ASwatch: An AS Reputation System to Expose Bulletproof Hosting ASes

Maria Konte
Georgia Tech
mkonte@cc.gatech.edu

Roberto Perdisci
University of Georgia
perdisci@cs.uga.edu

Nick Feamster
Princeton University
feamster@cs.princeton.edu

ABSTRACT

Bulletproof hosting Autonomous Systems (ASes)—malicious ASes fully dedicated to supporting cybercrime—provide freedom and resources for a cyber-criminal to operate. Their services include hosting a wide range of illegal content, botnet C&C servers, and other malicious resources. Thousands of new ASes are registered every year, many of which are often used exclusively to facilitate cybercrime. A natural approach to squelching bulletproof hosting ASes is to develop a reputation system that can identify them for takedown by law enforcement and as input to other attack detection systems (*e.g.*, spam filters, botnet detection systems). Unfortunately, current AS reputation systems rely primarily on data-plane monitoring of malicious activity from IP addresses (and thus can only detect malicious ASes after attacks are underway), and are not able to distinguish between *malicious* and *legitimate but abused* ASes.

As a complement to these systems, in this paper, we explore a fundamentally different approach to establishing AS reputation. We present *ASwatch*, a system that identifies malicious ASes using exclusively the *control-plane* (*i.e.*, routing) behavior of ASes. *ASwatch*'s design is based on the intuition that, in an attempt to evade possible detection and remediation efforts, malicious ASes exhibit "agile" control plane behavior (*e.g.*, short-lived routes, aggressive rewiring). We evaluate our system on known malicious ASes; our results show that *ASwatch* detects up to 93% of malicious ASes with a 5% false positive rate, which is reasonable to effectively complement existing defense systems.

CCS Concepts

•**Security and privacy** → **Network security;** •**Networks** → *Network monitoring;*

SIGCOMM'15, August 17–21, 2015, London, United Kingdom

DOI: http://dx.doi.org/10.1145/2785956.2787494

Keywords

AS Reputation; Bulletproof Hosting; Malicious Networks

1. INTRODUCTION

Today's cyber-criminals must carefully manage their network resources to evade detection and maintain profitable illicit businesses. For example, botmasters need to protect their botnet command-and-control (C&C) servers from takedowns, spammers need to rotate IP addresses to evade trivial blacklisting, and rogue online businesses need to set up proxies to mask scam hosting servers. Often, cyber-criminals accomplish these goals by hosting their services within a malicious autonomous system (AS) owned by an Internet service provider that willingly hosts and protects illicit activities. Such service providers are usually referred to as *bulletproof hosting* [7], due to their reluctance to address repeated abuse complaints regarding their customers and the illegal services they run. Notorious cases of malicious ASes include McColo [22], Intercage [19], Troyak [27], and Vline [2] (these ASes were taken down by law enforcement between 2008 and 2011). According to Hostexploit's reports [14], these types of ASes continue to appear in many regions around the world—mostly in smaller countries with lower levels of regulation, but also in the United States—to support activities ranging from hosting botnet command-and-control to phishing attacks [15]. For example, the Russian Business Network [31], one of the most notorious and still active cybercrime organizations, have decentralized their operations across multiple ASes. In most cases, nobody notices bulletproof hosting ASes until they have become hubs of illegal activities, at which point they are de-peered from their upstream providers. For example, Intercage [19] was de-peered more than ten times before it reached notoriety and was cut off from all upstream providers.

To defend against these crime-friendly ASes, the community has developed several AS reputation systems that monitor *data-plane* traffic for illicit activities. Existing AS reputation systems typically monitor network traffic from different vantage points to detect the presence of either malware-infected machines that contact their C&C servers, send spam, host phishing or scam websites, or perform other illicit activities. These systems establish AS reputa-

tion by measuring the "density" of malicious network activities hosted within an AS. For instance, FIRE [36] tracks the number of botnet C&C and drive-by malware download servers within an AS. ASes that host a large concentration of malware-related servers are then assigned a low reputation. Similarly, Hostexploit [14] and BGP Ranking [4] compute the reputation of an AS based on data collected from sources such as DShield [11] and a variety of IP and domain name blacklists.

Unfortunately, these existing AS reputation systems have a number of limitations: (1) They cannot distinguish between *malicious* and *legitimate but abused* ASes. Legitimate ASes often unwillingly host malicious network activities (*e.g.*, C&C servers, phishing sites) simply because the machines that they host are abused. For example, AS 26496 (GoDaddy) and AS 15169 (Google) repeatedly appeared for years among the ASes with lowest reputation, as reported by Hostexploit. Although these ASes are legitimate and typically respond to abuse complaints with corrective actions, they may simply be unable to keep pace with the level of abuse within their network. On the other hand, *malicious* ASes are typically unresponsive to security complaints and subject to law-enforcement takedown. (2) Because of the inability to distinguish between *malicious* and *legitimate but abused* ASes, it is not clear how to use the existing AS rankings to defend against *malicious* ASes. (3) Existing AS reputation systems require direct observation of malicious activity from many different vantage points and for an extended period of time, thus delaying detection.

We present a fundamentally different approach to establishing AS reputation. We design a system, *ASwatch*, that aims to identify malicious ASes using exclusively *control-plane* data (*i.e.*, the BGP routing control messages exchanged between ASes using BGP). Unlike existing *data-plane* based reputation systems, *ASwatch* explicitly aims to identify *malicious* ASes, rather than assigning low reputation to legitimate ASes that have unfortunately been abused.

Our work is motivated by the practical help that an AS reputation system, which accurately identifies malicious ASes, may offer: (1) Network administrators may handle traffic appropriately from ASes that are likely operated by cyber criminals. (2) Upstream providers may use reliable AS reputation in the peering decision process (e.g. charge higher a low reputation customer, or even de-peer early). (3) Law enforcement practitioners may prioritize their investigations and start early monitoring on ASes, which will likely need remediation steps.

The main intuition behind *ASwatch* is that malicious ASes may manipulate the Internet routing system, in ways that legitimate ASes do not, in an attempt to evade current detection and remediation efforts. For example, malicious ASes "rewire" with one another, forming groups of ASes, often for a relatively short period of time [20]. Only one AS from the group connects to a legitimate upstream provider, to ensure connectivity and protection for the group. Alternatively, they may connect directly to a legitimate upstream provider, in which case they may need to change upstream providers frequently, to avoid being de-peered and isolated from the rest of the internet. Changing providers is necessary because a legitimate upstream provider typically responds (albeit often slowly) to repeated abuse complaints concerning its customer ASes. Another example is that a malicious AS may advertise and use small blocks of its IP address space, so that as soon as one small block of IP addresses is blocked or blacklisted, a new block can be advertised and used to support malicious activities. To capture this intuition, we derive a collection of *control-plane features* that is evident solely from BGP traffic observed via Routeviews [32]. We then incorporate these features into a supervised learning algorithm, that automatically distinguishes malicious ASes from legitimate ones.

We offer the following contributions:

- We present *ASwatch*, an AS reputation system that aims to identify malicious ASes by monitoring their *control plane behavior*.
- We identify three families of features that aim to capture different aspects of the "agile" control plane behavior typical of malicious ASes. (1) *AS rewiring* captures aggressive changes in AS connectivity; (2) *BGP routing dynamics* capture routing behavior that may reflect criminal illicit operations; and (3) *Fragmentation and churn of the advertised IP address space* capture the partition and rotation of the advertised IP address space.
- We evaluate *ASwatch* on real cases of malicious ASes. We collect *ground truth* information about numerous malicious and legitimate ASes, and we show that *ASwatch* can achieve high true positive rates with reasonably low false positives. We evaluate our statistical features and find that the rewiring features are the most important.
- We compare the performance of *ASwatch* with BGP Ranking, a state-of-the-art AS reputation system that relies on data-plane information. Our analysis over nearly three years shows that *ASwatch* detects about 72% of the malicious ASes that were observable over this time period, whereas BGP Ranking detects only about 34%.

The rest of the paper is organized as follows. Section 2 offers background information about bulletproof hosting ASes. Section 3 describes the features we devised and an overview of our system. Section 4 discusses the evaluation of the system. Section 5 discusses various limitations of our work, Section 6 presents related work, and Section 7 concludes.

2. BACKGROUND

We define malicious and legitimate ASes and provide background information, with an emphasis on characteristics that are common across most confirmed cases of malicious ASes.

