The Abuse Sharing Economy: Understanding the Limits of Threat Exchanges

Kurt Thomas[1], Rony Amira[1], Adi Ben-Yoash[1], Ori Folger[1], Amir Hardon[1], Ari Berger[1], Elie Bursztein[1], and Michael Bailey[2]

[1] Google, Inc.
[2] University of Illinois, Urbana-Champaign

Abstract. The underground commoditization of compromised hosts suggests a tacit capability where miscreants leverage the same machine—subscribed by multiple criminal ventures—to simultaneously profit from spam, fake account registration, malicious hosting, and other forms of automated abuse. To expedite the detection of these commonly abusive hosts, there are now multiple industry-wide efforts that aggregate abuse reports into centralized *threat exchanges*. In this work, we investigate the potential benefit of global reputation tracking and the pitfalls therein. We develop our findings from a snapshot of 45 million IP addresses abusing six Google services including Gmail, YouTube, and ReCaptcha between April 7–April 21, 2015. We estimate the scale of end hosts controlled by attackers, expose underground biases that skew the abuse perspectives of individual web services, and examine the frequency that criminals re-use the same infrastructure to attack multiple, heterogeneous services. Our results indicate that an average Google service can block 14% of abusive traffic based on threats aggregated from seemingly unrelated services, though we demonstrate that outright blacklisting incurs an untenable volume of false positives.

Keywords: Threat exchanges, reputation systems, underground specialization

1 Introduction

The underground commoditization of compromised hosts enables miscreants to purchase, rent, or repurpose a glut of machinery in order to relay abusive traffic [1,32]. This suggests a tacit capability where miscreants leverage the same machine—subscribed by multiple criminal ventures—to simultaneously profit from spam, denial of service, malicious hosting, and other forms of automated abuse. Evidence to this effect includes the Torpig botnet which acted as an information stealer, SOCKS proxy, and HTTP proxy [30]; and the ZeroAccess botnet involved in search hijacking, automated click-fraud, and bitcoin mining [21]. More broadly, the *pay-per-install* business model enables miscreants to pay $10–180 for a thousand installs of an arbitrary payload [4]. Prolific botnets such as ZeroAccess, Mariposa, and Torpig provided similar install capabilities [20, 28, 30]. As a consequence, a single infected client may host multiple malware families, such as 6–15% of Confickr infections overlapping with Gameover Zeus [2], or 7–10% of search bots overlapping with spamming hosts [35].

In the absence of coordinated action among affected Internet services, each target must redundantly detect and filter commonly abusive hosts. While long-standing domain and IP blacklists have proven effective at bridging the information divide between email providers [12, 26, 36], there is no similarly mature system for globally tracking reputation across heterogeneous services such as cloud providers, mail servers, and social networks. To address this gap, an industry-wide effort has emerged in recent years to collate intelligence on active attacks and abusive clients into centralized *threat exchanges* [9, 19, 25, 27]. Under the mantra of "stronger together," these many-to-many exchanges have attracted participants across a spectrum of web services including Facebook, Bitly, Dropbox, Twitter, and Yahoo [9]. Despite momentum within industry, an important question remains for whether global threat intelligence will significantly improve current standalone anti-abuse pipelines, and if so, how best to reconcile, prioritize, and act upon warnings generated by algorithms and users rather than curated honeypots.

In this work, we design a threat exchange called Babel to explore the challenges and pitfalls inherent to any centralized reputation tracking of Internet devices. In particular, we measure: (1) the scale and network composition of infrastructure controlled by attackers; (2) the impact of network churn and evasion on long-term threat tracking; (3) the ratio of benign and abusive traffic originating from end hosts; and (4) ultimately whether commoditization has created a common substrate of abusive hosts that underpin multiple profit vectors. The answers to these questions serve to inform the nascent design of industry-lead threat exchanges and to illuminate any value in threat sharing between companies and government institutions.

To start, we develop an experimental threat exchange that collates hundreds of millions of real abuse incidents as reported by any of six federated Google services contending with spam, bulk account creation, fake engagement, and malware distribution over a 14 day period between April 7–April 21, 2015.[3] Each service relies on a specialized definition of abuse where incidents target semantically distinct entities (e.g., messages, accounts, domains) that are not immediately reconcilable—an inherent challenge for all threat exchanges. We decompose these application-specific, through context-rich abuse reports into a single repository of 45 million abusive IP addresses that serve as the launching point of attacks. We annotate each address with the volume of abuse per network, the services affected, and the duration of attacks.

We find that miscreants operate a vast apparatus of 8 million daily abusive hosts. Despite the sheer scale of infected devices in the wild, we find that the distribution of attacks across the Google services in our study is heavily skewed towards a small number of devices. The top 1% of abusive IP addresses generate 48–82% of abusive traffic per service, with email spam representing the most concentrated extreme. While this Zipf-like distribution holds for all abuse verticals, we also find evidence of regional specialization that biases the abuse perspective of individual Google services. In particular, we find the United States serves the majority of malware and drive-bys, Russian networks focus on fake YouTube engagement, and Indian networks create the most fake accounts. These non-uniform perspectives of abusive networks reduce the effectiveness of centralized threat exchanges, but as we will show does not render sharing inoperative.

[3] We opt for these reports over existing threat exchange data because the nascent (and invite-only) state of industry threat exchanges precludes a representative dataset for study.

Before actioning any network-based abuse intelligence, threat exchange consumers must contend with the possibility of coarse or stale abuse signals. To this end, we introduce a set of techniques to detect IP address re-assignment and quantify overlap between legitimate and abusive traffic on the same network. We find that a single device will cycle through an average of twenty IP addresses in two weeks. Translated into an abuse context, 66% of abusive IP addresses remain active for a single day before the associated device acquires a new IP address due to DHCP churn. As such, we find that while abuse lasts long enough to justify reporting, threat exchanges must enforce explicit time frames after which stale IP reputation expires.

Ultimately, by accounting for IP dynamism and taking advantage of skew where a small number of hosts are responsible for the vast majority of abuse, we find that at most 13% of Gmail spam and 43% of fake accounts can be caught due to simultaneous attacks on other products. These results illustrate that underground commoditization has yet to manifest into the purported ideal of miscreants maximizing the value of an infected host by engaging in all possible profit-generating activities. Nevertheless, we argue there is a value to threat exchanges that unify the abuse perspectives of heterogeneous web services. However, acting on this intelligence remains a challenge: outright blacklisting results in an unacceptable level of collateral damage as 62% of abusive IP addresses also relay legitimate content due to either NATing or simultaneous use by the device's owner. We discuss potential alternatives, such as incorporating centralized reputation signals into application-specific classifiers.

