Quantifying Cloud Misbehavior

Rajat Tandon, Jelena Mirkovic, Pithayuth Charnsethikul
University of Southern California Information Sciences Institute
rajattan@usc.edu, sunshine@isi.edu, charnset@usc.edu

***Abstract*—Clouds have gained popularity over the years as they provide on-demand resources without associated long-term costs. Cloud users often gain superuser access to cloud machines, which is necessary to customize them to user needs. But superuser access to a vast amount of resources, without support or oversight of experienced system administrators, can create fertile ground for accidental or intentional misuse. Attackers can rent cloud machines or hijack them from cloud users, and leverage them to generate unwanted traffic, such as spam and phishing, denial of service, vulnerability scans, drive-by downloads, etc.**

In this paper, we analyze 13 datasets, containing various types of unwanted traffic, to quantify cloud misbehavior and identify clouds that most often and most aggressively generate unwanted traffic. We find that although clouds own only 5.4% of the routable IPv4 address space (with 94.6% going to non-clouds), they often generate similar amounts of scans as non-clouds, and contribute to 22–96% of entries on blocklists. Among /24 prefixes that send vulnerability scans, a cloud's /24 prefix is 20–100 times more aggressive than a non-cloud's. Among /24 prefixes whose addresses appear on blocklists, a cloud's /24 prefix is almost twice as likely to have its address listed, compared to a non-cloud's /24 prefix. Misbehavior is heavy-tailed among both clouds and non-clouds. There are 25 clouds that contribute 90% of all the cloud scans, and 10 clouds contribute more than 20% of blocklist entries from clouds.

I. Introduction

The cloud computing industry has seen a growth of about 200 billion U.S. dollars over the past decade [26]. In this paper we focus on quantifying misuse of cloud resources for malicious Interent activity (e.g., sending spam). Clouds may offer content hosting (e.g. Wordpress) or may rent virtual machines to customers, giving them the liberty to install and run their own software and upload their own data (e.g. Amazon AWS). We focus on the latter category of clouds, since the ability to install custom software requires superuser access to machines, and is often necessary for misbehavior. It may appear that, since clouds charge per usage, the monetary aspect may deter misuse of their resources. Sadly, this is not so. Criminals often avoid paying by taking advantage of free trials, or hijacking legitimate accounts for their own use through Account Takeover Attacks [30]. Some clouds also willingly host malicious content or condone misuse [4].

Cloud owners are aware that their machines can be misused, and many work hard to prevent or detect and stop misuse [1], [22], [24], [25]. However, these protections can be bypassed by attackers [43], [49]. There is anecdotal evidence that clouds generate unwanted traffic in select incidents. For example, in 2014, Amazon AWS EC2 machines were used to launch DDoS attacks on a large regional US bank and a Japanese electronics maker [11], [12]. Further, Zhao et. al [51] observe that cloud service providers are being preferred by cyber-criminals to inflict harm on online services at a large scale. Link11 [5] finds that cloud services from leading providers including Amazon AWS, Microsoft Azure and Alibaba were used in 25% of all DDoS attacks in Europe from July 2017 to June 2018. Recent statistics from Akamai [34] reveal that cloud service providers account for a significant amount of DDoS traffic. Akamai points to trends in cloud providers towards greater capacity, flexibility, and availability as also favorable for misuse. Glazier et. al [50] mention that cloud hosting providers are being used as proxy farms by attackers, because they can spin up multiple virtual instances with different IP addresses and launch widely distributed attacks in a short period of time. There are many other documented incidents, e.g., [32], [44]–[46].

978-1-7281-9486-8/20/$31.00 ©2020 IEEE

A. Contributions

Our paper is the first to systematically analyze and quantify Internet-wide cloud misbehavior. We rely on 13 diverse datasets: three datasets document unwanted traffic (vulnerability scans) and ten contain IP addresses or URLs blocklisted for misbehavior by multiple providers. We classify all Internet prefixes as cloud or non-cloud using a multi-pass approach and relying on reverse DNS data, IPinfo.io prefix classification and manual verification. We then attribute each misbehavior (scan or appearance on a blocklist) from our datasets to a cloud or non-cloud, and analyze differences between clouds and non-clouds.

We find that **although clouds own only 5.4% of the routable IPv4 address space** (with 94.6% going to non-clouds), **they generate between 45–55% of all scans in our datasets and contribute to 22–96% of blocklist entries**.

Commensurate with this, we find that **an average /24 cloud prefix scans at a rate that is 20–100 times higher than that of non-cloud prefixes**. Observing the /24 prefixes whose IP addresses get blocklisted, **cloud prefixes appear on blocklists almost twice the rate of non-cloud prefixes**.

Clouds also play a pivotal role in spreading malware and hosting phishing URLs. They represent 80-96% of 22K listed entries on phishing and malware blocklists.

There is **a heavy tail in both cloud and non-cloud misuse**. Only 25 clouds account for 90% of the cloud-sourced scans, and 10 clouds account for 20% of the cloud addresses listed in blocklists. DigitalOcean and OVH are two of the most misused clouds, with an average rank of 5.3 and 5.8 respectively across *all the 13 datasets we use*.