2.1 Bulletproof Hosting ASes

In this section, we describe more precisely the differences between malicious (bulletproof hosting) and legitimate ASes. We also discuss how malicious ASes tend to

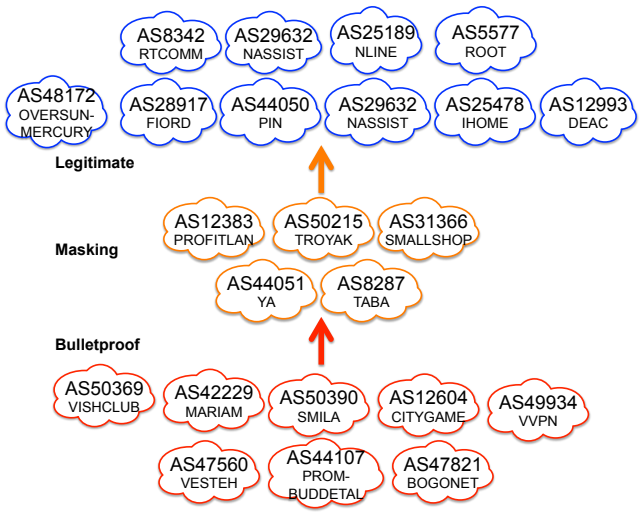
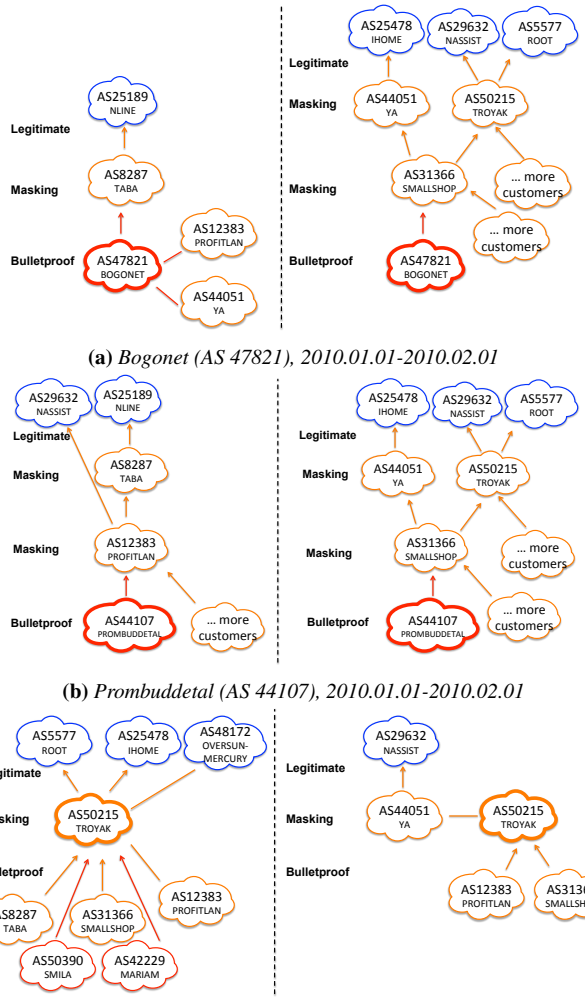


Figure 1: *The AS-TROYAK infrastructure (malicious ASes identified by blogs.rsa.com). The core of the infrastructure comprises eight bulletproof networks, which connect to legitimate ASes via a set of intermediate "masking" providers.*



(a) *Bogonet (AS 47821), 2010.01.01-2010.02.01*

(b) *Prombuddetal (AS 44107), 2010.01.01-2010.02.01*

(c) *Troyak (AS 50215), 2010.02.01-2010.04.01*

Figure 2: *Connectivity snapshots of three cases of ASes which are operated by cyber-criminals. All connected to a "masking" upstream provider. Directed edges represent customer-provider relationships; undirected edges represent peering relationships.*

connect with one another, and how some ISPs (some of which are themselves malicious) provide these ASes with upstream connectivity and protection. To illustrate this behavior, we explore a case study that shows how malicious ASes may be established and "rewired" in an attempt to evade current detection and takedown efforts.

Malicious vs. Legitimate ASes: We call an AS *malicious*, if it is managed and operated by cyber-criminals, and if its main purpose is to support illicit network activities (*e.g.*, phishing, malware distribution, botnets). In contrast, we refer to an AS as *legitimate*, if its main purpose is to provide legitimate Internet services. In some cases, a legitimate AS's IP address space may be abused by cyber-criminals to host malicious activities (*e.g.*, sending spam, hosting a botnet command-and-control server). Such abuse is distinct from those cases where cyber-criminals operate and manage the AS. *ASwatch* focuses on distinguishing between malicious and legitimate ASes; we aim to label *legitimate but abused* ASes as legitimate. Our approach is thus a significant departure from existing data-plane based AS reputation systems, which are limited to computing reputation by primarily focusing on data-plane abuse, rather than establishing if an AS is actually malicious.

Malicious AS Relationships: Bulletproof hosting ASes provide cyber-criminals with a safe environment to operate. Sometimes, malicious ASes form business relationships with one another to ensure upstream connectivity and protection. For example, they may connect to upstream providers that are themselves operated in part with criminal intent. In turn, these upstream ASes connect to legitimate ISPs, effectively providing cover for the bulletproof hosting ASes [2]. These "masking" upstream providers may not be actively engaged in cyber-criminal activity themselves (as observed from the data-plane). Consequently, network operators at legitimate ISPs may be unaware of the partnership among these "shady" upstream providers and bulletproof hosting ASes, making detection and remediation efforts more difficult.

Efforts to take down bulletproof hosting ASes have been ongoing since at least 2007, when upstream ISPs of the Russian Business Network (RBN) refused to route its traffic [21]. Many organizations track rogue ASes and report tens to hundreds of new rogue ASes every year [15]. Takedown efforts often result in a malicious AS moving to new upstream ISPs; for example, RBN now operates on many different ISP networks.

Case Study - Behavior of Malicious ASes: Figure 1 shows an example of a real network of eight bulletproof hosting ASes that connect to legitimate ASes via a set of intermediate "masking" providers. Notice that while we label the

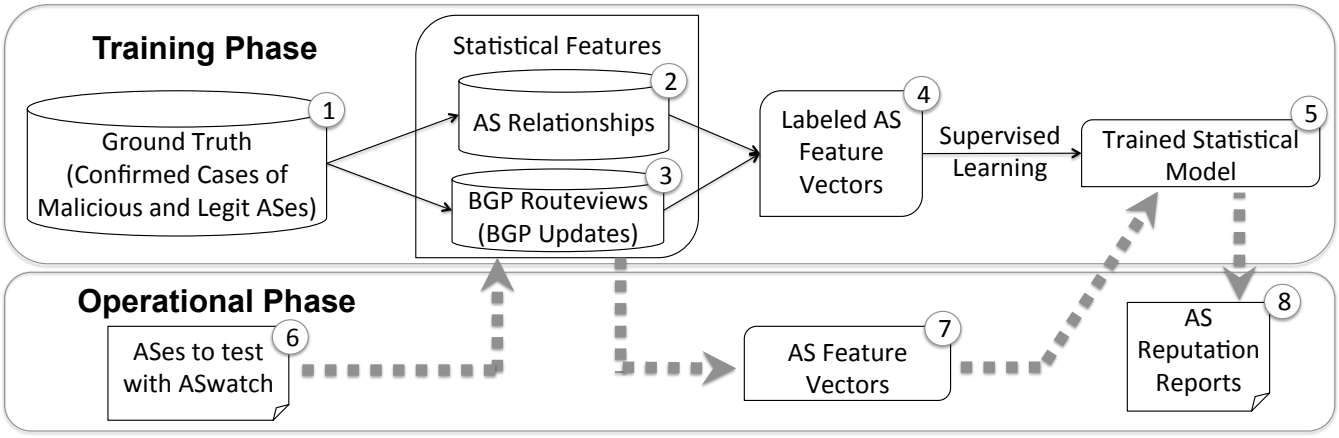


Figure 3: *ASwatch system architecture.*

malicious ASes in this case study, based on ground truth provided by blogs.rsa.com, we independently derive and analyze the relationships between the ASes from routing information. At the time they were reported by blogs.rsa.com (March 2010), the eight bulletproof ASes hosted a range of malware, including Zeus Trojans, RockPhish JabberZeus servers, and Gozi Trojan servers. We chose this as a case study because it represents one of the most well documented cases of known bulletproof hosting ASes, and is representative of other less well known incidents.

The bulletproof hosting ASes eventually switched between five upstream providers, which served as intermediaries to connect to the legitimate ASes. In turn, the upstream "masking" providers were customers of nine different legitimate ISPs.

To understand how malicious ASes form business relationships and how these relationships evolve over time, we tracked the upstream and downstream connectivity of the malicious ASes, as shown in Figures 1 and 2 (the figures show an activity period from January to April 2010; the malicious ASes went offline in March 2010).

We tracked the connectivity of one "masking" AS, Troyak (AS 50215), and two bulletproof hosting ASes, Bogonet (AS 47821) and Prombuddetal (AS 44107), that belong to the Troyak infrastructure. To track their upstream and downstream connectivity, we used a publicly available dataset from CAIDA, which provides snapshots of the AS graph, annotated with business relationships [25]. Figure 2 shows snapshots of the connectivity for the reported ASes.

All of these malicious ASes connected to a "masking" upstream provider, thus avoiding direct connectivity with legitimate ISPs, and also they change their connectivity between one another. For example, before takedown, Troyak had three upstream providers: Root, Ihome, and Oversun-Mercury. After the blog report on March 2010, Troyak lost *all* of its upstream providers and relied on a peering relationship with Ya for connectivity. After April 2010, Troyak and its customers went offline. Bogonet switched from Taba to Smallshop, and Prombuddetal switched from Profitlan to Smallshop, before going offline.