In summary, we highlight some of our key findings:

- **Exchanges must track millions of incidents:** We estimate miscreants control over 8 million daily IP addresses from a perspective of just six Google services.
- **Exchanges benefit even unrelated services:** Miscreants re-use underground infrastructure across abuse verticals. This allows an average service to catch 14% of abuse even when comparing spamming to fake account creation.
- **Exchanges must prioritize threats:** An inherent skew in miscreant strategies results in 1% of abusive devices generating 48–82% of attacks across services.
- **Exchanges must contend with transient abuse:** We find 66% of abusive IP addresses impact services for only a single day. Relying on stale incident reports results in unacceptable false positives.

2 Threat Exchanges: Design & Challenges

We begin by outlining current industry proposals for threat exchanges and potential challenges inherent to their design, the impact of which we evaluate throughout our work.

2.1 Existing Threat Exchanges

Threat exchanges are a community-driven, many-to-many broadcast platform for sharing abuse reports. This contrast with traditional domain and IP blacklists like Spamhaus

or Safe Browsing that rely on a curated one-to-many model for reporting abuse. Examples include Microsoft's Interflow [25], Facebook's ThreatExchange [9], IBM's X-Force Exchange [27], and Alien Vault's Open Threat Exchange [19]—all launched between 2012–2015. More historical examples also exist such as DShield, dating back to 2001, which serves as a bulletin of network intrusion incidents [7]. In practice, exchanges serve as a platform for alerting other participants to malicious IP addresses, URLs, binaries, extensions, email addresses, and even phone numbers: the support infrastructure underpinning digital fraud and abuse. Early adopters include Twitter, Pinterest, Tumblr, Dropbox, Bitly, and Yahoo [9] while IBM reports over 1,000 business participants [27].

2.2 Challenges

While threat exchanges are invaluable in theory for improving anti-abuse pipelines and providing access to training data, it remains up to participants to sift through the data deluge to identify credible threats. We summarize the potential challenges that arise from community-driven reporting.

Translating threats: The foremost challenge for threat exchange members is translating intelligence across abuse verticals. Email operates on messages, social networks on accounts and posts, URL shorteners on pages, and hosting providers on domains. Conveying threats between such web services requires decomposing abuse reports into universally recognizable subcomponents, potentially at the loss of rich contextual details such as collusion among accounts or domains all hosting the same spam template.

Competing policies: Participants in threat exchanges each have competing definitions of abuse (e.g., Terms of Service). For example, one social network may flag a host for aggressive account creation due to registering five accounts in a short window, while a second network might consider that typical behavior for a mobile endpoint. Similarly, a search engine may de-list URLs flagged for blackhat SEO while a URL shortener's abuse policy may restrict penalties to drive-by and phishing domains. Due to the arbitrary nature of many policies, threat exchange participants must learn which other members most closely match their rule sets.

Implicit bias: Abuse detection pipelines introduce an unmeasured bias due to the technology deployed, incomplete training data, and potentially skewed threats. Consequently, every reported (or unreported) entity carries an implicit false positive and negative rate that is unknown to all other participants absent longitudinal monitoring. While honeypots are highly curated to minimize false positives, any algorithm or user can report abuse to a threat exchange.

Stale identifiers: Abuse indicators such as IP addresses and domains, unlike file hashes, suffer from an innate instability introduced by network management (e.g., DHCP churn), takedown, and compromise remediation after which a host should no longer be treated as malicious. Reported entities are potentially credible only for a short time window before they become stale.

Coarse identifiers: Abuse indicators such as IP addresses and domains represent coarse identifiers for abusive hosts or pages. In particular, NATs, proxies, and middle-

boxes serve multiple simultaneous clients. Similarly, free hosting providers and URL shorteners serve content from a variety of owners all from the same domain. Blacklisting coarse identifiers inadvertently penalizes legitimate clients.

3 Building a Threat Exchange

With threat exchanges only recently launching, there is no agreed upon best practice for reconciling abuse incidents across web services. We present our approach for distilling application-specific abuse intelligence into a universal format. We apply this technique to hundreds of millions of abuse records collected by Google over a two week period ending on April 21, 2015. We note our limited collection window arises due to privacy restrictions.

3.1 Collating Abuse Reports

The greatest common divisor among services combating spam, malicious hosting, and account-related abuse is the IP address (and thus device) perpetrating the attack. Given a raw feed of labeled abusive and legitimate traffic belonging to a single web service, we aggregate threat intelligence into a tuple $\langle$*service, IP, date, volume, badness*$\rangle$ that contains a service identifier (e.g., Gmail), IP address, the date of abuse (restricted to 24 hour granularity), the volume of traffic originating from the address over the 24 hour period (e.g., email received), and the ratio of the traffic the service flagged as abusive. This approach allows us to identify which services are most impacted by mixed legitimate and abusive traffic and whether attacks persist for long periods. While other approaches exist, such as restricting analysis to domains or file hashes, we opt for IP addresses in our study because they are more expensive for miscreants to acquire and also universally applicable. We discuss limitations with this approach later in this section.

3.2 Abusive Traffic Dataset

In order to conduct our study, we rely on an abuse dataset that consists of 45 million IPv4 addresses reported by any of six Google services combating fraud and abuse. We detail each source of reports and highlight false positive and negative rates of respective feeds when previously published. A detailed breakdown of feeds is available in Table 1. As mentioned in Section 2, each of these datasets carries an implicit bias due to unknown accuracy, skewed threats, and orders of magnitude more traffic that are fundamental to threat exchanges.

Gmail Inbound Email [Spam]: Our Gmail dataset consists of SMTP relay IP addresses that sent an inbound spam email to a Gmail user, including messages blocked at the delivery layer and via content-based classification. Previous studies estimated this accuracy at above 99% [3]. We annotate each IP address with the total number of emails sent and the fraction Gmail blocked as spam.

Bulk Account Registration [Fake Accounts]: Our bulk account dataset includes IP addresses tied to automated account registration attempts blocked at creation or retroactively disabled upon detection where our timestamp reflects the original creation time

Table 1: Summary of abusive IPv4 addresses and the service targeted. We use the abbreviated names of services for all figures throughout our study.

Abbr	Reporting Service	IP Addresses	ASNs
–	Total Abusive IPs	45,171,301	40,069
GML	Gmail Spam	19,818,529	31,088
CAP	ReCaptcha Failures	14,892,992	34,018
YTB	YouTube Engagement	5,910,688	19,007
CMT	Comment Spam	4,233,722	21,607
SIG	Bulk Account Creation	1,616,067	15,627
SAF	Safe Browsing	49,117	3,348

of the account. We annotate each IP address with the number of registration attempts and the fraction Google identified as fraudulent.