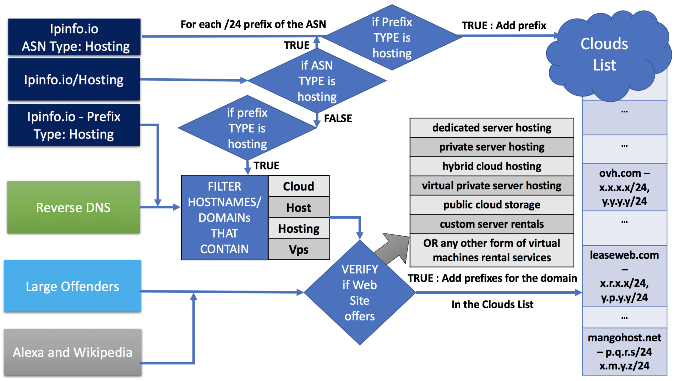

Ipinfo.io
ASN Type: Hosting
Ipinfo.io/Hosting
Ipinfo.io - Prefix Type: Hosting
Reverse DNS
Large Offenders
Alexa and Wikipedia
For each /24 prefix of the ASN
if Prefix TYPE is hosting
TRUE : Add prefix
Clouds List
TRUE
if ASN TYPE is hosting
FALSE
if prefix TYPE is hosting
TRUE
FILTER HOSTNAMES/ DOMAINs THAT CONTAIN
Cloud
Host
Hosting
Vps
dedicated server hosting
private server hosting
hybrid cloud hosting
virtual private server hosting
public cloud storage
custom server rentals
OR any other form of virtual machines rental services
VERIFY if Web Site offers
TRUE : Add prefixes for the domain
In the Clouds List
...
...
ovh.com – x.x.x.x/24, y.y.y.y/24
...
leaseweb.com – x.r.x.x/24, y.p.y.y/24
...
mangohost.net – p.q.r.s/24 x.m.y.z/24

Fig. 1. Identifying Clouds

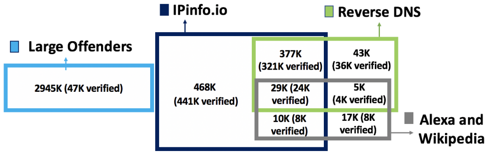


Fig. 2. Number of cloud prefixes identified from different sources and their overlaps.

II. Methodology

In this section we define the term “cloud” (Sec. II-A) and describe how we identify cloud prefixes on the Internet (Sec. II-B). We also discuss datasets we use to quantify misbehavior (Sec. III) and define how we measure maliciousness of a prefix or an organization (Sec. III-D).

A. Cloud Definition

We define a “cloud” as an organization that offers servers for rent, and allows users to install custom software on these servers. While there are other organizations that offer content hosting (e.g Wordpress), we focus on the server-based definition of a cloud, because the ability to install custom software is often necessary for misbehavior. For example, this is necessary for spoofed traffic generation, sending certain types of scans or hosting malware.

B. Identifying Clouds

Figure 1 outlines our methodology for identifying all /24 prefixes in the Internet that offer server hosting services, and are thus clouds by our definition. We focus on /24 prefixes and not entire organizations, because some large organizations (e.g., Microsoft, Amazon, Google) dedicate only a portion of their address space to cloud services. We also do not consider granularity smaller than a /24 prefix, because this is the smallest address space assigned to an organization by Regional Internet Registries.

We identify clouds through four methods: (1) reverse DNS lookup, (2) Alexa and Wikipedia, (3) IPinfo.io services and (4) Large offenders. Each of these methods provides us with a list of *cloud candidates*, and we follow it by our manual verification of each cloud candidate. We provide more details in the rest of this section.

Reverse DNS. We use reverse DNS to look up all routable IPv4 /24 prefixes from September 2019. To identify candidate cloud prefixes, we look for the keywords, “cloud”, “hosting”, “host” or ”vps” in the domain name part of the DNS names for hosts in the given prefix. If any of these terms are present, we proceed to manually verify if the given domain offers cloud services. At the end of the verification process, we find 385K /24 cloud prefixes using the reverse DNS approach. We manually verify 8 K organizations in 69 hours (out of which 3.1 K are clouds). Some of the clouds we identified through reverse DNS are *upcloud.com*, *dreamhost.com* and *greencloudvps.com*.

Alexa and Wikipedia. We identify cloud candidates from the Alexa listing of Top 500 hosting services [2], [23] and the Wikipedia’s page of well-known clouds [27]. We again manually verify each cloud to ensure they offer servers and not just content hosting. This step adds another 44 K /24 prefixes to our clouds list. We manually verify 636 organizations in 7 hours (out of which 208 are clouds).

IPinfo.io. The organization, IPinfo.io [36], collects information about IP addresses and prefixes. These details include information about the associated Autonomous System Number (ASN), AS type, AS domain, geolocation, prefix and prefix domain and type. AS type and prefix type can belong to one of the categories: “hosting”, “business”, “isp”, “education” or “inactive”. AS type and prefix type can be different. Thus, we rely on prefix type, as being more specific and thus likely more accurate. IPinfo.io provides information about all the IP prefixes owned by a given AS. We leverage IPinfo.io in two ways. First, we collect ASes that have type “hosting”, and include all their /24 prefixes, with prefix type “hosting”, in our list of clouds. For example, the AS26481 is associated with Rebel Hosting (*rebelhosting.net*) and has its ASN type “hosting”. So, we add all its /24 prefixes with prefix type “hosting” to our list of clouds. Second, we collect /24 prefixes whose type is “hosting”, while the AS type may be different. Such prefixes may correspond to internal hosting, or to hosting services offered to public. We keep only those prefixes whose domain name includes “cloud”, “hosting”, “host” or “vps”. We then manually verify these domains and add the confirmed prefixes to our cloud list. For example, the prefix 108.161.147.0/24 associated with *m5hosting.com* (M5 Hosting), and ASN AS395831, has AS type “business”, but has prefix type “hosting”. Its Web page confirms that it provides server hosting. We manually verify 1 K organizations in 9 hours (out of which 424 are clouds). Our use of IPinfo.io adds another 794 K /24 prefixes to our list of clouds. Using IPinfo.io, we find a total of 4 K clouds.