3. ASWATCH

ASwatch monitors globally visible BGP routing activity and AS relationships, to determine which ASes exhibit *control plane behavior* typical of malicious ASes. Because of the nature of their operations (criminal activity) and their need to fend off detection and possible take-down efforts, malicious ASes tend to exhibit control-plane behavior that is different from that of legitimate ASes. We now discuss how *ASwatch* works, including a detailed description of the features we used to differentiate between malicious and legitimate ASes, and our intuition for choosing each feature.

3.1 System Overview

Figure 3 presents an overview of *ASwatch*. The system has a training phase (Section 3.3.1) and an operational phase (Section 3.3.2). During the training phase, *ASwatch* learns the control-plane behavior of malicious and legitimate ASes. We provide the system with ① a list of known malicious and legitimate ASes (Section 4.1 describes this dataset). *ASwatch* tracks the control-plane behavior of the legitimate and malicious ASes over time using two sources of information: ② business relationships between ASes, and ③ BGP updates (from RouteViews). *ASwatch* then computes statistical features (Section 3.2 describes this process) from the previous inputs. Each AS is represented by a feature vector based on these statistical features ④. *ASwatch* uses these labeled feature vectors and a supervised learning algorithm to ⑤ train a statistical model. During the operational phase, we provide *ASwatch* with a list of new (not yet labeled) ASes ⑥ to be classified as legitimate or malicious using the same statistical features over the given time period. Then, *ASwatch* ⑦ computes the new AS feature vectors and ⑤ tests them against the previously trained statistical model. Finally, ⑧ the system assigns a reputation score to each AS.

3.2 Statistical Features

In this section, we describe the features we compute and the intuition for choosing them. Table 1 gives an overview of our feature families, and the most important group of features for each family. Given an AS, A, and time window, T, *ASwatch* monitors A's control-plane behavior and translates it into a feature vector consisting of three groups of features: rewiring activity, IP fragmentation and churn, and BGP routing dynamics.

Some of the behavioral characteristics we measure can be naturally described by a probability distribution, rather than a single numerical feature. In these cases, to capture the behavioral characteristics in a way that is more suitable for input to a statistical classifier, we translate each probability distribution into three numerical features that approximately describe the *shape* of the distribution. Specifically, we compute its 5th percentile, 95th percentile, and median. In the following, we refer to such features as *distribution characteristics*. We include these three values as features in the overall feature vector, and repeat this process for all behavioral characteristics that can be described as a probability distribution.

Notice that even though more values may more accurately summarize a distribution's shape, such a representation would significantly increase the overall size of the feature vector used to describe an AS. For this reason, we chose to only use three representative values, which we found to work well in practice.

We now explain in detail the features that *ASwatch* uses to establish AS reputation and motivate how we selected them.

3.2.1 Rewiring Activity

This group of features aims to capture the changes in A's connectivity. Our intuition is that malicious ASes have different connectivity behavior than legitimate ASes, because they tend to: (1) change providers more frequently to make detection and remediation more difficult; (2) connect with less popular providers, which may have less strict security procedures and may respond less promptly to abuse complaints, (3) have longer periods of downtime, possibly due to short-duration contracts or even de-peering from a legitimate upstream provider. In contrast, legitimate ASes tend to change their connectivity less frequently, typically due to business considerations (*e.g.*, a less expensive contract with a new provider).

To capture rewiring activity, *ASwatch* tracks changes to AS relationships (Step 2 in Figure 3). We use periodic snapshots of historic AS relationships, with one snapshot per month (Section 4.1 describes the data sets in more detail). A snapshot S_i contains the AS links annotated with the type of relationships, as observed at a given time t_i (*e.g.*, one snapshot is produced on the first day of each month).

AS presence and overall activity. Let A be the AS for which we want to compute our features. Given a sequence of N consecutive snapshots $\{S_i\}_{i=1}^{N}$, we capture the *presence* of an AS by measuring the total number of snapshots, C, and the maximum number of contiguous snapshots, M, in which A was present, the fraction C/N, and M/N (four features in total). To capture the overall activity of A, we measure the distribution (over time) of the number of customers, providers, and peers A links with for each snapshot. To summarize each of these distributions, we extract the distribution characteristics (5th percentile, 95th percentile, and median), as described earlier. This yields a total of nine features (three for each of the three types of AS relationships). We also count the total number and fraction (*i.e.*, normalized by C) of distinct customers, providers, and peers that A has linked with across all C snapshots when it was present, yielding another six features.

Feature Family	Description	Most Important Feature
Rewiring Activity	Changes in AS's connectivity (*e.g.*, frequent change of providers, customers or peers)	Link stability
IP Space Fragmentation & Churn	IP space partitioning in small prefixes & rotation of advertised prefixes	IP space fragmentation
BGP Routing Dynamics	BGP announcements patters (*e.g.*, short prefix announcements)	Prefix reachability

Table 1: *Overview of ASwatch feature families and the most important feature for each family.*

Link stability. We capture the *stability* of different types of relationships that an AS forms over time. For each of the C snapshots where A was present, we track all relationships between A and any other AS. Assuming A appeared as an upstream provider for another AS, say A^k, in v out of C snapshots, we compute the fraction $F^k = v/C$. We repeat this for all ASes where A appears as a provider at least once within C snapshots, thus obtaining a distribution of the F^k values. Finally, we summarize this distribution of the F^k values, computing the distribution characteristics as described above. We repeat this process, considering all ASes that appear as the upstream provider for A (*i.e.*, A is their customer), and for all ASes that have peering relationships with A. Overall, we compute nine features that summarize three different distributions (three features for each type of relationship).

Upstream connectivity. We attempt to capture *change* in the set of providers. Assume that from the i-th snapshot S_i we observed a total of M_i upstream providers for A, and call $\{A_i^k\}_{k=1}^{M_i}$ the set of upstream provider ASes. Then, for each pair of contiguous snapshots, S_i and S_{i+1}, we measure the Jaccard similarity coefficient $J_{i,i+1}$ between the sets $\{A_i^k\}$ and $\{A_{i+1}^k\}$. We repeat for all available $(N-1)$ pairs of consecutive snapshots, thus obtaining a distribution of Jaccard similarity coefficients. To summarize this distribution, we compute the distribution characteristics as described above, yielding three features. Figure 4 shows the CDF of the minimum Jaccard similarity, for the malicious and the legitimate

ASes. Overall, the legitimate ASes tend to have higher values of the Jaccard similarity metric, which indicates fewer changes in their upstream providers.

Attachment to popular providers. We aim to capture an AS's *preference* for "popular" providers. As previous work has shown [20], malicious ASes tend to connect more often with less prominent providers, which may have less strict security procedures and may respond less promptly to abuse complaints.

We compute the popularity of each provider per snapshot and across all snapshots. To this end, we first empirically derive the distribution of the number of customers per provider. We then consider a provider to be (a) *very popular*, if it belongs to the top 1% of all providers overall; (b) *popular*, if it belongs to the top 5%; (c) *very popular with respect to a snapshot* S_i, if it belongs to the top 1% in S_i, and (d) *popular with respect to a snapshot* S_i, if it belongs to the top 5% in S_i.

We then gather all upstream providers that A has used and compute the fraction of these providers that fall into each of the four categories described above (thus yielding four features). Finally, we compute the fraction of snapshots in which A has linked to at least one provider falling into one of the above categories; we do this for each category, thus obtaining four more features.

We capture the overall rewiring behavior of an AS with a total number of thirty five features.

3.2.2 IP Space Fragmentation and Churn

Malicious ASes tend to partition their IP address space into small BGP prefixes and to advertise only some of these prefixes at any given time. One possible explanation for this behavior may be that they attempt to avoid having their entire IP address space blacklisted at once. For example, if a number of IP addresses within a given BGP prefix are detected as hosting malicious activities, a blacklist operator (*e.g.*, Spamhaus [35]) may decide to blacklist the entire prefix where the IP addresses reside. By fragmenting the IP address space and advertising only a subset of their BGP prefixes, the operators of a malicious AS may be able to quickly move malicious activities to a "fresh" space. They perform this maneuver by leveraging not-yet-blacklisted IP addresses within newly advertised prefixes. On the other hand, legitimate ASes tend to consistently advertise their available IP address space in less fragmented prefixes, as they do not need to attempt to evade blacklisting.

IP Space Fragmentation and Churn Features. We attempt to capture IP address *fragmentation* with the following features. Given a snapshot, we group the advertised BGP prefixes into contiguous IP *blocks*. For each, AS we count the number of BGP prefixes and the number of distinct `/8`, `/16`, and `/24` prefixes within each IP block. To capture the *churn* in the advertisement of the IP address space, we proceed as follows. Given a pair of adjacent snapshots for an AS, we measure the Jaccard similarity among the sets of BGP prefixes advertised by the AS in the two snapshots. Similarly, we compute the Jaccard index among the sets of `/8`, `/16`, and `/24` prefixes. We summarize the above four distributions using the *distribution characteristics* that we described earlier, thus obtaining a total of twelve features.