YouTube Likes, Subscribes [Fake Engagement]: Our YouTube dataset contains IP addresses belonging to Google accounts polluting videos with fake "likes" and "subscribes", a form of signed-in abuse. We annotate each IP address with the total volume of likes and subscribes and the fraction YouTube flagged as abusive.

Comments [Spam]: Our comment dataset contains IP addresses tied to Google accounts that post spam comments to Blogger, YouTube, and Google+. We annotate each IP address with the total number of comments posted via the address and the fraction Google blocked as spam.

ReCaptcha [Automation]: Our ReCaptcha dataset consists of IP addresses that fail to correctly solve a CAPTCHA challenge. Unlike our other datasets where blocked activity is a concrete abuse verdict, we caution that CAPTCHA failure is a soft measure of abuse. We annotate each IP address with the number of CAPTCHA attempts and the fraction failed. We set a minimum failure threshold of 50%; we exclude IP addresses below this threshold from our dataset. We make no initial assumptions that unfiltered IP addresses are in fact abusive; instead, we rely on comparing CAPTCHA abuse to other verticals to draw conclusions.

Safe Browsing [Malicious Hosting]: Our final dataset consists of web server IP addresses reported by Safe Browsing for hosting malware and drive-by exploits [23]. We annotate each IP address with the number of web pages hosted on the IP and the fraction Safe Browsing flagged for distributing malware.

3.3 Inbound HTTP Requests Dataset

In order to contrast abuse with legitimate behavior, we rely on a second dataset that consists of de-identified HTTP logs restricted to signed-in users to the same subset of Google services we study for abuse. Log entries consist of $\langle$*PUID, User-Agent, service, IP, t*$\rangle$ containing a hashed, pseudo-anonymous account ID, User-Agent string, service identifier for what service the user interacted with, the user's IP address, and

fine-grained microsecond timestamp of the event. The logs contain hundreds of millions de-identified users from the a 14 day window ending on April 21, 2015. We use this data solely to estimate a lower bound on the aggregate number of users and User-Agents per IP; gauge the stability of $\langle IP, PUID \rangle$ pairs over time; and estimate the volume of legitimate traffic that a service would erroneously block with IP blacklists. For ethical and privacy reasons, all user data was handled by exclusively by Google, covered by their Terms of Service, and approved by their internal privacy review board.

3.4 Limitations

Our study suffers from a number of limitations that we lay out herein. First, our coverage of abuse is limited to IPv4 addresses. Based on inbound HTTP requests to Google, we estimate this covers 94% of signed-in traffic; the remaining 6% originates from IPv6 clients. This is higher than previous findings by Czyz et al. which reported IPv6 adoption at less than 1% of Internet traffic [6] or 4.8% adoption reported by Kreibich et al. for a sample of clients operating Netalyzer [13]. We make no claims our study of IPv4 abuse translates to IPv6.

Second, we caution that each service reporting malicious IP addresses relies on a specialized definition of abuse that is likely biased towards threats facing Google. We take such reports at face value—we cannot validate the precision or recall of the logic involved. This is identical to how participants in threat exchanges are blind to the accuracy of anti-abuse pipelines deployed by other members. As such, when we investigate the scale of abuse in Section 4 or examine the overlap of abusive IP addresses between services in Section 6 our results are biased towards the quality of abuse reports and their respective coverage.

4 Comparing Abuse Perspectives

Miscreants control a vast apparatus of hosts that, as a collective, encompasses 45 million IP addresses. We explore the scale of individual threats and the geographic specialization of attacks. We tie these into a broader understanding of bias introduced into threat exchanges due to participants with skewed abuse perspectives.

4.1 Scale of Abusive Networks

In aggregate, we estimate that miscreants control over eight million unique daily IP addresses as detailed in Figure 1. The strata between abuse verticals provides a lens into the allocation of compromised hosts on the Internet. Our email spam dataset contains the largest volume of abuse totaling nearly five million daily IP addresses. This is an order of magnitude larger than the number of hosts involved in account-based abuse affecting Google such as fake engagement, comment spam, or fraudulent account creation. Even more, email spam represents a 200x increase over hosts serving drive-by downloads and malware. While we avoid characterizing the volume of IP addresses as a reflection of the most pressing abuse challenges (or the criminal profit involved),

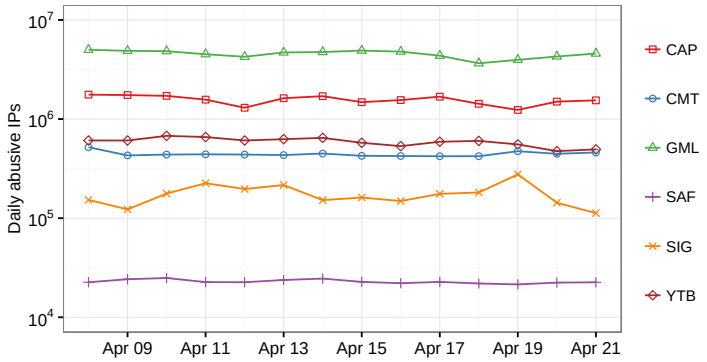


Fig. 1: Daily volume of IP addresses reported by various Google services for abuse. We observe an order of magnitude more spam bots than all other threats.

we argue (infected) hosts targeting Google during our collection period heavily skew towards email spam—a timeless staple of the underground monetization [15, 17].

Despite drastically different counts of abusive IP addresses between verticals, we find that the distribution of abusive traffic across IP addresses all follow a Zipf-like distribution as detailed in Figure 2. At the most concentrated extreme we find the top 1% of abusive email IP addresses relay 82% of inbound Gmail spam. While email is biased towards large SMTP relays active for all 14 days of our collection, we nevertheless observe a similar pattern with IP addresses linked also to failed CAPTCHAs that appear more transiently throughout our collection (median of five days). At the most distributed end of the spectrum, we find 48% of fake YouTube engagement originates from the top 1% of abusive IP addresses. We find similar patterns for other signed-in fraud. We suspect that miscreants favor this more decentralized approach to avoid services that cluster abusive accounts based on IP addresses. Nevertheless, our results present an opportunity to systematically block a significant volume of abuse from only a few hundred thousand hosts—assuming the IP addresses do not also relay legitimate traffic due to either re-allocation or over subscription as we explore in Section 5.