Large offenders. We collect top 20 organizations that contribute the most to scans and blocklist entries, for each dataset. We then manually verify if these organizations are clouds or non-clouds. If an organization is a cloud we add all its /24 prefixes to our list of clouds (e.g., *hinet.net* AS3462, provides multiple other services along with server hosting). That adds 47 K /24 prefixes to our list. In some cases the Web site may offer minimal information, such as a vague offer of services and a phone number (e.g., *ipvolume.net*). We will include such an organization on the cloud list if its tendency to participate in misbehavior is documented (e.g., by *badpackets.net* or by news articles.

Manual verification. Our manual verification consists of visit to the Web server of the given organization associated with a cloud candidate prefix, and examination of offered services. If the services include one of the following options, we conclude that the organization is a cloud: dedicated server hosting, private server hosting, hybrid cloud hosting, virtual private server hosting, custom server rental or virtual machine rental. We use Google Translate to translate foreign-language Web sites into English. If we conclude that the organization is a cloud, we verify all its prefixes on our candidate list.

In the end, our list of cloud prefixes includes 889 K unique /24 IP prefixes from 6 K unique organizations, which contribute to 5.4% of the publicly routable IPv4 address space [16]. The number of cloud prefixes identified from different sources and their overlaps are shown in Figure 2.

We attempted to quantify the accuracy of our overall cloud identification process by randomly selecting 100 organizations from our list of clouds, and visiting their Web pages for manual verification. Out of 100, 97 were indeed clouds and 3 were misclassified, containing 13,429 and 4 /24 prefixes respectively. We thus conclude that our cloud identification precision is 97% for organizations and 99.97% for prefixes.

Limitations. One limitation of our work is that there is no ground truth that we could use to evaluate accuracy of our cloud identification. Another limitation is that our manual verification may be subjective, since it is performed by a human, and thus it may in some cases be inaccurate. Further, if an organization offers both content and server hosting, and is not added through IPinfo.io source, we add all the hosting prefixes to our list as we cannot distinguish between them, based on the information we have. Yet another limitation is that we may miss some clouds, because our candidate identification process misses them. Since our manual verification of large offenders adds 5.33% of the total prefixes to our clouds list we estimate that our accuracy of candidate identification (without large offender dataset) is 94.66%. Moreover, as per IPinfo.io, 1%–5% of hosting provider allocations may change every month and IP address allocations across hosting providers can change frequently too. However, 9 out of 13 datasets that we analyze overlap with the time frame when we extract the prefixes from IPinfo.io, which contributes to 89.31% of our clouds list. Our list of cloud prefixes is available as open source [6], and we hope that other researchers can help improve its accuracy.

III. Datasets

We work with 13 diverse datasets, which document misbehavior. We group these datasets into two major categories (1) network traces, and (2) IP addresses or URLs listed on blocklists. Table I summarizes all our datasets.

A. Network Traces

CAIDA real-time network telescope data [19] captures all traffic to an unused /8 prefix owned by CAIDA. The traffic to these unassigned addresses (darknet) is unsolicited, and results from a wide range of events, including scanning (unwanted traffic looking for vulnerabilities) and backscatter (replies by the victims of DDoS attacks to randomly spoofed traffic, including darknet space). We analyze TCP SYN traffic sent to darknet, because such traffic indicates vulnerability scanning and it cannot be part of backscatter [48], [31]. We analyze traffic from March 12, 2020 to April 28, 2020. Each compressed hourly file is close to 80 GBs, and contains a few billion packets. The traffic is not anonymized. This dataset can only be accessed on CAIDA machines, and it is too large to be processed live in entirety. Instead, we analyze the first twenty minutes of traffic from each hour, and produce a list of top-10 destination ports. These ports are computed by analyzing 113 archived files, randomly sampled from December 1, 2019 to February 28, 2020. The port numbers are 22, 23, 80, 81, 443, 3389, 3390, 5222, 8545 and 8080. Our analysis then focuses only on scans sent to these select port numbers, which enables us to process CAIDA traces in real time, using Berkeley packet filters and parallel processes.

Merit network real-time network telescope data [17] also captures full packet traces, using a /13 dark prefix. We analyze their data from March 11, 2020 to March 19, 2020. Each compressed hourly file is close to 2 GBs, and usually contains fewer than 0.1 billion packets, so we can process it fully. The traffic is not anonymized and can be analyzed only on Merit's machines.

Regional optical network RONX dataset contains sampled Netflow records from a mid-size US regional network, connecting educational, research, government and business institutions to the Internet. To detect scans, we identify flows with only TCP SYN flag set. A short flow may thus be misclassified as a scan, because other packets on this flow are not sampled. To address this we also, for each source prefix, keep track of flows that have TCP PSH or TCP ACK flags set, but not TCP SYN flag. We assume that these flows indicate an established connection. We then calculate the difference of SYN flows and PSH/ACK flows. If the difference is positive, we use it as the estimate of scan flows from the given source prefix. Our dataset consists of 5-minute long Netflow record collections, from February 24, 2020 to April 30, 2020. To protect user privacy, the IP addresses in records are anonymized in a prefix-preserving manner, using CryptoPAN [9]. With the consent of RONX operators, we have obtained a list of anonymized and original /24 prefixes in the dataset, so we could use them to classify a prefix as cloud or non-cloud.