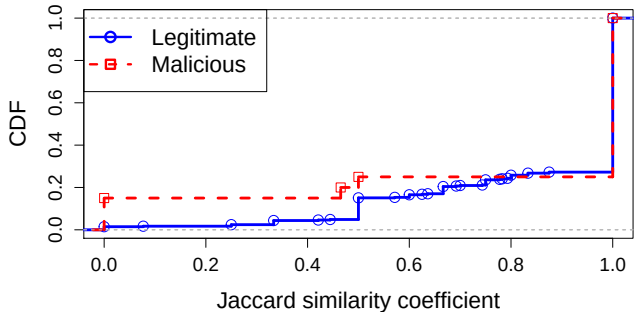


Figure 4: Malicious ASes rewire more frequently. *Distribution of the 5th percentile of the Jaccard similarity coefficient between consecutive snapshots of an AS's upstream providers. Higher values indicate fewer changes in upstream connectivity.*

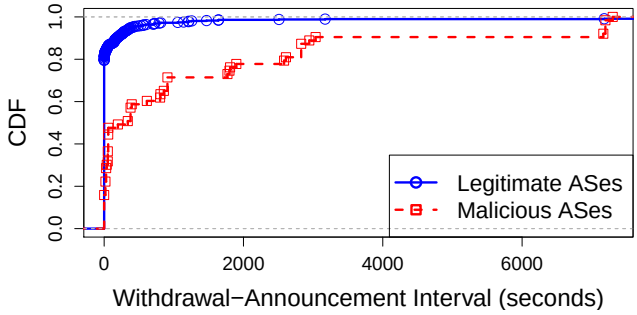


Figure 5: Malicious ASes withdraw prefixes for longer periods. *The distribution of the median interval between a prefix withdrawal and re-announcement across 15 contiguous epochs.*

3.2.3 BGP Routing Dynamics

These features attempt to capture abnormal BGP announcement and withdrawal patterns. For example, to support aggressive IP address space fragmentation and churn and avoid easy blacklisting, malicious ASes may periodically announce certain prefixes for short periods of time. On the contrary, the pattern of BGP announcements and withdrawals for legitimate ASes is mainly driven by normal network operations (*e.g.*, traffic load balancing, local policy changes), and should thus exhibit BGP routing dynamics that are different to those of malicious ASes.

Prefix reachability. We aim to capture the fraction of time that prefixes advertised by A remain reachable, which we define as *reachability*. First, we measure the time that elapses between an announcement and a withdrawal for every advertised prefix. Given the distribution of these time intervals, we extract the distribution characteristics as described above. Second, we track the time for a prefix to become reachable again after a withdrawal. Third, we measure the inter-arrival time (IAT) between withdrawals, for each of the prefixes that A announces, and compute the IAT distribution.

As before, we extract the distribution characteristics for each of the three distributions, yielding a total of nine features. Figure 5 shows the CDF of the median reachability value for the malicious and the legitimate ASes over the course of one day, and over 15 days. Higher values of this feature suggest that malicious ASes tend to re-advertise their prefixes after longer delays.

Topology and policy changes. We track the *topology and policy changes*, defined as in Li *et al.* [24], that are associated with each prefix. We define a *policy change* as follows: after a path to a destination is announced, a second BGP announcement is observed with the same AS path and next-hop, yet one or more of the other attributes (such as MED or community) is different. Similarly, we define a *topology change* event as follows: after a path to a destination is announced, a second announcement follows with an alternate route (implicit withdrawal) or after a route to a destination is explicitly withdrawn, a different route (with different AS path or next-hop attributes) to the same destination is announced (explicit withdrawal).

To capture and summarize the topology and policy changes per AS, we group the prefixes per origin AS (the origin AS appears as the last AS in the AS path). We track the policy change events for each prefix, and we measure the inter-arrival time between the events per prefix. Then, we analyze the collection of inter-arrival times of the policy events for all prefixes advertised by the same AS. For each AS, we form the distribution of such intervals, and we extract the distribution characteristics as described above. We also compute the total number of events and the total number of events divided by the total prefixes advertised by the AS. We repeat this process for the topology change events. We compute a total of ten features.

3.3 System Operation

We now describe *ASwatch*'s training and operation.

3.3.1 Training Phase

To train the classifier (Steps 6 and 7 in Figure 3), we first prepare a training dataset with labeled feature vectors related to known malicious and legitimate ASes. We start with a ground truth dataset that includes confirmed cases of *malicious* ASes, and legitimate ASes (described in more details is Section 4.1).

We compute the statistical features for each labeled AS using two sources of data: BGP announcements and withdrawals from Routeviews [32], and information from a publicly available dataset [25] about the relationships between ASes. We compute the feature vectors over m contiguous epochs (in our experiments, each epoch is one day). More specifically, we maintain a sliding window of size m epochs, which advances one epoch at a time. Using this sliding window, we can compute multiple feature vectors for each AS (one per window). Then, we associate a label to each feature vector, according to the ground truth related to the AS from which a vector was computed.

Finally, to build the statistical classifier, we use the Random Forest (RF) algorithm. We experimented with different algorithms, but we chose RF because it can be trained efficiently and has been shown to perform competitively with respect to other algorithms for a variety of problems [6].

3.3.2 Operational Phase

Once the statistical classifier has been trained, *ASwatch* can assign a reputation score to new ASes (*i.e.*, ASes for which no ground truth is yet available). *ASwatch* computes a reputation score for each new AS observed in the BGP messages from Routeviews. Suppose that we want to compute the reputation of an AS, A, over some time period, T. First, we compute A's features (as explained in Section 3.2) over period T, using a sliding window procedure as in the training phase. Namely, a feature vector is computed for each window within T. Second, we classify an AS as malicious, if *ASwatch* consistently assigns it a low reputation score for several days in a row.

More specifically, let T_i be the current day of observations, f_{A,T_i} be the corresponding feature vector for A, and $s(f_{A,T_i})$ be the *bad reputation* score output by the classifier at the end of T_i. Also let $W_i = (T_i, T_{i+1}, \ldots, T_{(i+m-1)})$ be a period of m consecutive days. We report A as malicious if: (a) score $s(f_{A,T_i}) > \theta$ for 90% of the days in period W_i, where θ is a predefined threshold that can be learned during the training period; and (b) condition (a) holds for at least l consecutive periods $W_i, W_{i+1}, \ldots, W_{i+l}$.

We note that we have experimented with multiple values for m and l (see Section 4.3 for detailed discussion on parameter selection).

4. EVALUATION

We now describe the data we collected and the setup for our evaluation of *ASwatch*, where we evaluate the system's accuracy. Our results show that *ASwatch* achieves a high detection rate for a reasonably low false positive rate, can detect malicious ASes before they are publicly reported by others, and can complement existing AS reputation systems that rely solely on data-plane observations. Furthermore, we find that *ASwatch* detects nearly double the fraction of confirmed cases of malicious ASes compared to BGP Ranking, a data-plane based AS reputation system.

4.1 Data

Labeling malicious ASes. Collecting reliable ground truth about malicious ASes is extremely challenging, due to the utter lack of public information available about such cases. Nonetheless, through extensive manual search and review efforts, we managed to collect a set of ASes for which there exists publicly available evidence of malicious behavior. For example, we identified a reasonable set of malicious ASes that were at some point seized by law enforcement or disconnected by other network operators.

To obtain our dataset of malicious ASes, we searched through websites that are operated by cyber-security professionals (*e.g.*, www.abuse.ch, blogs.rsa.com [1, 2, 10, 13, 16,

Informex, AS20564	Infium, AS40965	Vpnme, AS51354
Ecatel, AS29073	Egis, AS40989	Lyahov, AS51554
Volgahost, AS29106	K2KContel, AS43181	Taba, AS8287
RapidSwitch, AS29131	Phorm, AS48214	Retn, AS9002
Riccom, AS29550	IT-Outsource, AS48280	Vesteh, AS47560
Naukanet, AS31445	Vlaf, AS48984	Prombuddetal, AS44107
PromiraNet, AS31478	Moviement, AS49073	Citygame, AS12604
Ys-IX, AS31506	Interactive-3D, AS49544	Bogonet, AS47821
Vakushan, AS34229	Vvpn, AS49934	Troyak, AS50215
Euroaccess, AS34305	Softnet, AS50073	Vishclub, AS50369
SunNetwork, AS38197	Onlinenet, AS50722	Gaxtranz/Info, AS29371
Vline, AS39150	Digernet, AS50818	Group3, AS50033
Realhosts, AS39458	Proxiez, AS50896	Smila, AS50390
UninetMd, AS39858	Gorby, AS51303	

Figure 6: *Malicious ASes we collected from blogs.*

23]) and carefully reviewed articles about ASes known to be operated by cyber-criminals.

We observed the following common characteristics across all articles and blog reports we considered: (1) the reported ASes hosted a variety of cyber-criminal activities (*e.g.*, botnet C&C hosting, malware domains, phishing), (2) several ASes were associated with each other, either directly (*e.g.*, customer-provider relationship) or indirectly (*e.g.*, they shared the same upstream provider), (3) the operators of these ASes were uncooperative and unresponsive (*e.g.*, would not respond to abuse complaints or attempts by other AS operators to communicate with them), (4) some ASes were prosecuted by law enforcement and taken down, (5) many of these disappeared only for a relatively short time before resurfacing. From each blog report, we extracted the ASes involved and the dates when they were active. Overall, we collected forty one known malicious ASes. We provide our list of ASes in Figure 6.

