounts from a single merchant. Without canonicalization, we could not specify what relationship exists between Wendy and Kristina due to differing text and lengths. When searching, we employ both exact pattern matching as well as partial matches (e.g. neff found in brindagtgneff for *buyaccs.com*). We use the search results to construct invariants for both strings that must repeat as well as strings that never repeat.

Clustering Similar Accounts Once we know the Σ-Seq, repeatedNames, and email domain of every account from a merchant, we cluster accounts into non-overlapping groups with identical patterns, as described in Algorithm 1. We do this on a per-merchant basis rather than for every merchant simultaneously to distinguish which merchant an account originates from. We prune small clusters based on a empirically determined τ to reduce false positives, with our current implementation dropping clusters with fewer than 10 associated accounts.

Bounding Character Lengths The final phase of our algorithm strengthens the invariants tied to Σ-Seq transitions by determining a minimum length $min(\sigma_i)$ and maximum length $max(\sigma_i)$ of each character class σ_i. We use these to define a bound $\{l_{min}, l_{max}\}$ that captures all accounts with the same Σ-Seq. Returning to our examples in Table 6, we group the account names from *accs.biz* and produce an expression $p\{Lu\}\{1,1\}p\{Ll\}\{5,8\}\ \{1,1\}p\{Lu\}\{1,1\}p\{Ll\}\{4,4\}$. We combine these patterns with the invariants produced by repeatedNames to construct a regular expression that fingerprints a cluster. We refer to these patterns for the rest of this paper as *merchant patterns*.

5.2 Pattern Refinement

We refine our merchant patterns by including abuse-orientated signals that detect automated signup behavior

[5] We use Java character class notation, where p{*} indicates a class of letters and Lu indicates uppercase, Ll lowercase, and Lo non-case.

Seller	Popularity	Name	Screenname	Email
victoryservices	57%	Trstram Aiken	Trstramsse912	KareyKay34251@hotmail.com
		Millicent Comolli	Millicentrpq645	DanHald46927@hotmail.com
accs.biz	46%	Wendy Hunt	WendyHunt5	imawzgaf7083@hotmail.com
		Kristina Levy	KristinaLevy6	exraytj8143@hotmail.com
formefor	43%	ola dingess	olawhdingess	TimeffTicnisha@hotmail.com
		brinda neff	brindagtgneff	ScujheShananan@hotmail.com
spamvilla.com	38%	Kiera Barbo	Kierayvydb	LinJose344@hotmail.com
		Jeannine Allegrini	Jeanninewoqzg	OpheliaStar461@hotmail.com

Table 6: Obfuscated sample of names, screennames, and emails of purchased accounts used to automatically generate merchant patterns. Popularity denotes the fraction of accounts that match the pattern for an individual merchant.

based on the registration process, user-agent data, and timing events.

Signup Flow Events We begin our refinement of merchant patterns by analyzing the activities of purchased accounts during and immediately after the signup work flow. These activities include events such as a user importing contacts and accessing a new user tutorial. The complete list of these events is sensitive information and is omitted from discussion. Many of these events go untriggered by the automated algorithms used by account sellers, allowing us to distinguish automated registrations from legitimate users.

Given a cluster of accounts belonging to a single merchant, we generate a binary feature vector $\mathbf{e}_{sig} = \{0,1\}^n$ of the n possible events triggered during signup. A value of 1 indicates that at least ρ accounts in the cluster triggered the event e. For our experiments, we specify a cutoff $\rho = 5\%$ based on reducing false positives. Subsequently, we determine whether a new account with event vector $\mathbf{e}$ matches a seller's signup flow signature $\mathbf{e}_{sig}$ by computing whether $\mathbf{e} \subseteq \mathbf{e}_{sig}$ holds. The majority of legitimate accounts have $|\mathbf{e}| \gg |\mathbf{e_{sig}}|$, so we reject the possibility they are automated even though their naming conventions may match a merchant's.

User Agents A second component of signups is the user agent associated with a form submission. Direct matching of user agents used by a seller with new subsequent signups is infeasible due to sellers randomizing user agents. For instance, *buytwitteraccounts.info* uses a unique (faked) agent for every account in our purchased dataset. Nevertheless, we can identify uniformity in the naming conventions of user agents just as we did with account names and screennames.