B. Blocklists

Blocklists usually represent IP addresses, ASNs or DNS names that are sources of some observed misbehavior (e.g., sending spam). We describe our datasets below. When a blocklist includes ASN we disaggregate the corresponding autonomous system into its /24 prefixes and include the prefixes in our dataset. When a blocklist includes DNS names, we use DNS queries to obtain corresponding IP addresses associated and include them in our dataset.

Scamalytics: IP fraud risk lookup tool [35] is known for maintaining the largest shared anti-fraud database, mainly dedicated to the online dating industry. It computes the fraud score for all known IPs, and publishes the top 100 IPs each month, and their fraud scores. Our dataset includes top 102 IP addresses for March 2020 and April 2020.

F5 Labs [18] maintains the list of the top 50 malicious autonomous systems and attacker IPs, which are associated with 14 million different attacks across the globe. Majority of these are Web application attacks. Organizations and attacker IPs are ranked by the number of attacks generated. This dataset is spread over 90-day period from Aug 2019 to Oct 2019.

BLAG [47] project produces a publicly available master blocklist [38], obtained by aggregating content from 157 publicly available, popular blocklists. A new master list is published whenever any of the blocklists updates its contents. We use the data for the entire 2018 and 2019, separated into two datasets: 0.5 billion IP addresses in 2018, and 5 billion IP addresses for 2019 and January 2020.

Google Safe Browsing [14] examines billions of URLs per day looking for unsafe websites, and publishes their list. We collected this list from May 8, 2020 to May 16, 2020 from maltiverse.com [14].

COVID-19 phishing URL's list from maltiverse.com [7] contains 239 phishing URLs related to COVID-19 content, from March 13, 2020 to May 16, 2020.

COVID-19 malicious hostnames/URL's List from maltiverse.com [8] contains 9,874 malicious hostnames/URLs from January 2020 to May 2020 that contain the words "COVID-19" or "corona" and are known to generate different types of unwanted traffic.

Openphish [20] maintains the list of autonomous systems associated with phishing. We collect 54 snapshots, from May 15, 2020 to May 22, 2020, each containing top 10 ASNs associated with phishing. The unique snippets of the Web page that we downloaded are available at [21].

Cybercrime Tracker [10] maintains a public list of IP addresses that are known to spread malware at [10]. We use data from January 1st, 2019 to May 12th, 2020, which comprised 3,471 IP addresses that spread malwares. This dataset overlaps with the BLAG dataset.

udger.com [15] maintains a list of known user agent strings, and also maintains a list of source IPs of known attacks, which is updated every half an hour. We collected a total of 101 snapshots from this list including a total of 3,030 IP addresses, on April 13, 2020, and also from May 4, 2020 to May 20, 2020. Attacks include installations of vulnerable versions of popular web applications, brute-force login attempts, and floods.

BGP Ranking [3] provides a ranking model to rank the ASNs from the most malicious ASN to the least malicious ASN using data from compromised systems and other publicly available blocklists. We collected the snapshot of Top 100 most malicious ASNs on August 21st, 2020.

C. Limitations

The darknet datasets can contain spoofed traffic, which may lead us to misclassify the origin of the scans. We employ techniques outlined in [37] to identify and remove randomly spoofed traffic, but these techniques cannot guarantee that all spoofed traffic is removed. Further, blocklist datasets are produced using proprietary algorithms, and they may contain some false positives. We have no way to independently establish accuracy of the blocklists. However, we believe that spoofing and false positives are equally likely to affect any prefix, regardless of whether it belongs to a cloud or a non-cloud. Thus we expect that spoofing and false positives do not affect relative comparison between clouds and non-clouds, which is the focus of our paper.

D. Misbehavior Metrics

We define several measures of misbehavior in this section. To quantify misbehavior of a /24 prefix in a network trace dataset, we define $mal^{d}_{scans}(t)$, which is the number of scans this prefix sends during time t in dataset d. Similarly, we define mal^{d}_{bl}, which is the number of times the prefix appears on a given blocklist. To quantify maliciousness of an organization, we calculate their rank for each dataset, ordered in the decreasing order of their contribution to the dataset (number of scans or number of appearances on the blocklist). We then report the average rank for an organization $rank_{avg}$. As organizations own address spaces of different sizes, it may seem that an organization's likelihood to host misbehaving nodes grows with its address size. To account for this, we devise another, normalized measure of maliciousness $malorg^{d}_{score}$, as:

$$malorg^{d}_{score} = \frac{prefs_{d} \cdot r_{d}}{prefs_{tot}} \tag{1}$$

where $prefs_{tot}$ is the number of /24 prefixes owned by the given organization, and $prefs_{d}$ is the number out of these prefixes that appear in the dataset d. The measure r_{d} is the fraction of the contribution of this organization to the total scans or entries in the dataset. We then report average across datasets – $malorg_{score}$

IV. Results

In this section we report our findings from datasets, organized into findings from network traces and findings from blocklists. In both categories of datasets, clouds appear 2–100 times more aggressive than non-clouds. We also observe heavy tail misbehavior from both clouds and non-clouds. Thus, if defenders focused their efforts on this handful of organizations they could eliminate most of malicious behaviors.