Labeling legitimate ASes. To collect a set of legitimate ASes, we proceeded as follows. Every day for one year, we collected the list of top one million domain names from `alexa.com`. For each of these domains, we calculated the average daily ranking; we selected the domain names that had an average ranking above 10,000. In other words, we selected only those domains that were consistently very popular. Finally, we mapped each domain name to its resolved IP addresses and mapped those IP addresses to the AS that hosted them. Overall, we collected a total of 389 ASes, which we label as *legitimate*.

Although we cannot be absolutely certain that our labeling of legitimate ASes contains no noise, we rely on two reasonable assumptions. First, we assume that websites that are consistently popular are unlikely to be offering malicious services. Intuitively, a malicious site that becomes highly popular would also have a high number of victims, and would rapidly attract attention for take-down. As a result, the site would be quickly blocked or taken down and would thus not remain consistently popular. Second, we assume that the administrators of the most popular websites are unlikely to host their services within malicious ASes. Intuitively, if they relied on malicious ASes, they would risk damaging their own reputation, not to mention extended downtimes if the hosting ASes were taken down due to abuse complaints.

Finally, to ensure that our set of legitimate ASes consists of ASes that are similar in size to the malicious ASes, we keep only those legitimate ASes that have no customers, or whose customers are all stub ASes.

AS rewiring and relationships data (CAIDA). To track how malicious ASes change their connectivity, we use a publicly available dataset that reports AS business relationships. The dataset reports one snapshot of the AS graph per month, from 1998 to 2013.

Luckie *et al.* [25] provide an AS graph built by inferring business relationships among ASes, based on AS customer cones. Although this dataset has its own limitations (see Section 5), it provides a reasonably accurate view of AS relationships, allowing us to estimate our rewiring features that we presented in Section 3.2.

BGP routing dynamics (Routeviews). To further capture the control-plane behavior of malicious and legitimate ASes, we monitored the BGP messages that originate from these ASes using the Routeviews dataset. We use this dataset to measure both the dynamics of BGP updates and the IP fragmentation and churn features.

4.2 Experiment Setup

In the following section, we describe the training and the evaluation of our system. The training period extends from January 2010 to March 2010, while the evaluation experiments extend from January 2011 to December 2013.

Computing AS feature vectors. Given a period of time (*i.e.*, m contiguous epochs) over which we want to capture the behavior of an AS, we construct the AS feature vector as follows: (1) *Rewiring activity*: We compute the rewiring features over the most recent k snapshots of the AS relationships dataset, prior to the period of interest. Our source of AS relationships provides only one snapshot per month. Given this limitation, we select a reasonable number of snapshots to capture the most recent rewiring activity of an AS. For our experiments we set $k = 4$ (see Section 4.3 on parameter selection); (2) *BGP routing activity*: To compute BGP routing dynamics features, IP address space fragmentation and churn, we collect the BGP announcements and withdrawals originating from the AS during the period of interest. We note that BGP Routeviews offers a large number of monitors. Our pilot experiments over a number of different monitors indicated that changing the monitor selection did not significantly affect the overall performance of our classifier. Therefore, to compute our routing activity features, we select one monitor and consistently use it, throughout all the experiments.

Training the AS reputation model. Because our data is derived from cases of malicious ASes publicly reported by others, we rely on the report dates for an approximate period of time when the ASes were likely to be actively used by the attackers. For example, if an AS was reported as malicious on a given day d, we assume the AS was operated by criminals for at least a few months before d (in fact, it typically takes time for security operators to detect, track, confirm, and take down a malicious AS). For the purpose of computing our labeled feature vectors and training our system, we

selected a period of time with the highest concentration of active malicious ASes. This period extends from January–March 2010, during which we identified a total of 15 active malicious ASes. Even though this period may appear somewhat dated, it allows us to capture the agile behavior of several known malicious ASes within one consistent time frame, enabling a "clean" evaluation setup. Our evaluation detects a large fraction of malicious ASes that we have observed over a longer, more recent time period (2011–2013). In the future, we plan to investigate more sources of ground truth and identify additional periods of time that can be used to train our model (see Section 5 for further discussion).

Performing cross-validation tests. During the three-month training period mentioned above, we maintain a sliding window of fifteen contiguous days (epochs), sliding the window one day at a time (*i.e.*, two consecutive windows overlap by 14 days). For each sliding window, we compute the feature vector for each AS and we perform three-fold cross-validation as follows: (1) We separate the ASes into three subsets, using two subsets to train our reputation model, and one for testing. (2) For each training subset, we balance the two classes by oversampling from the underrepresented class. After balancing, the number of feature vectors of the two classes are equal. (3) We train the model using a Random Forest classifier [6]. (4) Finally, we test all feature vectors that belong to the third fold against the model, as we described in Section 3.3. Cross-validation yields the scores from the testing phase and the true label for each AS feature vector. We plot the receiver operating characteristic (ROC), which illustrates the performance of the classifier for different values of the detection threshold. Because we perform our testing once for each sliding window, we plot a similar ROC for each sliding window. The results are reported in Section 4.3.

Evaluating ASwatch across a nearly three-year period. After the cross-validation experiments, we use our model to test new ASes whose BGP behavior was observed outside the training period over nearly three years, from 2011 to 2013. We perform this evaluation for two reasons: a) to test how well *ASwatch* performs to detect new malicious ASes (outside of the training period), and b) to compare the performance of *ASwatch* with other AS reputation systems (*e.g.*, BGP Ranking) over an extended period of time. For each (previously unseen) AS we want to test against *ASwatch*, we classify it as malicious if it has *multiple* feature vectors that are *consistently* assigned a "bad reputation" score (see Section 3.3). The results are reported in Section 4.3.

4.3 Results

How accurate is ASwatch? Evaluation with cross-validation: Figure 7 shows the detection and false positive rates for one cross-validation run. The detection rate and false positives reported on the ROC correspond to the fraction of *malicious* feature vectors that are correctly classified and *legitimate* feature vectors that are incorrectly classified, respectively. As shown by the ROC curve, *ASwatch* can achieve a detection rate of 93.33% (correctly classifying 14 out of 15 ASes as malicious), with a reasonably low false positive rate of 5.25% (20 falsely detected ASes). In practice, we believe this false positive rate is manageable, as it represents 20 falsely detected ASes over a three-month period, or one every few days. Although this false positive rate is clearly too high to automate critical decisions such as take-down efforts, *ASwatch* can still be used to significantly narrow down the set of ASes for further investigation considerably, and can thus help both law enforcement focus their investigation efforts, and network administrators make decisions on who to peer with or which abuse complaints to prioritize.

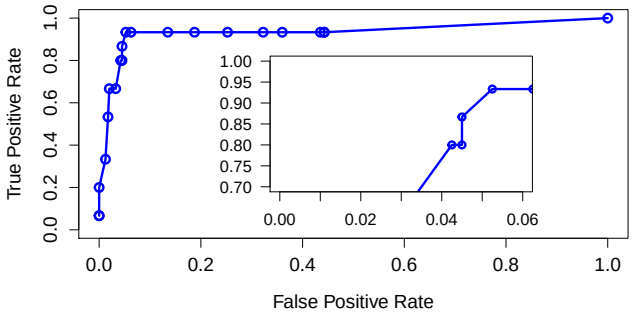
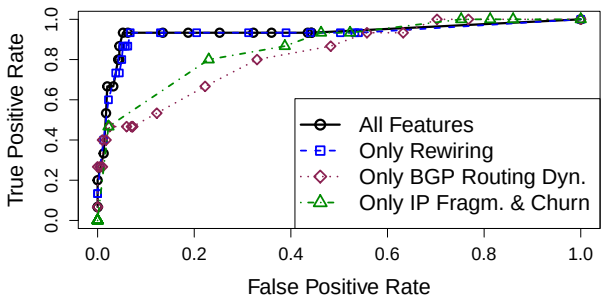


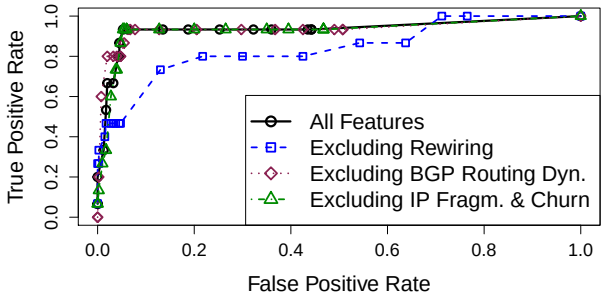
Figure 7: *The cross-validation detection and false positive rates of ASwatch.*

Evaluation outside the training period, over nearly three years: As described in Section 4.1, we use our model to test new ASes observed after the training period, over nearly three years, from 2011 to 2013. It is important to notice that, from a control-plane point of view, malicious ASes may not always be behaving maliciously across a three year period of time. Our ground truth information does not allow us to distinguish between the periods of activity and periods of "dormancy". Nonetheless, over time an AS operated by cyber-criminals will likely behave in a noticeably different way, compared to legitimate ASes, allowing us to detect it. Figure 10 shows the *cumulative* true positive rate of detected ASes over the testing period. At the end of this nearly three years period, *ASwatch* reached a true positive rate of 72% (21 out of 29 ASes correctly flagged as malicious).