Given a cluster of accounts from a single seller, we generate a *prefix tree* containing every account's user agent. A node in the tree represents a single character from a user agent string while the node's depth mirrors the character's position in the user agent string. Each node also contains the fraction of agents that match the substring terminated at the given node. Rather than find the longest common substring between all accounts, we prune the tree so that every substring terminating at a node has a fraction of at least ϕ accounts in the cluster (in practice, 5%). We then generate the set of all substrings in the prefix tree and use them to match against the agents of newly registered accounts. The resulting substrings include pattens such as Mozilla/5.0 (X11; Linux i686 which, if not truncated, would include multiple spurious browser toolbars and plugins and be distinct from subsequent signups. While in theory the resulting user agent substrings can be broad, in practice we find they capture browser variants and operating systems before being truncated.

Form Submission Timing The final feature from the signup process we use measures the time between Twitter serving a signup form to the time the form is submitted. We then compute a bound $\{min_{ts}, max_{ts}\}$ for each seller to determine how quickly a seller's algorithm completes a form. To counter outliers, we opt for the 99% for both minimum and maximum time. For instance, the Fiverr merchant *kathlyn* registers accounts within $\{0,1\}$ seconds. A newly minted account can match a seller's algorithm if its form completion time is within the sellers bound.

5.3 Alternative Signals

There were a number of alternative signals we considered, but ultimately rejected as features for classification. We omitted the delay between an account's registration and subsequent activation as we lacked training data to measure this period; all our accounts remain dormant after purchase (minus the small fraction that were resold). We also analyzed both the timing of registra-

tions as well as the interarrival times between successive registrations. We found that merchants sell accounts in blocks that sometimes span months, preventing any interarrival analysis. Furthermore, merchants register accounts at uniformly random hours and minutes. Finally, as merchants create accounts from IP addresses around the globe, no subnet or country accurately captures a substantive portion of abusive registrations.

5.4 Evaluation

To demonstrate the efficacy of our model, we retroactively apply our classifier to *all* Twitter accounts registered in the last year. In total, we identify several million[6] distinct accounts that match one of our merchant patterns and thus are potentially fraudulent. We validate these findings by analyzing both the *precision* and *recall* of our model as well measuring the impact of time on the model's overall accuracy.

Precision & Recall Precision measures the fraction of identified accounts that are in fact fraudulent (e.g., not misclassified, legitimate users), while recall measures the fraction of all possible fraudulent accounts that we identify, limited to the merchants that we study. To estimate the precision of each merchant pattern, we select a random sample of 200 accounts matching each of 26 merchant patterns,[7] for a total of 4,800 samples. We then manually analyze the login history, geographic distribution of IPs, activities, and registration process tied to each of these accounts and label them as spam or benign. From this process, we estimate our overall precision at 99.99%, with the breakdown of the most popular merchant pattern precisions shown in Table 7. In a similar vein, we estimate recall by calculating the fraction of all accounts we purchase that match our classifier. In total, we correctly identify 95% of all purchased accounts; the remaining 5% of missed accounts did not form large enough clusters to be included in a merchant's pattern, and as a result, we incorrectly classified them as legitimate.

Performance Over Time The performance of our model is directly tied to accurately tracking adaptations in the algorithms used by merchants to register accounts. To understand how frequently these adaptations occur, we evaluate the performance of our classifier as a function of time. Figure 5 shows the overall recall of each of our merchant patterns for the sellers we track over the longest period of time. For each merchant, we train a classifier on accounts acquired up to time t and evaluate it on all accounts from the merchant, regardless of when we purchased the account. We find that some sellers such as *alexissmalley* rarely alter their registration algorithm throughout our study, allowing only two purchases to suffice for accurate detection. In contrast, we see a shift in registration algorithms for a number of merchants around October and January, but otherwise patterns remain stable for long periods. The several million accounts we identify as fraudulent should thus be viewed as a lower bound in the event we missed an adaptation.