Dataset	Source	Format	Description	Start Date	End Date	Size
CAIDA	[19]	PCAP files	Real-Time Network Telescope Data	12-Mar-20	28-Apr-20	80 GBs hourly compressed files (a few billion packets)
Merit	[17]	PCAP files	Real-Time Network Telescope Data	11-Mar-20	19-Mar-20	2 GBs hourly compressed files (fewer than 0.1 billion packets)
RONX	Anon.	Netflow files	Live ISP data (all 5-minutes long Netflow records)	24-Feb-20	26-Apr-20	1.15 TB
Scamalytics	[35]	IP addresses	Top 100 monthly IPs with the maximum fraud in online dating	1-Mar-20	30-Apr-20	204 IP addresses
udger.com	[15]	IP addresses	Source IP addresses associated with different attacks	4-May-20	5-May-20	3,030 IP addresses
Cybercrime Tracker	[10]	IP addresses	IPs that spread malwares	1-Jan-19	12-May-20	3,471 IP addresses
Google Safebrowsing	[14]	URLs	Malicious URLs List from maltiverse.com	8-May-20	16-May-20	7,886 URLs (that belong to non-bogon IPv4 address range)
COVID-19 Hostnames	[8]	Hostnames / URLs	Malicious hostnames/URLs that contain the word ”COVID-19”/”corona” and are associated with generating different variants of unwanted traffic	1-Jan-20	16-May-20	9,874 malicious hostnames
COVID-19 Phishing	[7]	URLs	Phishing URLs related to COVID-19 content	13-Mar-20	16-May-20	239 phishing URLs
Openphish	[20]	ASNs and ASNs Domains	Top 10 ASNs associated with phishing	15-May-20	22-May-20	54 snippets of top 10 ASNs
BGP Ranking	[3]	ASNs	Maintains top 100 malicious ASNs	21-Aug-20	21-Aug-20	101 ASNs
BLAG (2018)	[47], [38]	IP addresses	Publicly available blacklisted IPs list collected daily using 157 popular blocklists	1-Jan-18	31-Dec-18	0.5 billion IP addresses (14.5 million unique IPs)
BLAG (2019)	[47], [38]	IP addresses	Publicly available blacklisted IPs list collected daily using 157 popular blocklists	1-Jan-19	31-Jan-20	5 billion IP addresses
F5 Labs: Attack Traffic	[18]	IP addresses / ASNs	Top 50 malicious ASNs and IPs ranked by number of attacks for more than 14 million attacks globally	1-Aug-19	31-Oct-19	50 ASNs and 50 IPs

TABLE I
DATASETS SUMMARY.

A. *Findings from Network Traces*

Cloud prefixes are more aggressive than non-cloud prefixes across trace datasets. The average measure mal_{scans} for a cloud prefix is 20 - 100 times higher than the average mal_{scans} for a non-cloud prefix. Figures 3(a), 3(d) and 3(g) show the average mal_{scans} for CAIDA, Merit and RONX respectively, for cloud and non-cloud prefixes.

At the same time, the number of /24 prefixes is 30 – 60 times higher for non-clouds than for clouds. Figures 3(b), 3(e) and 3(h) show the number of /24 prefixes for CAIDA, Merit and RONX respectively, for clouds and non-clouds.

Although cloud prefixes make only 2-3% of prefixes in the network trace datasets, total scans per hour from clouds are similar to scans per hour from non-clouds (the remaining 97–98% of prefixes). The percentage contribution of unwanted traffic by clouds and non-clouds stays roughly equal throughout the full duration of the datasets. Both clouds and non-clouds contribute about 50% of the scans, as shown in Figures 3(c), 3(f) and 3(i).

We note that in all three datasets, port 8545, associated with Ethereum, cryptocurrency and e-wallet services, is frequently scanned by clouds but far less often by non-clouds. Port 8545 contributes to 1.7%, 1.3% and 3.7% of cloud scans in CAIDA, Merit and RONX datasets, respectively. Conversely it only contributes to 0.3%, 0.4% and 1% of non-cloud scans, respectively. **This supports an observation that attackers also misuse clouds for possible monetary gains [13].**

Larger clouds seem to be less malicious than smaller clouds in terms of the total mal_{scans}. Figure 5(a) shows the $rank_{avg}$ of top clouds and non-clouds across *Network Traces* datasets. Organizations are ordered by organization-wise total mal_{scans}. There seems to be a slight downward trend in the number of scans generated per /24 prefix as we go from smaller to larger organizations. The misbehavior associated with organization as a whole, i.e. the $malorg_{score}$, is higher for small-sized clouds and non-clouds than for large clouds. Figure 5(b) shows the $rank_{avg}$ of top clouds and non-clouds for *Network Traces* datasets ordered by organization-wise $malorg_{score}$. 90% of these top organizations have fewer than 10 /24 prefixes.

Misbehavior is heavy-tailed. We also investigate contribution of each organization to the total number of scans sent by clouds and non-clouds. In all three datasets there is a heavy-tailed distribution of organization's contributions, as shown in Figure 6, i.e., **around 25 clouds and 200 non-clouds, generate 90% of the malicious traffic**. Thus if defenders focused their efforts on these heavy hitters, they could eliminate a large amount of misbehavior on the Internet.