To compute the false positives, for each month we count the number of distinct ASes that were detected as malicious. The false positives reach at most ten to fifteen ASes per month, which we believe is a manageable number, because these cases can be further reviewed by network operators and law enforcement. For instance, the upstream providers of an AS that is flagged as malicious by *ASwatch* may take a closer look at its customer's activities and time-to-response for abuse complaints. Furthermore, the output of *ASwatch* could be combined with the reputation score assigned by existing data-plane based AS reputation systems. The intuition is that if an AS behaves maliciously both at the control plane (as detected by *ASwatch*) and at the data plane (as detected by existing reputation systems), it is more likely that the AS is in fact operated by cyber-criminals.



(a) *Considering each feature family separately.*



(b) *Excluding one feature family at a time.*

Figure 8: Relative importance of different types of features. *The rewiring features contribute the most to the overall detection rate; other features contribute to lower false positive rates.*

How early can ASwatch detect malicious ASes before they are widely noticed? We want to evaluate if *ASwatch* can detect malicious ASes *before* they were reported by blog articles. For each of the 14 malicious ASes that *ASwatch* detected during the cross-validation experiments discussed earlier, we took note of the day that *ASwatch* first detected the malicious AS, and we measured the number of days between the time *ASwatch* detected the AS and the day the blog story was published. About 85% of the detected malicious ASes were detected by *ASwatch* 50 to 60 days before their story became public.

Which features are the most important? We evaluate the strength of each family of features that *ASwatch* uses. To understand which features are most important for *ASwatch*, we evaluate each family's contribution to the overall true and false positive rates. In particular, we want to study the effect of each family of features on the detection of malicious ASes, independently from the other families, and the effect of each family on the false positives when those features are excluded. To this end, we repeated the experiment described previously by excluding one family of features at a time. We repeated the experiment four times, once for each family of features, and we calculated the overall detection and false positive rates. Figure 8 shows the results of our experiments, which suggest that the rewiring features are very important, because excluding them significantly lowers the detection rate. The BGP dynamics and IP address space churn and fragmentation features help reduce the false positives slightly (the "Only Rewiring" ROC in Figure 8a is slightly shifted to the right). We followed a similar procedure to identify which features are most important for each family of features. Table 1 shows the most important features for each family.

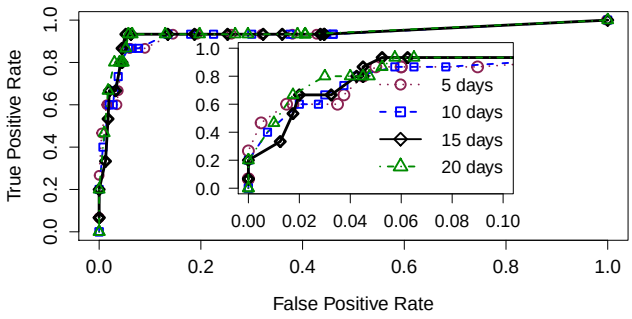


Figure 9: *The detection and false positive rates for ASwatch, if we vary the size of the sliding window. Our experiments show that the performance is not greatly affected.*

Is ASwatch sensitive to parameter tuning? As explained in Sections 3.3.2, 4.2 we use the following parameters to classify an AS as malicious: (1) *feature window size*: we compute feature vectors for an AS for a window of m consecutive days (one feature vector per day), and we repeat the feature computation over l consecutive sliding windows of size m. (2) *number of most recent snapshots of AS relationships*: we compute the rewiring features for an AS over the k most recent snapshots.

To tune our parameters, we performed several pilot experiments, rather than an exhaustive search over the entire parameter space. Our pilot experiments showed that *ASwatch*'s performance is robust to both parameters m and l. Due to space limitations, we only show our experiments for the parameter m. Figure 9 shows the performance for window sizes of 5, 10, 15, and 20 days. Our results show that the accuracy of *ASwatch* is not overly sensitive to the choice of window size m. The ROC plots in Figure 9 show that $m = 15$ gives a higher true positive rate with a reasonable false positive rate. We therefore set $m = 15$. Using a similar approach, we set $l = 5$. We classify an AS as malicious, if it scores lower than the detection threshold over five consecutive periods of 15 days.

After we have selected parameters m and l, we proceed to set parameter k. Suppose that we want to compute the reputation of an AS A, over period T. Then, parameter k is the number of most recent AS relationship snapshots, prior to T, over which we compute the rewiring features for A (notice that our AS relationships dataset consists of one snapshot per month, as mentioned in Section 4.1). In other words, k denotes "how much" history we consider, to capture the rewiring behavior for A. Ideally, we want to accurately capture A's rewiring behavior while using a small number of snapshots. We performed experiments using different values of k (*i.e.*, 1, 2, 3, 4). We then selected $k = 4$, because further increasing its value did not produce a significant increase in classification accuracy.

4.4 Comparison to BGP Ranking

We now compare *ASwatch* with BGP Ranking. In contrast to *ASwatch*, BGP Ranking is an AS reputation system based on *data-plane* features (*e.g.*, observations of attack traffic enabled by machines hosted within an AS). Clearly, BGP Ranking is an AS reputation system that is designed differently from *ASwatch*, because it aims to report ASes that are most heavily abused by cyber-criminals, but not necessarily operated by cyber-criminals. We compare the two systems for two reasons: (1) to test how many of the malicious ASes that are operated by cyber-criminals show enough data-plane evidence of maliciousness and get detected by existing data-plane based AS reputation systems; and (2) to evaluate whether the control-plane based approach can effectively complement data-plane based AS reputation systems.

Results summary. We found that *ASwatch* detected 72% of our set of malicious ASes over a three year period, and BGP Ranking detected about 34%. Both systems reported the same rate of false positives (on average 2.5% per month, which is ten to fifteen ASes per month). Combining the two systems we were able to detect only 14% of the malicious ASes, but we were able to reduce the false positives to 0.08% per month (12 ASes in total across the three year period).

BGP Ranking reports. BGP Ranking [5] has been making its AS reputation scores publicly available since 2011, along with a description of the approach used to compute the scores. BGP Ranking currently has information for a total of 14k ASes, and they announce a daily list of the worst 100 ASes by reputation score. The BGP Ranking score has a minimum value of 1 (which indicates that the AS hosts benign activity) but no maximum value (the more malicious traffic hosted by the AS, the higher the score).

Using our list of confirmed cases of malicious ASes (Section 4.1), we checked which ASes are visible from BGP Routeviews starting from 2011. We found a total of 29 ASes. We chose to check which ASes are active since January 2011, because this is the oldest date for which BGP Ranking has data available. Then, we tracked these ASes until November 2013, because the historic AS relationships dataset from CAIDA has a gap from November 2013 to August 2014. Therefore, we collected the historical scores for each active known malicious AS from BGP Ranking, from January 2011 until the end of 2013.

ASwatch setup. Using *ASwatch*, we generate the feature vectors for each AS in our list, starting from January 2011 until November 2013. To generate the feature vectors, we follow the same procedure as described in Section 3.3.2. We train *ASwatch* as previously described (on training data collected in 2010) and test the ASes observed from 2011 to 2013 against the model.

Comparing BGP Ranking with ASwatch. As mentioned earlier, BGP Ranking is not a detection system *per se*, in that it aims to report ASes that host a high concentration of malicious activities, and does not focus on distinguishing between abused ASes and ASes that are instead owned and operated by cyber-criminals. Nonetheless, for the sake of comparison it is possible to obtain a detection system by setting a threshold on the score output by BGP Ranking. BGP Ranking publishes the set of "worst" 100 ASes and their scores, which are updated daily (to obtain the historic scores for any other non-top-100 AS, one has to make explicit queries through the web portal). It also reports the average AS score per country or region, and ranks the countries that host the ASes with the lowest reputation. The four top ("worst") countries are Russia, Ukraine, Hong Kong, and the US. Using the above information we consider five distinct detection thresholds as follows: (1) average score for ASes in Russia (BGP Ranking Russia cut-off), (2) average score for ASes in Ukraine (BGP Ranking Ukraine cut-off), (3) average score for Hong Kong (BGP Ranking Hong Kong cut-off), and (4) average score for ASes in the US (BGP Ranking US cut-off). We also set a threshold based on the average score of the 100th worst AS (BGP Ranking top 100) collected from the daily reports. Figure 10 shows the detection results using these thresholds.