Service	Rel. Volume	P	R
buuman	0.00%	100.00%	70.67%
smokinbluelady	0.08%	100.00%	98.91%
danyelgallu	0.12%	100.00%	100.00%
twitter007	0.13%	100.00%	97.33%
kathlyn	0.13%	100.00%	93.24%
jackhack08	0.41%	100.00%	100.00%
twitfollowers	0.72%	100.00%	92.50%
denial93	2.18%	100.00%	100.00%
ghetumarian	3.05%	100.00%	85.94%
formefor	4.75%	100.00%	100.00%
shivnagsudhakar	–	–	–
kamalkishover	29.90%	99.60%	92.73%
naveedakhtar	0.24%	100.00%	98.40%
webmentors2009	0.85%	100.00%	99.64%
truepals	1.02%	100.00%	93.08%
alexissmalley	1.68%	100.00%	98.62%
victoryservices	6.33%	99.70%	99.03%
spamvilla.com	0.71%	99.00%	98.70%
getbulkaccounts.com	2.97%	100.00%	100.00%
xlinternetmarketing.com	3.12%	100.00%	95.13%
accs.biz	4.48%	100.00%	97.62%
buytwitteraccounts.biz	6.10%	100.00%	84.27%
quickaccounts.bigcartel	10.91%	100.00%	99.73%
buytwitteraccounts.info	20.45%	99.60%	81.85%
dataentryassistant.com	24.01%	100.00%	96.57%
buyaccountsnow.com	30.75%	99.10%	95.10%
buyaccs.com	58.39%	100.00%	91.66%
Total	100.00%	99.99%	95.08%

Table 7: Breakdown of the merchants, the relative volume of all detected accounts in the last year that match their pattern, precision (P) and recall (R).

Pattern Overlap & Resale The simultaneous adaptation of merchant patterns in Figure 5 around October and other periods leads us to believe that a multitude of merchants are using the same software to register accounts and that an update was distributed. Alternatively, the

[6] Due to operational concerns, we are unable to provide exact numbers on the volume of spam accounts registered. As such, we reference merchants and the impact they have on Twitter as a *relative volume* of all several million accounts that we detect.

[7] We omit accounts purchased from the Freelancer merchant *shivnagsudhakar* as these were registered over a year ago and thus lay outside the range of data to which we had access.

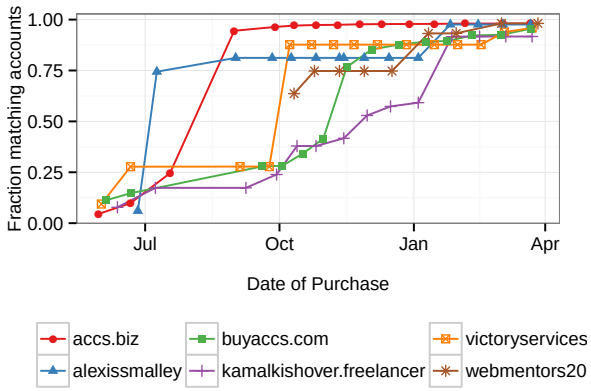

Figure 5: Recall of generated merchant patterns for all purchased accounts as a function of training the classifier on data only prior to time t.

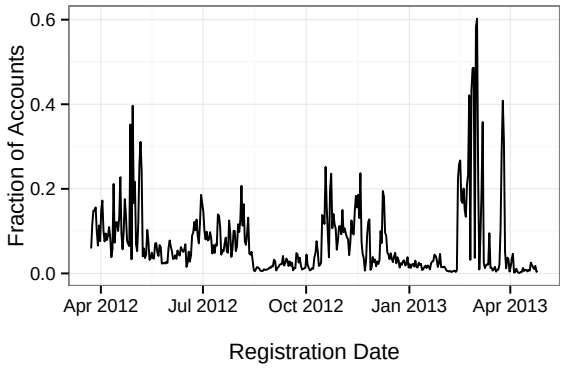

Figure 6: Fraction of all suspended accounts over time that originate from the underground market.

account marketplace may have multiple levels of resale (or even arbitrage) where accounts from one merchant are resold by another for an increased cost, leading to correlated adaptations. Further evidence of correlated patterns appears in the merchant patterns we construct, where a classifier for one merchant will accurately detect accounts sold to us by a second merchant. For instance, the accounts sold by *kamalkishover* from Freelancer overlap with the patterns of 9 other merchants, the most popular of which is *buyaccountsnow.com*. We find most Fiverr sellers are independent with the exception of *denial93*, *ghetumarian*, and *formefor*, whose patterns overlap with the major account web storefronts. This would explain why these three Fiverr sellers appear to be much larger (from the perspective of Table 7) compared to other Fiverr merchants. As a result, our estimates for the number of accounts registered by each merchant may be inflated, though our final total counts only unique matches and is thus globally accurate.

6 Impact of the Underground Market

We analyze the several million accounts we flag as registered by merchants operating in the underground market and estimate the fraction that have been sold and used to generate Twitter spam. We find that, during active months, the underground market was responsible for registering 10–20% of all accounts that Twitter later flagged as spam. For their efforts, we estimate that merchants generated a combined revenue between $127,000– $459,000.