4.2 Network Locality and Specialization

While we observe abuse from networks around the globe, six countries in particular host the largest volume of abusive IP addresses: the United States (12.5%), Brazil (5.9%), Germany (4.6%), Russia (3.7%), India (3.6%), and China (3%). Combined, these regions cover 27–64% of all abusive IP addresses per service. We find some attacks are niche to specific localities as illustrated in Figure 3. For instance, networks in the United States serve the majority of malware and drive-bys (41%), followed by China (10%). With respect to bulk account, Indian networks create the most fake accounts (10%). This suggests that while miscreants rely on access to any compromised host possible, we find hints of bias potentially introduced by regional specialization within the underground or greater geo-political threats. This observation is consistent with prior work that shows regional biases in other attack vectors [36]. One potential root cause is underground market dynamics: hosts outside of Europe and the United States are less

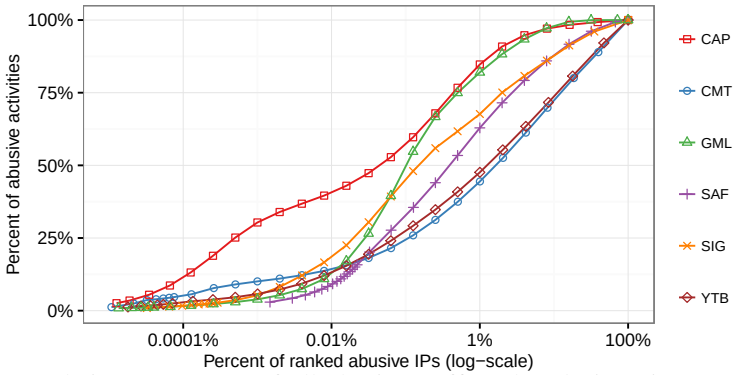


Fig. 2: Cumulative percentage of all abusive traffic relayed via unique IP addresses (ranked by contribution). The top 1% of abusive IP addresses contribute 48–82% of abusive activity.

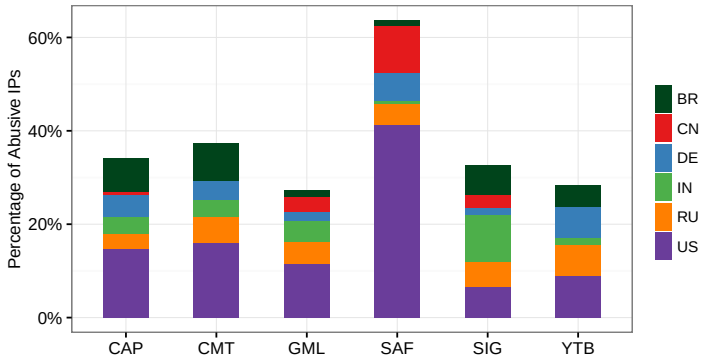


Fig. 3: Geolocation of abusive IP addresses for the top 6 offending regions. We observe a geographic bias in threats: malicious hosting in the United States; fake engagement from Russia; and bulk account creation in India.

expensive on the pay-per-install market and may be favored by miscreants for abuse with minimal bandwidth requirements [4].

5 Characterizing Abusive IP Addresses

We characterize the network-level behaviors of abusive IP addresses including the impact of DHCP churn and NAT on reconciling abuse reports. Given the diverse infrastructure and geographic distribution in each abuse vertical, we also examine whether any particular threat is more amenable to outright IP blacklisting.

5.1 Stability of IP-Device Pairs

One of the primary challenges of IP reputation is the stability of IP addresses as identifiers for abusive clients. While we observe a roughly constant daily volume of abusive

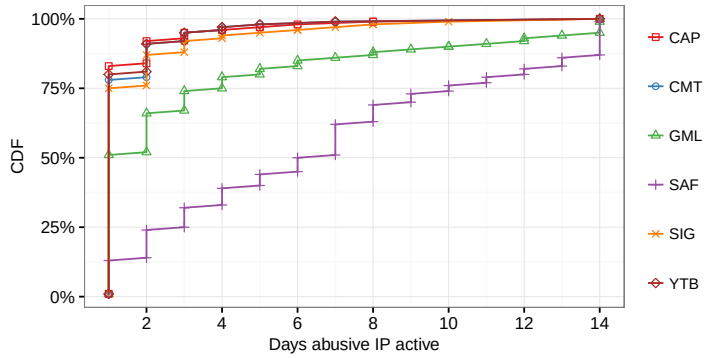


Fig. 4: CDF of the total number of days in the last two weeks a service reported an IP for abuse. We find 66% of abusive hosts persist for only a single day and preclude long term reputation tracking.

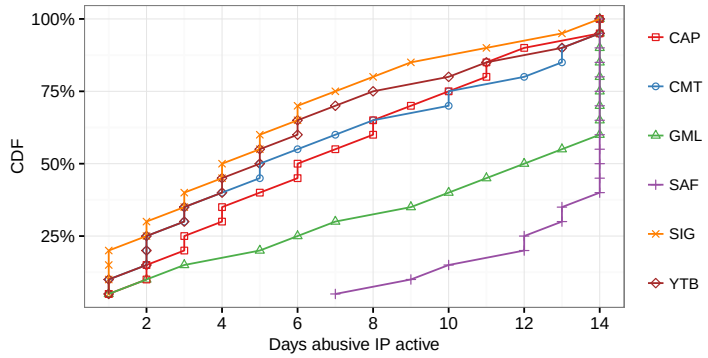


Fig. 5: CDF of the total number of days in the last two weeks a service reported an IP for abuse, restricted to the top 1% most abusive IP addresses. The most abusive IP addresses remain stable for longer periods, enabling longer term reputation tracking.

hosts as previously discussed in Section 4, we, somewhat surprisingly, find 66% of the hosts in our dataset actively relay abuse for only a single day over a two week period. One potential culprit for these observations is IP dynamism. We provide a more detailed breakdown of the duration of abuse per service in Figure 4. Absent emails spam and malicious hosting, 75–80% of abusive IP addresses persist for a single day. We find 14% of spam SMTP relays persist for 7 days as well as 49% of servers hosting malware, allowing for more stable IP reputation.

If we restrict our analysis to the top 1% of abusive IP addresses, a different picture emerges as shown in Figure 5. We find 52% of the top abusive SMTP relays actively send spam every day. Bulk account creation appears the least amenable to long-term IP reputation: 75% of IP addresses appear for fewer than seven days and 50% fewer than four days. We observe a similar pattern for the other top IP addresses involved in signed-in abuse.

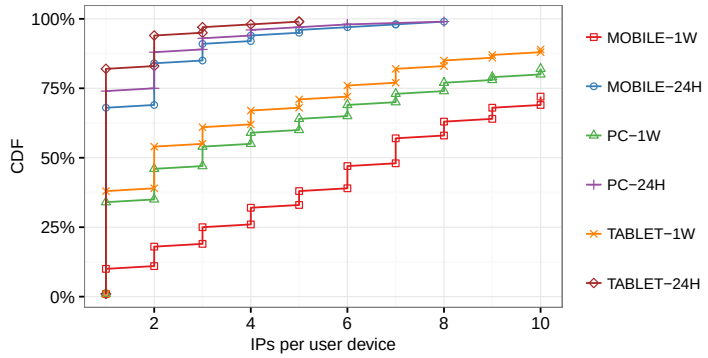


Fig. 6: Unique IP addresses per device for successively larger periods of time, broken down by device class. IP-device pairings rarely survive beyond 24 hours.