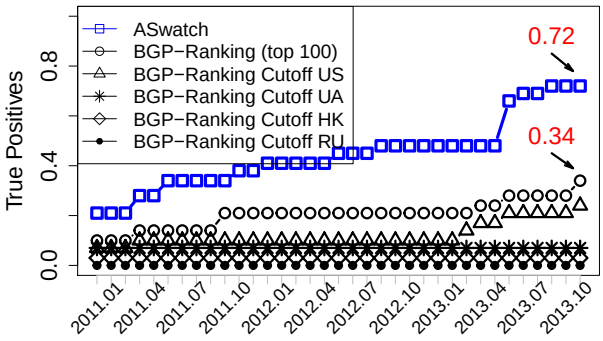


Figure 10: *True positive rates for ASwatch and BGP Ranking. Accumulation of detected ASes over nearly three years.*

We then compared BGP Ranking's detection with that of *ASwatch*. Figure 10 shows the fraction of ASes that *ASwatch* and BGP Ranking detected. We show the cumulative fraction of detected ASes, from January 2011 to November 2013. At the end of the 35-month period, *ASwatch* detected about 72% of the set of ASes we tracked, while BGP Ranking detected about 34%. *We found that 72% of the malicious ASes were detected by monitoring their control-plane behavior, but only 34% of the malicious ASes showed enough data-plane activity to be detected by BGP Ranking.* BGP Ranking may have only limited visibility of malicious activities in the data plane across the entire Internet, and thus may completely miss the malicious activities of certain ASes. Naturally, it is challenging to deploy a large number of sensors dedicated to detecting malicious network communications over the entire Internet. On the other hand, *ASwatch* monitors BGP behavior, and may therefore compensate the limited visibility of data-plane based approaches.

We also compared the false positive rates of BGP Ranking and *ASwatch*. Our motivation is to see if the false positives are manageable within a reasonable period of time

(*e.g.*one month). We collected the *ASwatch* scores and the BGP Ranking scores for our set of legitimate ASes (see Section 4.1). For each system, we counted the number of legitimate ASes that *ASwatch* detected per month. We found that both systems produce only ten to fifteen false positives per month on average over the total of 389 known legitimate ASes in our dataset. As we have mentioned earlier, BGP Ranking is designed differently from *ASwatch*. Although the rate we calculated does not represent the actual false positive rate for BGP ranking, it does provide an estimate of the false positive that an operator would need to deal with, if BGP Ranking were used to detect malicious ASes.

Combining control-plane with data-plane. Finally, we evaluated how the two systems would perform if we used them together. To this end, we label an AS as malicious if it was reported by both systems, with each two report dates to be at most six months apart from each other. For BGP Ranking we used the BGP Ranking top 100 threshold. We found that combining the two systems, we were able to detect 14% of our malicious ASes. This means that of 14% of the known malicious ASes exhibited both control plane and data plane malicious behavior within six months. The fraction of legitimate ASes that both systems detected as malicious is only 3% (*i.e.*, 12 ASes out of 389) for the whole three year period (which is on average 0.08% per month). Finally, five out of the 29 known malicious ASes that were active in the three year observation period were missed by both systems. For example, AS 49544 (Interactive 3D) and AS 39858 (UninetMd, now Comstar Volga Arzamas) are among the top worst ASes that both systems detected.

5. DISCUSSION

ASwatch reputation scores in practice. *ASwatch* may help the work of network operators and security practitioners as follows: (1) *Prioritize traffic*: knowing what ASes have suspicious (low reputation) control-plane behavior may help administrators to appropriately handle traffic originating from such ASes; (2) *Peering decisions*: Upstream providers could use AS reputation scores as an additional source of information to make peering decisions, for example by charging higher costs to compensate for the risk of having a low reputation customer or even de-peer early if reputation scores drop significantly; (3) *Prioritize investigations*: law enforcement and security practitioners may prioritize their investigations and start early monitoring on low reputation ASes; (4) *Complement data-plane based systems*: *ASwatch* could be used in combination with data-plane based reputation systems, so that ASes that exhibit malicious behavior both from the control-and data-plane points of view can be prioritized first; (5) *Strengthen existing defenses*: furthermore, reputation could be used as input to other network defenses (*e.g.*, spam filters, botnet detection systems) to improve their detection accuracy.

Working with limited ground truth. We briefly summarize the challenges that we faced due to limited ground truth, and how we mitigated them. (1) *Highly unbalanced dataset*: The ratio of malicious ASes to legitimate ASes produced a highly unbalanced dataset. Before training we used well-known data mining approaches to balance the dataset, by oversampling the underrepresented class of malicious ASes (Section 4.1). (2) *Limited time period for training*: We relied on the date of the ground truth reports to estimate the period of time in which the ASes were likely to be actively used by the attackers. We were not able to obtain additional information about the activity periods (or dormancy periods) outside the report dates. Therefore, we designed AS *ASwatch* so that it does not make a final decision for an AS based on a single observation (*i.e.*, a single feature vector). Instead, we introduced parameters to ensure that we label an AS as malicious only if it is assigned consistently low scores for an extended period of time. (3) *Model update with adaptive training*: Because of the lack of information on the activity periods (or dormancy periods) outside the report dates, we were not able to periodically update our model. Therefore, we performed a one-time training on our model using a period of time (January–March 2010) for which we had "clean" data. Even though *ASwatch* uses observations of cases of malicious ASes in 2010, we believe that it effectively models fundamental characteristics of malicious ASes that are still reflected on today's cases. This belief is supported in part by the results of correlating *ASwatch*'s output with recent BGP Ranking reports (see Section 4). In our future work, we plan to investigate more sources of ground truth and identify other periods of time that could be included in our training.

Limitations of the AS relationships dataset. To measure our rewiring features, we relied on a dataset that provides snapshots of AS relationships over years (see Section 4.1). The relationship inference algorithm is based on the idea of customer cones—the set of ASes an AS can reach through its customer links. This dataset has its own set of limitations. For example, each pair of ASes is assigned only a single relationship, and visibility is limited to the monitoring points publicly available via Routeviews. It is possible that some business relationships may be missing, or that some false relationships are reported. Moreover, since the dataset is provided in snapshots (one snapshot per month), we are not able to observe rewiring activity that may be happening at a finer time scales. Nevertheless, this AS relationships dataset has the largest validated collection of AS relationships gathered to date, with about 44,000 (34.6%) of the inferences validated, and it reports the AS relationships over years, which allowed us to track our ground truth ASes over an extended period of time.

Evasion. Naturally, as for any other detection system, *ASwatch* may face the challenge of sophisticated attackers who attempt to evade it. For example, an attacker may attempt to manage her AS to mimic the BGP behavior of legitimate ASes. However, we should notice that *ASwatch* relies heavily on rewiring features, which capture how an AS connects with other ASes, including upstream providers. Mimicking legitimate behavior to evade *ASwatch* would mean that the malicious AS has to become "less agile". In turn, being less agile may expose the AS to de-peering by its up-

stream providers as a consequence of accumulating abuse complaints. For example, if McColo (which was taken down in 2008) had not changed ten upstream providers before it was taken down, it might have been taken down much sooner.

Future work. We plan to expand our set of features to capture other types of behavior, such as making peering arrangements for specific prefixes. We intend to expand our sources of bullet-proof hosting ASes, so that we test *ASwatch* over larger datasets and longer periods of time. We also plan to explore how we may combine our set of control plane features with data plane features.

6. RELATED WORK

We review studies of "unclean" ASes and existing AS reputation systems, as well as applications of machine learning and signal processing to detect BGP anomalies.

Studies of "unclean" ASes. Previous studies have attempted to identify "unclean" ASes, which are ASes with a high concentration of low reputation IP addresses. In contrast, we attempt to understand the behavior of ASes that are *controlled and managed by attackers*, rather than ASes which are heavily abused. Collins [9] first introduced the term network uncleanliness as an indicator of the tendency for hosts in a network to become compromised. They gathered IP addresses from datasets of botnets, scan, phishing, and spam attacks to study spatial and temporal properties of network uncleanliness; this work found that compromised hosts tend to cluster within unclean networks. Kalafut *et al.* [18] collected data from popular blacklists, spam data, and DNS domain resolutions. They found that a small fraction of ASes have over 80% of their routable IP address space blacklisted. Konte *et al.* [20] studied ASes that are reported by Hostexploit and how they changed their upstream connectivity. Johnson *et al.* introduced metrics for measuring ISP badness [17]. Moura *et al.* studied Internet bad neighborhoods aggregation. Earlier papers have looked into IP addresses that host scam websites or part of spamming botnets are organized intro infrastructures [8, 12, 38]. Finally, Ramachandran *et al.* found that most spam originates from a relatively small number of ASes, and also quantified the extent to which spammers use short-lived BGP announcements to send spam [29, 30]. These studies suggest that it is possible to develop an AS reputation system based on analysis of control-plane features, which is the focus of our work.

AS reputation systems. The state of the art in AS reputation systems is to use features that are derived from data-plane information, such as statistics of attack traffic. Current systems correlate data from multiple sources such as spam, malware, malicious URLs, spam bots, botnet C&C servers, phishing servers, exploit servers, cyber-warfare provided by other organizations or companies. Then, then rank ASes based on the concentration of low reputations IP addresses. Organizations, such as Hostexploit [34], Sitevet [34], and BGP Ranking [4] rate each AS with an index based on the activity of the AS weighted by the size of its allocated address space. FIRE [36] examines datasets of IRC-based botnets, HTdetection-based botnets, drive-by-download and phishing hosts and scores ASes based on the longevity of the malicious services they host and the concentration of bad IP addresses that are actively involved. ASMATRA [37] attempts to detecting ASes that provide upstream connectivity for malicious ASes, without being malicious themselves.