6.1 Impact on Twitter Spam

From our seed set of 121,027 accounts purchased from 27 merchants, we are able to identify several million fraudulent accounts that were registered by the same merchants. Of these, 73% were sold and actively tweeting or forming relationships at one point in time, while the remaining 37% remained dormant and were yet to be purchased.

In cooperation with Twitter, we analyzed the total fraction of all suspended accounts that appear to originate from the merchants we track, shown in Figure 6. At its peak, the underground marketplace was responsible for registering 60% of all accounts that would go on to be suspended for spamming. During more typical periods of activity, the merchants we track contribute 10–20% of all spam accounts. We note that the drop-off around April does not indicate a lack of recent activity; rather, as accounts are stockpiled for months at a time, they have yet to be released into the hands of spammers, which would lead to their suspension. The most damaging merchants from our impact analysis operate out of blackhat forums and web storefronts, while Fiverr and Freelancer sellers generate orders of magnitude fewer accounts.[8]

6.2 Estimating Revenue

We estimate the revenue generated by the underground market based on the total accounts sold and the prices charged during their sale. We distinguish accounts that have been sold from those that lay dormant and await sale based on whether an account has sent tweets or formed relationships. For sold accounts, we identify which mer-

[8]The exception to this is a Freelancer merchant kamalkishover, but based on their merchant pattern overlapping with 9 other merchants, we believe they are simply reselling accounts.

chant created the account and determine the minimum and maximum price the merchant would have charged for that account based on our historical pricing data.[9] In the event multiple merchants could have generated the account (due to overlapping registration patterns), we simply take the minimum and maximum price of the set of matching merchants.

We estimate that the total revenue generated by the underground account market through the sale of Twitter credentials is between the range of $127,000– $459,000 over the course of a year. We note that many of the merchants we track simultaneously sell accounts for a variety of web services, so this value likely represents only a fraction of their overall revenue. Nevertheless, our estimated income is far less than the revenue generated from actually sending spam [17] or selling fake antivirus [25], where revenue is estimated in the tens of millions. As such, account merchants are merely stepping stones for larger criminal enterprises, which in turn disseminate scams, phishing, and malware throughout Twitter.

7 Disrupting the Underground Market

With Twitter's cooperation, we disable 95% of all fraudulent accounts registered by the 27 merchants we track, including those previously sold but not yet suspended for spamming. Throughout this process, we simultaneously monitor the underground market to track fallout and recovery. While we do not observe an appreciable increase in pricing or delay in merchant's delivering new accounts, we find 90% of all purchased accounts immediately after our actioning are suspended on arrival. While we successfully deplete merchant stockpiles containing fraudulent accounts, we find that within two weeks merchants were able to create fresh accounts and resume selling working credentials.

7.1 Suspending Identified Accounts

In order to disrupt the abusive activities of account merchants, we worked with Twitter's Anti-spam, SpamOps, and Trust and Safety teams to manually validate the accuracy of our classifier and tune parameters to set an acceptable bounds on false positives (legitimate users incorrectly identified as fraudulent accounts). Once tuned, we applied the classifier outlined in Section 5 to every account registered on Twitter going back to March, 2012, filtering out accounts that were already suspended for abusive behavior.

From the set of accounts we identified[10], Twitter iteratively suspended accounts in batches of ten thousand and a hundred thousand before finally suspending all the remaining identified accounts. At each step we monitored the rate of users that requested their accounts be unsuspended as a metric for false positives, where unsuspension requests require a valid CAPTCHA solution. Of the accounts we suspended, only 0.08% requested to be unsuspended. However, 93% of these requests were performed by fraudulent accounts abusing the unsuspend process, as determined by manual analysis performed by Twitter. Filtering these requests out, we estimate the final precision of our classifier to be 99.9942%. The tuned classifier has a recall of 95%, the evaluation of which is identical to the method presented in Section 5. Assuming our purchases are a random sample of the accounts controlled by the underground market, we estimate that 95% of all fraudulent accounts registered by the 27 merchants we track were disabled by our actioning.

7.2 Marketplace Fallout and Recovery

Immediately after Twitter suspended the last of the underground market's accounts, we placed 16 new orders for accounts from the 10 merchants we suspected of controlling the largest stockpiles. Of 14,067 accounts we purchased, 90% were suspended on arrival due to Twitter's previous intervention. When we requested working replacements, one merchant responded with:

> All of the stock got suspended ... Not just mine .. It happened with all of the sellers .. Don't know what twitter has done ...