B. *Findings from Blocklists*

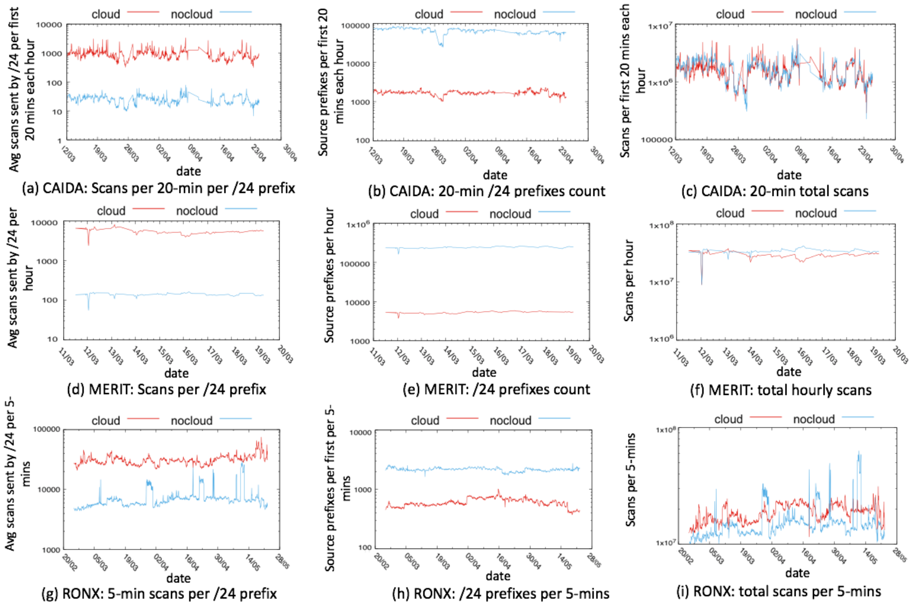
Figure 4 shows the percentage of contribution to blocklist entries by clouds and non-clouds in different datasets. **Clouds contribute anywhere from 22% to 96% of entries, in spite of being only 5.4% of the routable address space**.

On the average, the total mal_{bl} per cloud is 1.82 times higher than the total mal_{bl} per non-cloud. Thus cloud prefixes are **almost twice as likely** to engage in misbehavior that lands them on a blocklist than non-clouds. **Clouds play a vital role in spreading malware and supporting phishing URLs** as evident from COVID-19 datasets, Openphish Top ASNs, Google Safe Browsing and Cybercrime Tracker. In each of these datasets, clouds account for at least **80%** of the entries.

Clouds play a vital role in attacks. Around **61%** of the global attacks from the F5 Labs Research dataset originated from clouds, i.e., **8.6 M** out of the**14 M** attacks. There is *heavy tail in contributions of both clouds and non-clouds to the blocklists.* Figure 6 shows that the top 10 clouds ranked by the total mal_{bl}, account for 20% of the total blocklist entries that list cloud addresses. Most frequently blocklisted clouds are OVH and Digital Ocean, shown in Figures 5(c) and 5(d). OVH, Digital Ocean and Namecheap also have a high $malorg_{score}$.

V. RELATED WORK

In addition to news articles mentioned in Section I , there are research articles examining cloud misbehavior. Zhao et. al [51] observe that cloud service providers are being preferred by cyber-criminals to inflict harm on online services at a large scale. Our findings complement their observations. Vlajic et. al [49] show that the virtual private servers of cloud providers are vulnerable to a number of misuse attempts, which involve the use of IP spoofing. Liao et al. [39], [41] discover that cloud web hosting services are being used as a platform for long-tail Search Engine Optimization (SEO) spam attacks, and measure the effectiveness of these attacks. Liao et. al [40] show



(a) CAIDA: Scans per 20-min per /24 prefix

(b) CAIDA: 20-min /24 prefixes count

(c) CAIDA: 20-min total scans

(d) MERIT: Scans per /24 prefix

(e) MERIT: /24 prefixes count

(f) MERIT: total hourly scans

(g) RONX: 5-min scans per /24 prefix

(h) RONX: /24 prefixes per 5-mins

(i) RONX: total scans per 5-mins

Fig. 3. Network Traces Datasets

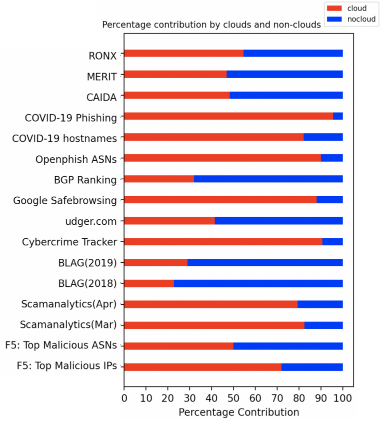
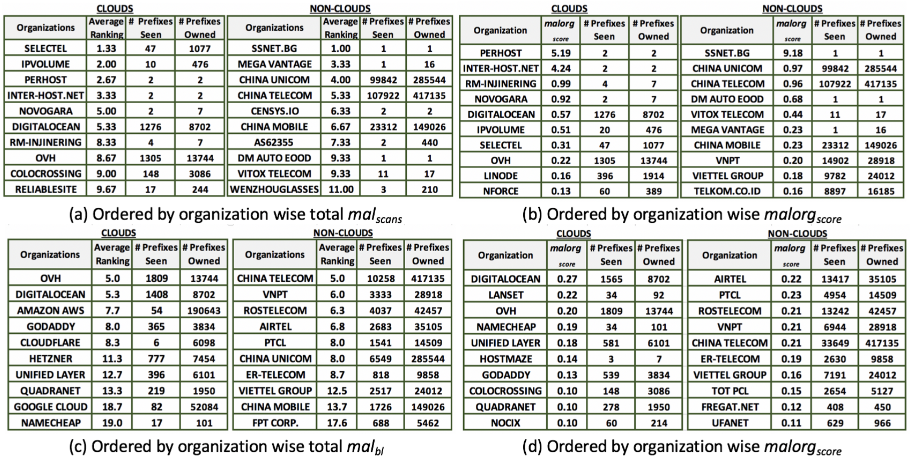