The unsuitability of IP addresses as long-term stable identifiers is well documented. For example, Maier et al. observed that ISPs would re-assign 50% of IP addresses to at least 2 different customers in the course of 24 hours [16]. We re-visit this issue and estimate the duration that device, IP pairs remain stable. Using our HTTP request logs, we first approximate a unique device identifier as a $\langle$*PUID, User-Agent*$\rangle$ pair and then calculate the number of IP addresses per device for 24 hours and one week.[4] Given the possibility that mobility rather than reallocation explains short IP leases, we segment devices by class (e.g., mobile, tablet, personal computer) as gleaned from the OS family of a User-Agent.

We present our findings in Figure 6. We observe that 74% of PCs maintain the same IP over a 24 hour period compared to 68% of mobile phones and 82% of tablets. After one week, only 10% of mobile devices retain their original IP compared to 34% of PCs and 38% of tablets—clients likely behind static IP addresses. This is a strict underestimate as we may not observe clients during every DHCP lease window.

On average, we find devices that send at least one request for every day in our two week collection period cycle through 20 distinct IP addresses. Furthermore, we find that 50% of all IP addresses remain active for the entirety of our collection period. As such, even though we observe short DHCP leases, ISPs quickly allocate the IP to a new set of devices. Our results differ drastically from previous measurements by Casado et al. who found that only 8% of clients (identified by cookies) used more than three IP addresses over 2–4 weeks [5]. In summary, we argue that IP intelligence carries value for only a limited time frame, after which ISPs may re-allocate a previously abusive IP to a benign set of clients. For our own work, we restrict all subsequent IP analysis to 24 hour windows.

5.2 Diverse Device Traffic

A second challenge of IP reputation is the coarse granularity of addresses compared to the diverse user populations they potentially service. Based on our HTTP request logs,

[4] While clients may report spoofed User-Agents, we assume that the majority of non-abusive users accurately report their device information.

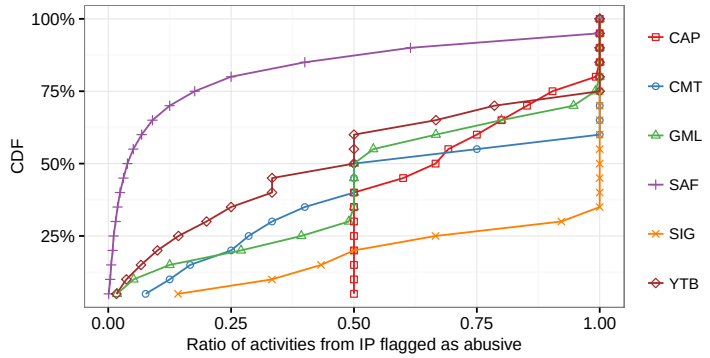


Fig. 7: CDF of the ratio of traffic from each IP flagged as abusive. Most IP addresses exhibit a mixture of legitimate and abusive activity.

we find 67% of IP addresses service at least two unique ⟨*PUID, User-Agent*⟩ devices in a 24 hour window and 21% of IP addresses service at least 5 devices. We caution this is limited to signed-in users and thus likely underestimates the number of devices per IP. Our findings are higher than a prior report in 2011 by Ihm et al. who observed only a single User-Agent for 83–94% of IP addresses depending on geographic region [11]. One explanation—matching the observations of Ihm et al.—is that networks have densified over time. The consequence for IP reputation is that NATs are becoming increasingly coarse approximations of the devices served.

In the presence of large NATs or SMTP relays, exclusively abusive IP addresses will be a rarity. Indeed, most IP addresses in our dataset are unsuitable for outright blacklisting: only 38% exclusively relay abusive content. However, in aggregate these exclusively abusive IP addresses carry 16%–49% of all malicious activity per service. We provide a more detailed breakdown of the fraction of traffic per IP flagged as abusive in a 24 hour window in Figure 7. Surprisingly, we find the top 1% of abusive IP addresses exhibit a lower likelihood of exclusive abuse compared to all abusive IP addresses as shown in Figure 8. This precludes the possibility of outright blacklisting many of the top offending IP addresses.

Malicious hosting represents one extreme where website content on 87% of all abusive IP addresses is more likely to be innocuous than harmful. Only 5% of malicious hosting IP addresses exclusively serve harmful content. This mirrors previous findings by Provos et al. who found drive-by downloads predominantly relied on compromised websites that otherwise serve legitimate content [24]. Bulk account registration presents an opposite extreme where 69% of all IP addresses exclusively create fake accounts. Intuitively, account creation should be a rare event per IP with limited exceptions for mobile gateways. Other abuse verticals fall between these extremes and provide a less pristine signal for filtering, reinforcing our observation that the complex interplay between NATs and devices obstructs any path towards daily blacklist deployment.

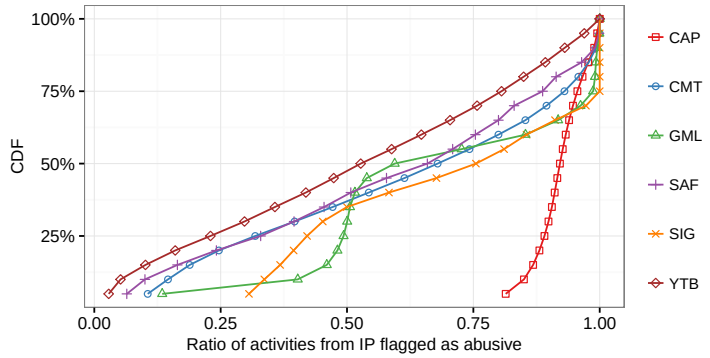


Fig. 8: CDF of the ratio of activity from each IP flagged as abusive for the top 1% of harmful IP addresses. Even the worst offending hosts relay significant legitimate activity and cannot be blacklisted.

5.3 Reputation Across IP Re-Assignment

While DHCP churn impedes long-term reputation tracking, a significant question remains whether the same device continuously abuses services across days. In the absence of device-level identifiers, we instead examine a microcosm of the same problem: the number of days static IP addresses that remain abusive. We isolate static (or at least minimally churning) IP addresses by examining rare instances where we see at least one (legitimate) PUID make a HTTP request from the same IP for a minimum of seven days. This approach is aided in part by NATing: so long as one user behind the NAT remains active, we can approximate that ISPs never re-assigned the IP. We note this is a strict subset of static IP addresses as (1) not all static IP addresses will have user-traffic (e.g., web servers) and (2) not all users on static IP addresses will exhibit constant activity. In total, we identify between 26,000—530,000 static IP samples per abuse vertical with the exception of Safe Browsing where only 485 malicious hosts also carried consistent user traffic.