Zhang *et al.* [39] find that there is a correlation between networks that are mismanaged and networks that are responsible for malicious activities. The authors use a mismanagement metric to indicate which ASes may be likely to exhibit malicious behaviors (e.g. spam, malware infections), which does not necessarily indicate if an AS is actually operated by cyber-criminals or not. In contrast, we focus on detecting ASes that are operated by attackers, rather than ASes that are mismanagement and likely abused. Also, [39] examined short-lived BGP announcements as an indication of BGP misconfigurations. Even though we also examine the duration of prefix announcements, this is only one of the features we use to capture control plane behavior. Our analysis shows that this feature alone is not enough to distinguish between legitimate and malicious ASes.

Roveta *et al.* [33] developed BURN, a visualization tool, that displays ASes with malicious activity, with the purpose to identify misbehaving networks. In contrast to these reputation systems that rely on data-plane observations of malicious activity from privileged vantage points, *ASwatch* establishes AS reputation using control-plane (*i.e.*, routing) features that can be observed without privileged vantage points and often before an attack.

Machine learning and signal processing approaches. These approaches detect BGP anomalies (*e.g.*, burstiness), with the goal to help system administrators diagnose problematic network behaviors, but they do not provide a connection between BGP anomalies and criminal activity. In contrast to these approaches, *ASwatch* attempts to capture suspicious control-plane behavior (*e.g.*, aggressive change of connectivity, short BGP announcements) with the goal to detect malicious ASes. Prakash *et al.* developed BGPlens, which monitors anomalies by observing statistical anomalies in BGP updates based on analysis of several features, including self-similarity, power-law, and lognormal marginals [28]. Similarly, Mai [26], Zhang [40] and Al-Rousan [3] have examined BGP update messages using tools based on self-similarity and wavelets analysis hidden Markov models to design anomaly detection mechanisms.

7. CONCLUSION

This paper presented *ASwatch*, the first system to derive AS reputation based on control-plane behavior. *ASwatch* is based on the intuition that malicious ASes exhibit "agile" control-plane behavior (*e.g.*, short-lived routes, aggressive rewiring). We evaluated *ASwatch* on known malicious ASes and found that it detected 93% of malicious ASes with a 5% false positive rate. When comparing to BGP Ranking, the current state-of-the-art AS reputation system, we found that *ASwatch* detected 72% of reported malicious ASes, whereas

BGP ranking detected only 34%. These results suggest that *ASwatch* can better help network operators and law enforcement take swifter action against these ASes that continue to remain sources of malicious activities. Possible remediations could be assessing the risk of peering with a particular AS, prioritizing investigations, and complementing existing defenses that incorporate other datasets.

Acknowledgments

We thank our shepherd, Walter Willinger, and the anonymous reviewers for their helpful comments and guidance. This material is supported in part by National Science Foundation awards CNS-1149051, CNS-1539923, and CNS-1531281.

REFERENCES

[1] abuse.ch. And Another Bulletproof Hosting AS Goes Offline, Mar. 2010. http://www.abuse.ch/?p=2496.

[2] abuse.ch. 2011: A Bad Start For Cybercriminals: 14 Rogue ISPs Disconnected. http://www.abuse.ch/?tag=vline-telecom, Jan. 2011.

[3] N. M. Al-Rousan and L. Trajkovic. Machine learning models for classification of BGP anomalies. In *High Performance Switching and Routing (HPSR)*, pages 103–108, 2012.

[4] BGP Ranking. http://bgpranking.circl.lu/.

[5] Bgp ranking reports. http://bgpranking.circl.lu/.

[6] L. Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001.

[7] Bulletproof Hosting. http://en.wikipedia.org/wiki/Bulletproof_hosting. Wikipedia.

[8] K. Chiang and L. Lloyd. A case study of the Rustock rootkit and spam bot. In *Workshop on Understanding Botnets*, 2007.

[9] M. P. Collins, T. J. Shimeall, S. Faber, J. Janies, R. Weaver, M. De Shon, and J. Kadane. Using uncleanliness to predict future botnet addresses. In *ACM SIGCOMM Internet Measurement Conference*, pages 93–104, 2007.

[10] Criminal service providers. http://cyberthreat.wordpress.com/category/criminal-service-providers/.

[11] DShield: Internet Storm Center - Internet Security. www.dshield.org/?

[12] F.Li and M. Hsieh. An empirical study of clustering behavior of spammers and group-based anti-spam strategies. In *Conference on Email and Anti-Spam (CEAS)*, 2006.

[13] Hostexploit. AS50896-PROXIEZ Overview of a crime server, May 2010. http://goo.gl/AYGKAQ.

[14] Hostexploit, June 2011. http://hostexploit.com/.

[15] Hostexploit. World Hosts Report. Technical report, Mar. 2014. http://hostexploit.com/downloads/summary/7-public-reports/52-world-hosts-report-march-2014.html.

[16] Crimeware-firendly ISPs. http://hphosts.blogspot.com/2010/02/crimeware-friendly-isps-cogent-psi.html.

[17] B. Johnson, J. Chuang, J. Grossklags, and N. Christin. Metrics for Measuring ISP Badness: The Case of Spam. In *Financial Cryptography and Data Security*, pages 89–97. Springer, 2012.

[18] A. J. Kalafut, C. A. Shue, and M. Gupta. Malicious Hubs: Detecting Abnormally Malicious Autonomous Systems. In *IEEE INFOCOM*, pages 1–5. IEEE, 2010.

[19] J. Kirk. ISP Cut Off from Internet After Security Concerns. http://www.pcworld.com/article/153734/mccolo_isp_security.html, Nov. 2008. PC World.

[20] M. Konte and N. Feamster. Re-wiring Activity of Malicious Networks. In *Passive and Active Measurement*, pages 116–125. Springer, 2012.

[21] B. Krebs. Russian Business Network: Down, But Not Out. http://goo.gl/6ITJwP, Nov. 2007. Washington Post.

[22] B. Krebs. Host of Internet Spam Groups Is Cut Off. http://goo.gl/8J5P89, Nov. 2008. Washington Post.

[23] B. Krebs. Dozens of ZeuS Botnets Knocked Offline, Mar. 2010. http://krebsonsecurity.com/2010/03/dozens-of-zeus-botnets-knocked-offline/.

[24] J. Li, M. Guidero, Z. Wu, E. Purpus, and T. Ehrenkranz. BGP Routing Dynamics Revisited. *ACM SIGCOMM Computer Communication Review*, 37(2):5–16, 2007.

[25] M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, et al. AS Relationships, Customer Cones, and Validation. In *Proceedings of ACM SIGCOMM Internet Measurement Conference*, pages 243–256. ACM, 2013.

[26] J. Mai, L. Yuan, and C.-N. Chuah. Detecting BGP anomalies with wavelet. In *IEEE Network Operations and Management Symposium*, pages 465–472. IEEE, 2008.

[27] R. McMillan. After Takedown, Botnet-linked ISP Troyak Resurfaces. http://goo.gl/5k0OV1, Mar. 2010. Computer World.

[28] B. A. Prakash, N. Valler, D. Andersen, M. Faloutsos, and C. Faloutsos. BGP-lens: Patterns and Anomalies in Internet Routing Updates. In *ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 1315–1324, 2009.

[29] A. Ramachandran and N. Feamster. Understanding the Network-Level Behavior of Spammers. In *ACM SIGCOMM*, 2006.

[30] A. Ramachandran, N. Feamster, and S. Vempala. Filtering Spam with Behavioral Blacklisting. In *ACM Conference on Computer and Communications Security (CCS)*, 2007.

[31] The Russian Business Network. http://en.wikipedia.org/wiki/Russian_Business_Network.

[32] The RouteViews Project. www.routeviews.org/.

[33] F. Roveta, G. Caviglia, L. Di Mario, S. Zanero, F. Maggi, and P. Ciuccarelli. Burn: Baring unknown rogue networks. In *International Symposium on Visualization for Cyber Security (VizSec)*, 2011.

[34] Sitevet. http://sitevet.com/.

[35] Spamhaus. www.spamhaus.org.

[36] B. Stone-Gross, C. Kruegel, K. Almeroth, A. Moser, and E. Kirda. FIRE: Finding rogue networks. In *IEEE Computer Security Applications Conference (ACSAC)*, pages 231–240. IEEE, 2009.

[37] C. Wagner, J. François, R. State, A. Dulaunoy, T. Engel, and G. Massen. ASMATRA: Ranking ASes providing transit service to malware hosters. In *IFIP/IEEE International Sympososium on Integrated Network Management*, pages 260–268, 2013.

[38] Y. Xie, F. Yu, K. Achan, R. Panigrahy, and G. Hulten. Spamming Botnets: Signatures and Characteristics. In *SIGCOMM*, 2008.

[39] J. Zhang, Z. Durumeric, M. Bailey, M. Karir, and M. Liu. On the Mismanagement and Maliciousness of Networks. In *Proceedings of the 21st Annual Network & Distributed System Security Symposium (NDSS '14)*, San Diego, California, USA, February 2013.

[40] J. Zhang, J. Rexford, and J. Feigenbaum. Learning-based anomaly detection in BGP updates. In *ACM SIGCOMM Workshop on Mining Network Data (MineNet)*, pages 219–220. ACM, 2005.