Similarly, immediately after suspension, *buyaccs.com* put up a notice on their website stating "Временно не продаем аккаунты Twitter.com", translating via Google roughly to "Temporarily not selling Twitter.com accounts".

While Twitter's initial intervention was a success, the market has begun to recover. Of 6,879 accounts we purchased two weeks after Twitter's intervention, only 54% were suspended on arrival. As such, long term disruption of the account marketplace requires both increasing the cost of account registration (as outlined in Section 4) and integrating at-signup time abuse classification into the account registration process (similar to the classifier

[9] Determining the exact time of sale for an account is not possible due to the potential of miscreants stockpiling their purchases; as such, we calculate revenue for both the minimum and maximum possible price.

[10] Due to operational concerns, we cannot specify the exact volume of accounts we detect that were not previously suspended by Twitter's existing defenses.

outlined in Section 5). We are now working with Twitter to integrate our findings and existing classifier into their abuse detection infrastructure.

8 Summary

We have presented a longitudinal investigation of the underground market tied to fraudulent Twitter credentials, monitoring pricing, availability, and fraud perpetrated by 27 account merchants. These merchants specialize in circumventing automated registration barriers by leveraging thousands of compromised hosts, CAPTCHA solvers, and access to fraudulent Hotmail, Yahoo, and mail.ru credentials. We identified which registration barriers positively influenced the price of accounts and distilled our observations into a set of recommendations for how web services can improve existing barriers to bulk signups. Furthermore, we developed a classifier based on at-registration abuse patterns to successfully detect several million fraudulent accounts generated by the underground market. During active months, the 27 merchants we monitor appeared responsible for registering 10–20% of all accounts later flagged by Twitter as spam. For their efforts, these merchants generated an estimated revenue between $127,000–$459,000. With Twitter's help, we successfully suspended 95% of all accounts registered by the 27 merchants we track, depleting the account stockpiles of numerous criminals. We are now working with Twitter to integrate our findings and existing classifier into their abuse detection infrastructure.

Acknowledgments

This work was supported by the National Science Foundation under grants 1237076 and 1237265, by the Office of Naval Research under MURI grant N000140911081, and by a gift from Microsoft Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Alexa. Alexa top 500 global sites. http://www.alexa.com/topsites, 2012.

[2] D.S. Anderson, C. Fleizach, S. Savage, and G.M. Voelker. Spamscatter: Characterizing internet scam hosting infrastructure. In *USENIX Security*, 2007.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, 2010.

[4] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring pay-per-install: The commoditization of malware distribution. In *USENIX Security Symposium*, 2011.

[5] CBL. Composite Blocking List. `http://cbl.abuseat.org/`, 2012.

[6] G. Danezis and P. Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2009.

[7] J. Franklin, V. Paxson, A. Perrig, and S. Savage. An inquiry into the nature and causes of the wealth of Internet miscreants. In *Proceedings of ACM Conference on Computer and Communications Security*, pages 375–388, October 2007.

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

[9] C. Grier, L. Ballard, J. Caballero, N. Chachra, C.J. Dietrich, K. Levchenko, P. Mavrommatis, D. McCoy, A. Nappa, A. Pitsillidis, et al. Manufacturing compromise: The emergence of exploit-as-a-service. 2012.

[10] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2010.

[11] T. Holz, C. Gorecki, F. Freiling, and K. Rieck. Detection and mitigation of fast-flux service networks. In *Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS)*, 2008.

[12] C.Y. Hong, F. Yu, and Y. Xie. Populated ip addresses—classification and applications. 2012.

[13] Heather Kelley. 83 million facebook accounts are fakes and dupes. `http://www.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/index.html`, 2012.

[14] Brian Krebs. Spam volumes: Past & present, global & local. `http://krebsonsecurity.com/2013/01/spam-volumes-past-present-global-local/`, 2012.

[15] S. Lee and J. Kim. Warningbird: Detecting Suspicious URLs in Twitter Stream. In *Symposium on Network and Distributed System Security (NDSS)*, 2012.

[16] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G.M. Voelker, and S. Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the 32nd IEEE Symposium on Security and Privacy*, 2011.

[17] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G.M. Voelker, S. Savage, and K. Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the 21st USENIX conference on Security symposium*. USENIX Association, 2012.

[18] A. Metwally and M. Paduano. Estimating the number of users behind ip addresses for combating abusive traffic. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

[19] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G.M. Voelker, and S. Savage. Re: Captchas–understanding captcha-solving services in an economic context. In *USENIX Security Symposium*, volume 10, 2010.