Even without DHCP churn we find that multi-day abusive IP addresses are in fact rare as shown in Figure 9. With the exception of email spam and malicious hosting, miscreants use 72–80% of static IP addresses for only a single day in the last 14 compared to 75–80% of all IP addresses. Our results indicate that even were we able to track IP re-assignments, the likelihood of abuse in the next 24 hours is only loosely predicted by previous abuse (20–28% of IP addresses outside email and hosting). This leaves anti-abuse pipelines only a short window in which to detect abuse before miscreants migrate to entirely different devices and networks. However, we cannot rule out the possibility that miscreants continuously abuse dynamic IP addresses with full knowledge that churn provides greater anonymity to anti-abuse detection compared to static IP addresses. Furthermore, the rare exceptions matter: as we pointed out in previously in Figure 5 the top 1% of abusive IP addresses tend to be active for multiple days.

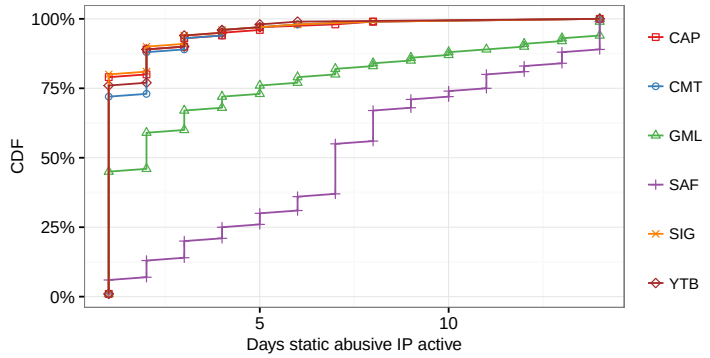


Fig. 9: CDF of the number of days exclusively static IP addresses remain abusive. Even without DHCP churn, attacks appear transient (potentially to avoid detection.)

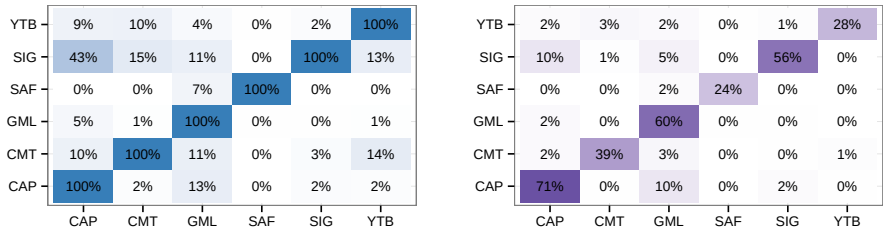
5.4 Subnet Abuse Affinity

We find that spatial qualities of networks provide a weak predictor of abusive IP addresses. For each /24 and /16 subnet containing at least one abusive IP, we calculate the likelihood miscreants abuse other IP addresses in the same subnet. Mechanistically, we calculate the ratio of observed IP addresses in HTTP request logs versus IP addresses reported by any service for abuse over 2 weeks. We find a median of 15% of observed IP addresses per /24 relay abuse. This increases to 24% at the /16 subnet level. If we restrict ourselves to the top 1% abusive IP addresses, this actually falls to 0.7% for /24 networks and 0.08% for /16 networks due to omitting more dynamic, short lived IP addresses and subnets.

The limited effectiveness of network topology in identifying abuse stems in part from devices migrating across subnet boundaries upon DHCP lease expiration. Using our HTTP request logs, if we compare a $\langle$*PUID, User-Agent*$\rangle$ device tuples' original and subsequent IP, we find 90% of devices cross a /24 boundary and 70% cross a /16 boundary. Only 6% of clients cross an ASN boundary and 0.6% appear in an entirely different geolocation after switching IP addresses.[5]

Translated into an abuse context, we observe 78–96% of /24 subnets contain only a single abusive IP per service in a 14 day window and 91–100% at most two abusive IP addresses. If we restrict our analysis to the top 1% of abusive IP addresses, we still find 76–92% of /24 subnets contain a single abusive IP. We note we cannot precisely identify where abusive clients migrate upon DHCP lease expiration as not all abuse verticals require account credentials and, more problematic, miscreants may access the same abusive account via multiple compromised devices and networks while reporting a spoofed User-Agent. Our findings illustrate that spatial properties of networks have little bearing on how ISPs re-assign devices to IP addresses. As a result, we advocate the most effective reputation system must operate on a per-IP rather than subnet granularity.

[5] ASN transitions may also occur due to a single network operator controlling multiple AS numbers, or alternatively, users may log in from duplicate devices (in terms of User-Agents) in different networks. Geolocation variations are within the predicted error of geolocation services.



(a) $\tau >= 0$ (b) $\tau >= 0.9$

Fig. 10: Fraction of abusive traffic from a service (y-axis) overlapping with a list of abusive IP addresses from a second service (x-axis) with an abuse likelihood greater than τ.

6 Cross-Vertical Abuse

We now turn our attention to investigating the impact of sharing intelligence across heterogeneous web services. In terms of the absolute number of abusive IP addresses, we find that cross-vertical abuse is a rare event: miscreants use only 6% IP addresses to attack at least two services in a 24 hour window. However, in aggregate these IP addresses relay 5–43% of all abuse per service. We examine which services benefit the most from sharing abuse intelligence due to miscreants re-using underground infrastructure as well as limitations of global IP reputation tracking.

6.1 Overlapping Abuse Verticals

To gauge the value of threat exchanges, we estimate the percentage of all abusive traffic per service S_1 that overlaps with malicious IP addresses reported by a second service S_2. We rely on an asymmetric calculation for the total volume of abuse caught:

$$\frac{|S_1 \cap S_2|}{|S_1|}$$

We present our results in Figure 10a. As many abusive IP addresses relay a significant volume of legitimate traffic, we present a thresholded calculation restricted to IP addresses in S_2 with an abuse likelihood greater than 90% in Figure 10b. Even absent thresholding, we find the majority of abusive IP addresses exclusively target individual services. These findings are consistent with high-level observations in related work that identified little cross-abuse IP intersections between spam, phishing, and network scanning IP reputation lists [36]. However, those few IPs that do overlap generate a significant volume of abusive traffic. We discuss a few salient instances where abuse intelligence sharing has the strongest impact.

Email Spam: Of all threats, IP addresses flagged for email spam provide the strongest predictor of abuse affecting other services. Coverage varies between 4–13% of all abusive traffic, dropping to 2–10% if we examine only SMTP relays where 90% of all email

sent is spam. Nevertheless, even with Gmail reporting five million daily IP addresses for spam, we find a significant fraction of abusive hosts engage solely in spam-based monetization rather than other forms of abuse.