Fig. 4. Percentage of clouds and non-clouds in the different *Network Traces* datasets and *Blocklists* datasets



CLOUDS

Organizations	Average Ranking	# Prefixes Seen	# Prefixes Owned
SELECTEL	1.33	47	1077
IPVOLUME	2.00	10	476
PERHOST	2.67	2	2
INTER-HOST.NET	3.33	2	2
NOVOGARA	5.00	2	7
DIGITALOCEAN	5.33	1276	8702
RM-INJINERING	8.33	4	7
OVH	8.67	1305	13744
COLOCROSSING	9.00	148	3086
RELIABLESITE	9.67	17	244

NON-CLOUDS

Organizations	Average Ranking	# Prefixes Seen	# Prefixes Owned
SSNET.BG	1.00	1	1
MEGA VANTAGE	3.33	1	16
CHINA UNICOM	4.00	99842	285544
CHINA TELECOM	5.33	107922	417135
CENSYS.IO	6.33	2	2
CHINA MOBILE	6.67	23312	149026
AS62355	7.33	2	440
DM AUTO EOOD	9.33	1	1
VITOX TELECOM	9.33	11	17
WENZHOUGLASSES	11.00	3	210

(a) Ordered by organization wise total mal_{scans}

CLOUDS

Organization	$malorg_{score}$	# Prefixes Seen	# Prefixes Owned
PERHOST	5.19	2	2
INTER-HOST.NET	4.24	2	2
RM-INJINERING	0.99	4	7
NOVOGARA	0.92	2	7
DIGITALOCEAN	0.57	1276	8702
IPVOLUME	0.51	20	476
SELECTEL	0.31	47	1077
OVH	0.22	1305	13744
LINODE	0.16	396	1914
NFORCE	0.13	60	389

NON-CLOUDS

Organization	$malorg_{score}$	# Prefixes Seen	# Prefixes Owned
SSNET.BG	9.18	1	1
CHINA UNICOM	0.97	99842	285544
CHINA TELECOM	0.96	107922	417135
DM AUTO EOOD	0.68	1	1
VITOX TELECOM	0.44	11	17
MEGA VANTAGE	0.23	1	16
CHINA MOBILE	0.23	23312	149026
VNPT	0.20	14902	28918
VIETTEL GROUP	0.18	9782	24012
TELKOM.CO.ID	0.16	8897	16185

(b) Ordered by organization wise $malorg_{score}$

CLOUDS

Organizations	Average Ranking	# Prefixes Seen	# Prefixes Owned
OVH	5.0	1809	13744
DIGITALOCEAN	5.3	1408	8702
AMAZON AWS	7.7	54	190643
GODADDY	8.0	365	3834
CLOUDFLARE	8.3	6	6098
HETZNER	11.3	777	7454
UNIFIED LAYER	12.7	396	6101
QUADRANET	13.3	219	1950
GOOGLE CLOUD	18.7	82	52084
NAMECHEAP	19.0	17	101

NON-CLOUDS

Organizations	Average Ranking	# Prefixes Seen	# Prefixes Owned
CHINA TELECOM	5.0	10258	417135
VNPT	6.0	3333	28918
ROSTELECOM	6.3	4037	42457
AIRTEL	6.8	2683	35105
PTCL	8.0	1541	14509
CHINA UNICOM	8.0	6549	285544
ER-TELECOM	8.7	818	9858
VIETTEL GROUP	12.5	2517	24012
CHINA MOBILE	13.7	1726	149026
FPT CORP.	17.6	688	5462

(c) Ordered by organization wise total mal_{bl}

CLOUDS

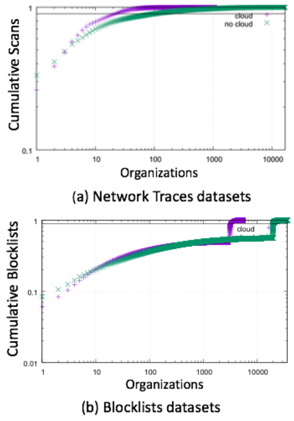
Organization	$malorg_{score}$	# Prefixes Seen	# Prefixes Owned
DIGITALOCEAN	0.27	1565	8702
LANSET	0.22	34	92
OVH	0.20	1809	13744
NAMECHEAP	0.19	34	101
UNIFIED LAYER	0.18	581	6101
HOSTMAZE	0.14	3	7
GODADDY	0.13	539	3834
COLOCROSSING	0.10	148	3086
QUADRANET	0.10	278	1950
NOCIX	0.10	60	214

NON-CLOUDS

Organization	$malorg_{score}$	# Prefixes Seen	# Prefixes Owned
AIRTEL	0.22	13417	35105
PTCL	0.23	4954	14509
ROSTELECOM	0.21	13242	42457
VNPT	0.21	6944	28918
CHINA TELECOM	0.21	33649	417135
ER-TELECOM	0.19	2630	9858
VIETTEL GROUP	0.16	7191	24012
TOT PCL	0.15	2654	5127
FREGAT.NET	0.12	408	450
UFANET	0.11	629	966