[20] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G.M. Voelker. An analysis of underground forums. In *Proceedings of the Internet Measurement Conference (IMC)*. ACM, 2011.

[21] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G.M. Voelker. Dirty jobs: The role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX Security Symposium*, 2011.

[22] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G.M. Voelker, V. Paxson, N. Weaver, and S. Savage. Botnet Judo: Fighting spam with itself. 2010.

[23] P. Prasse, C. Sawade, N. Landwehr, and T. Scheffer. Learning to identify regular expressions that describe email campaigns. 2012.

[24] W.E. Ricker. *Computation and interpretation of biological statistics of fish populations*, volume 191. Department of the Environment, Fisheries and Marine Service Ottawa, 1975.

[25] B. Stone-Gross, R. Abman, R. Kemmerer, C. Kruegel, D. Steigerwald, and G. Vigna. The Underground Economy of Fake Antivirus Software. In *Proceedings of the Workshop on Economics of Information Security (WEIS)*, 2011.

[26] G. Stringhini, C. Kruegel, and G. Vigna. Detecting Spammers on Social Networks. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2010.

[27] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and Evaluation of a Real-time URL Spam Filtering Service. In *Proceedings of the 32nd IEEE Symposium on Security and Privacy*, 2011.

[28] K. Thomas, C. Grier, and V. Paxson. Adapting social spam infrastructure for political censorship. In *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats*. USENIX Association, 2012.

[29] K. Thomas, C. Grier, V. Paxson, and D. Song. Suspended Accounts In Retrospect: An Analysis of Twitter Spam. In *Proceedings of the Internet Measurement Conference*, November 2011.

[30] Twitter. The Twitter Rules. `http://support.twitter.com/entries/18311-the-twitter-rules`, 2010.

[31] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B.Y. Zhao. Serf and Turf: Crowdturfing for Fun and Profit. In *Proceedings of the International World Wide Web Conference*, 2011.

[32] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: Signatures and characteristics. *Proceedings of ACM SIGCOMM*, 2008.

[33] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing Spammers' Social Networks for Fun and Profit: a Case Study of Cyber Criminal Ecosystem on Twitter. In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[34] H. Yu, M. Kaminsky, P.B. Gibbons, and A. Flaxman. Sybilguard: defending against sybil attacks via social networks. *ACM SIGCOMM Computer Communication Review*, 2006.

[35] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum. Botgraph: Large scale spamming botnet detection. 2009.

A Legal and Ethical Guidelines

To minimize the risk posed to Twitter or its users by our investigation of the account market, we follow a set of policies set down by our institutions and Twitter, reproduced here to serve as a note of caution to other researchers conducting similar research.

Twitter & Users Some of the account merchants we deal with work in an on-demand fashion, where purchases we place directly result in abusive registrations on Twitter (e.g. harm) in violation of the site's Terms of Services. Even purchases from existing stockpiles might be misconstrued as galvanizing further abuse of Twitter. As such, we directly contacted Twitter to receive permission to conduct our study. In the process, we determined that any interactions with the underground market should not result in harm to Twitter's user base. In particular, accounts we purchased should *never* be used to tweet or form relationships while under our control. Furthermore, we take no special action to guarantee our accounts are not *suspended* (e.g disabled) by Twitter; our goal is to observe the natural registration process, not to interact with or impede Twitter's service in any way.

Account Merchants We do not interact with merchants anymore than necessary to perform transactions. To this end, we only purchased from merchants that advertise their goods publicly and never contact merchants outside the web sites or forums they provide to conduct a sale (or to request replacement accounts in the event of a bad batch). Our goal is not to study the merchants themselves or to collect personal information on them; only to analyze the algorithms they use to generate accounts.

Sensitive User Data Personal data logged by Twitter is subject to a multitude of controls, while user names and passwords sold by merchants also carry controls to prevent fraud, abuse, and unauthorized access. First, we *never* log into accounts; instead, we rely on Twitter to verify the authenticity of credentials we purchase. Furthermore, all personal data such as IP addresses or activities tied to an account are never accessed outside of Twitter's infrastructure, requiring researchers involved in this study to work on site at Twitter and to follow all relevant Twitter security practices. This also serves to remove any risk in the event an account is *compromised* rather than registered by an account merchant, as no personal data ever leaves Twitter. To our knowledge, we never obtained credentials for compromised accounts.