Account-based Abuse: We find that miscreants registering fake accounts nominally re-use the same infrastructure for comment spam (15%) and YouTube fake engagement (13%). This falls to 0–1% if we examine thresholded abusive IP addresses. The weak correlation suggests that miscreants either stockpile accounts for more than 24 hours prior to abuse, or that vertically integrated account creation and monetization may in fact be rare due to specialized account merchants [33]. Furthermore, even though comments and fake engagement both require Google accounts, we find infrastructure involved in either vertical rarely overlaps in a 24 hour period: only 10–14% of spam comments and fake likes and subscribes originate from the same IP address. Consequently, once a fake account evades initial detection upon registration, each service contending with account-based fraud must detect specialized threats even though other verticals may have more mature abuse prevention systems.

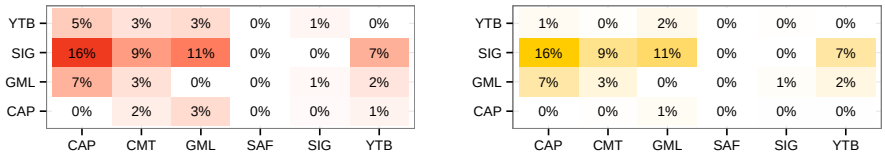
CAPTCHAs: The wide-spread adoption of CAPTCHAs across services including account creation and commenting creates a strong tendency where IP addresses that fail CAPTCHAs also tend to abuse other services. In particular, 43% of all bulk registered accounts overlap with an IP that failed a CAPTCHA. However, as CAPTCHA failure is only a weak signal of abuse, if we restrict our analysis to IP addresses that fail over 90% of CAPTCHA attempts (e.g., potentially automated solving), IP reputation catches only 10% of abusive accounts.

Hosting Exclusivity: We observe a negligible overlap between web servers that miscreants compromise for malicious hosting and other forms of abuse. As a small exception, we find 7% of malicious webpages overlap with IP addresses also serving as spam SMTP relays. Consequently, it appears that miscreants rarely re-use compromised web servers to proxy traffic for other abuse verticals. Our findings indicate that web security scanners for malware hosting cannot expedite their search coverage by scanning hosts also involved in email spam or other abuse.

6.2 Limitations of Intelligence Sharing

The coarse granularity of IP addresses as identifiers of abuse comes at a cost of false positives in the event of outright blacklisting. We estimate the volume of legitimate traffic in our signed-in HTTP request logs to a service S_1 that erroneously overlaps with abusive IP addresses reported by a service S_2. Given abuse appears in our request logs, we first optimistically filter all requests originating from IP addresses reported as abusive by S_1. Furthermore, as our request logs are limited to signed-in activity, we restrict our false positive estimates to YouTube engagement, account registration, outbound email, and CAPTCHAs solved by signed-in users.

We present our results for all abusive IP addresses in Figure 11a and the same calculation restricted to IP addresses with an abuse likelihood greater than 90% in Figure 11b. Without any threshold, IP blacklists would block 6–16% of all newly created, legitimate users and 0–5% of YouTube engagement. Even with thresholding, we find IP



(a) $\tau >= 0$ (b) $\tau >= 0.9$

Fig. 11: Fraction of benign IP addresses from a service (y-axis) that erroneously overlap with abusive IP addresses reported by a secondary service (x-axis) with an abuse likelihood greater than τ.

blacklists negatively impact account growth. We note our estimates are upper bounds on the volume of false positives as some erroneously blocked traffic may in fact reflect abuse currently unreported by an affected service. Similar trade-offs between blacklist false positives and false negatives with respect to the ratio of benign (so called ham) to spam traffic for email spam blacklists were demonstrated by Sinha et al. [29]. A key observation here is that individual organizations may have differing sensitivities to false positives, and that this sensitivity may even vary by abuse type.

Our findings indicate that while per-service, per-IP reputation captures a significant volume of abuse (as reported in Section 5), translating that abuse intelligence across services remains a challenge. This stems from the diverse user bases and devices served by a single IP address; while only a fraction of those devices may connect to one service, other services experience an entirely distinct user distribution. The end result is a limited utility for outright blacklisting based on cross-service intelligence.

7 Related Work

7.1 Characterizing IP Addresses

Research has spent a significant amount of effort to understand the network-level behaviors of (abusive) clients. This includes estimating the number of clients behind NATed IP addresses based on usage patterns [18], measuring the duration of DHCP leases and traffic patterns [16], examining the densification of networks over time [11], and understanding the behaviors of edge networks [5, 13]. Within an abuse context, Xie et al. examined how to automatically detect dynamic IP addresses and the prevalence of abuse among such hosts [34], with an extension to automatically classify large NATed IP addresses as exclusively spam relays [10]. Ramachandran et al. examined the network behavior of spam bots and the confinement of abusive hosts to a subset of networks and autonomous systems [26]. We compared these prior estimates to our own findings and examined the impact of NAT and DHCP churn on the effectiveness of IP reputation tracking.

7.2 Blacklist Efficacy

IP blacklists enable web services to identify and penalize abusive hosts in the absence of overt user identifiers (e.g., authenticated session cookies). A number of studies previously examined the efficacy and global applicability of spam-based blacklists. The closest themes to our own work include estimating the coverage of various blacklists with respect to spamvertised abuse and affiliate programs [22], the performance of email spam blacklists when applied to alternative domains such as social network spam [8, 31], the lack of overlap of abusive domains and IP addresses for blacklists targeting spam, phishing, and malware [36], instances where popular blacklists identify the same threats [12], and the efficacy of public versus commercial malware blacklists when applied to major malware families [14]. We provided a Google-centric view of abuse exclusivity across products which mirror these previous findings: abuse is an incredibly large and diverse space where miscreants are highly specialized. However, examining only the overlap of abusive IP addresses fails to capture skewed traffic emanating from rare co-occurrences.

8 Summary

In this work, we measured the effectiveness of centralized reputation tracking as a tool for identifying miscreants who leverage the same machine for spam, denial of service, malicious hosting, and other forms of automated abuse. We focused initially on the scale of these individuals threats and found they differ by two orders of magnitude, with email spam reigning as our top source of abusive hosts. Despite 8 million IP addresses reported every day by one of six Google services for abuse, blocking only 1% of these sources can prevent 48–82% of all harmful traffic. We found some of these threats exhibit a local specialization: malicious hosting in the United States; fake engagement in Russia; and bulk account registration in India.

Transforming this intelligence into actionable reputation data proved challenging: 66% of abusive IP addresses remained active for only a single day, in part driven by dynamic reallocation where the average (compromised) device cycled through 20 IP addresses over the course of two weeks. Equally problematic, NATs representing large user populations polluted IP reputation to the point where only 38% of abusive IP addresses exclusively relayed harmful traffic, while the rest shared some overlap with benign devices. Nevertheless, we found this minority of hosts delivered 16–49% of all abuse per service. Ultimately, we found that hosts involved in cross-service abuse were in fact rare: only 6% of abusive IP addresses negatively affected at least two services within a 24 hour window. However, combined, these hosts generated 5–43% of all abuse per service. In the end, we argued less mature anti-abuse pipelines for new products or Internet services could tap into such an intelligence feed to benefit from threat exchanges of seemingly unrelated attacks. However, these benefits must come from machine learning pipelines—outright blacklisting remains out of reach due to collateral damage to legitimate users.

Acknowledgments

This work was supported in part by the National Science Foundation under contracts CNS 1409758, CNS 1111699, and CNS 1518741. 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. Ross Anderson, Chris Barton, Rainer Böhme, Richard Clayton, Michel J.G. van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. Measuring the cost of cybercrime. In *Proceedings of the Workshop on Economics of Information Security (WEIS)*, 2012.
2. Hadi Asghari, Michael Ciere, and Michel JG Van Eeten. Post-mortem of a zombie: conficker cleanup after six years. In *Proceedings of the USENIX Security Symposium*, 2015.
3. Brad Taylor. It's not about the spam. `http://goo.gl/zzAL4N`, 2007.
4. Juan Caballero, Chris Grier, Christian Kreibich, and Vern Paxson. Measuring pay-per-install: The commoditization of malware distribution. In *USENIX Security Symposium*, 2011.
5. Martin Casado and Michael J Freedman. Peering through the shroud: The effect of edge opacity on ip-based client identification. In *Proceedings of the Symposium on Networked Systems Design and Implementation*, 2007.
6. Jakub Czyz, Mark Allman, Jing Zhang, Scott Iekel-Johnson, Eric Osterweil, and Michael Bailey. Measuring ipv6 adoption. In *Proceedings of the ACM conference on SIGCOMM*, 2014.
7. DShield. DShield. `https://www.dshield.org/`, 2015.
8. Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the ACM Conference on Computer and Communications Security*, 2010.
9. Mark Hammell. ThreatExchange: Sharing for a safer internet. `http://on.fb.me/1zvuPdS`, 2015.
10. Chi-Yao Hong, Fang Yu, and Yinglian Xie. Populated ip addresses: Classification and applications. In *Proceedings of the Conference on Computer and Communications Security*, 2012.
11. Sunghwan Ihm and Vivek S Pai. Towards understanding modern web traffic. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2011.
12. Jaeyeon Jung and Emil Sit. An empirical study of spam traffic and the use of dns black lists. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2004.
13. Christian Kreibich, Nicholas Weaver, Boris Nechaev, and Vern Paxson. Netalyzr: illuminating the edge network. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2010.
14. Marc Kührer, Christian Rossow, and Thorsten Holz. Paint it black: Evaluating the effectiveness of malware blacklists. In *Proceedings of Research in Attacks, Intrusions and Defenses*. 2014.
15. Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, and Chris Kanich et al. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2011.
16. Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. On dominant characteristics of residential broadband internet traffic. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2009.

17. Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M Voelker, Stefan Savage, and Kirill Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the 21st USENIX conference on Security symposium*, 2012.
18. Ahmed Metwally and Matt Paduano. Estimating the number of users behind ip addresses for combating abusive traffic. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.
19. Ron Miller. AlienVault Announces More Social Threat Exchange. `http://tcrn.ch/1FL7E8A`, 2015.
20. Alan Neville and Ross Gibb. ZeroAccess Indepth. `http://goo.gl/j0eMHr`, 2013.
21. Paul Pearce, Vacha Dave, Chris Grier, Kirill Levchenko, Saikat Guha, Damon McCoy, Vern Paxson, Stefan Savage, and Geoffrey M Voelker. Characterizing large-scale click fraud in zeroaccess. In *Proceedings of the Conference on Computer and Communications Security*, 2014.
22. Andreas Pitsillidis, Chris Kanich, Geoffrey M Voelker, Kirill Levchenko, and Stefan Savage. Taster's choice: a comparative analysis of spam feeds. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2012.
23. Niels Provos. Safe Browsing - Protecting Web Users for 5 Years and Counting. `http://goo.gl/psdXkP`, 2012.
24. Niels Provos, Panayiotis Mavrommatis, Moheeb Abu Rajab, and Fabian Monrose. All your iFRAMEs point to us. In *Proceedings of the USENIX Security Symposium*, 2008.
25. Tim Rains. Microsoft Interflow: a new Security and Threat Information Exchange Platform. `http://bit.ly/1SKpcs2`, 2015.
26. Anirudh Ramachandran and Nick Feamster. Understanding the network-level behavior of spammers. In *Proceedings of the ACM conference on SIGCOMM*, 2006.
27. Michael Rowinski. More than 1,000 Organizations Join IBM to Battle Cybercrime. `https://www-03.ibm.com/press/us/en/pressrelease/46856.wss`, 2015.
28. Prosenjit Sinha, Amine Boukhtouta, Victor Heber Belarde, and Mourad Debbabi. Insights from the analysis of the mariposa botnet. In *Proceedings of the International Conference on Risks and Security of Internet and Systems (CRiSIS)*, 2010.
29. Sushant Sinha, Michael Bailey, and Farnam Jahanian. Improving spam blacklisting through dynamic thresholding and speculative aggregation. In *Proceedings of the Network & Distributed System Security Symposium*, 2010.
30. Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. Your botnet is my botnet: analysis of a botnet takeover. In *Proceedings of the ACM conference on Computer and Communications Security*, 2009.
31. Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the Internet Measurement Conference*, 2011.
32. Kurt Thomas, Danny Yuxing Huang, David Wang, Elie Bursztein, Chris Grier, and Thomas J. Holt et al. Framing dependencies introduced by underground commoditization. In *Proceedings of the Workshop on the Economics of Information Security*, 2015.
33. Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *Proceedings of the USENIX Security Symposium*, 2013.
34. Yinglian Xie, Fang Yu, Kannan Achan, Eliot Gillum, Moises Goldszmidt, and Ted Wobber. How dynamic are ip addresses? In *Proceedings of the ACM conference on SIGCOMM*, 2007.
35. Fang Yu, Yinglian Xie, and Qifa Ke. Sbotminer: large scale search bot detection. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2010.
36. Jing Zhang, Ari Chivukula, Michael Bailey, Manish Karir, and Mingyan Liu. Characterization of blacklists and tainted network traffic. In *Passive and Active Measurement*, 2013.