(d) Ordered by organization wise $malorg_{score}$

Fig. 5. $rank_{avg}$ of top clouds and non-clouds



(a) Network Traces datasets

(b) Blocklists datasets

Fig. 6. Cumulative scans ratio distribution.

that clouds are being used to aid malicious online activities, as repositories for malicious content. Alrwais et. al [29] present a study on the BulletProof Hosting ecosystem [4], reporting 39 K malicious sub-allocations, distributed across 3,200 autonomous systems.Ahmad et al. [28] show that cloud abuse threats have not been addressed yet. They showcase security vulnerabilities associated with clouds, provide anecdotal evidence, and mention some known solutions for different scenarios. Doelitzscher et al. [33] express concern for cloud abuse. They present an anomaly detection system for IaaS clouds based on customers' usage patterns. Lindemann [42] highlights the problem of cloud abuse and presents a survey of possible abuses from seven clouds. All the related works focus on a limited set of clouds and misbehaviors, while we study a larger set of clouds and leverage diverse data sources to systematically analyze and quantify misbehavior.

VI. Conclusion

Clouds can be misused, either due to the negligence or because they explicitly permit misuse. In this paper, we quantify misbehavior from clouds and measure it in terms of participation in scanning activities, and being listed on blocklists. In all of our 13 datasets, cloud prefixes misbehave much more than non-clouds: they generate 20 – 100 times more scans and are twice more likely to end on a blocklists. Clouds also play a vital role in spreading malware and supporting phishing. Some small clouds misbehave more per /24 prefix than larger clouds, which may indicate lack of security resources, or higher inclination to allow malicious activities. Both clouds and non-clouds misbehave in heavy-tailed manner. Top 25 clouds account for 90% of the unwanted scans from clouds, and 10 clouds contribute more than 20% of blocklisted cloud addresses. Thus, if efforts are focused on securing these clouds, Internet attacks can be greatly reduced.

VII. Acknowledgement

We are grateful to CAIDA, Merit Network, RONX and IPinfo.io for allowing us to use their data in our research. This material is based on research sponsored by the Department of Homeland Security (DJ-IS) Science and Technology Directorate, Homeland Security Advanced Research Projects Agency (HSARPA), Cyber Security Division (DHS S&T/HSARPA CSD) BAA HSHQDC-14-R-B0005, and the Government of United Kingdom of Great Britain and Northern Ireland via contract number D15PC00184. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Department of Homeland Security, the U.S. Government, or the Government of United Kingdom of Great Britain and Northern Ireland.

References

[1] "Akamai: What are the Security Risks of Cloud Computing?" https://tinyurl.com/y6c96lxq.

[2] "Alexa Top 500 sites," https://tinyurl.com/y5j8lkrc, Retrieved on October 5, 2019.

[3] "BGP Ranking," https://bgpranking.circl.lu/.

[4] "Bulletproof Hosting," https://en.wikipedia.org/wiki/Bulletproof_hosting.

[5] "Cloud services from leading providers including AWS, Microsoft Azure and Alibaba used in 25% of all DDoS attacks in Europe from July 2017 to June 2018," https://www.link11.com/en/blog/threat-landscape/public-cloud-services-increasingly-exploited-to-supercharge-ddos-attacks-new-link11-research/.

[6] "Clouds List," https://steel.isi.edu/Projects/Cloud_Misbehavior/.

[7] "COVID-19 Phishing URL's," https://maltiverse.com/collection/TYErBHEB8jmkCY9eFseO.

[8] "Covid19 malicious Hostnames/URL's," https://maltiverse.com/collection/_IyQknEB8jmkCY9ehYUn.

[9] "Crypto-PAn," https://en.wikipedia.org/wiki/Crypto-PAn.

[10] "Cybercrime Tracker," http://cybercrime-tracker.net/.

[11] "Cybercriminals Abuse Amazon Cloud to Host Linux DDoS Trojans," https://tinyurl.com/yyxwgddt.

[12] "DDoS-ers Launch Attacks From Amazon EC2," https://www.infosecurity-magazine.com/news/ddos-ers-launch-attacks-from-amazon-ec2/.

[13] "Experts warn hackers have already stolen over $20 Million from Ethereum clients exposing interface on port 8545," https://tinyurl.com/yxv3e29d.

[14] "Google Safebrowsing - Malicious URL," https://maltiverse.com/collection/8nxipXAB8jmkCY9euMUp.

[15] "IP Addresses of Attack Sources," https://udger.com/resources/ip-list/known_attack_source.

[16] "IPv4 Private Address Space and Filtering," https://www.arin.net/reference/research/statistics/address_filters/.

[17] "Merit Network, Inc.'s network telescope," https://www.merit.edu/initiatives/#1598371817757-ee63e0a6-5330 (Accessed Oct. 31, 2020).

[18] "Regional Threat Perspectives, Fall 2019: United States," https://www.f5.com/labs/articles/threat-intelligence/regional-threat-perspectives--fall-2019--united-states.

[19] "The UCSD Network Telescope," https://www.caida.org/projects/network_telescope/.

[20] "Top 10 ASNs," https://openphish.com/phishing_activity.html.

[21] "Top 10 ASNs," https://tinyurl.com/y36n5lqs.

[22] "Top 5 Risks of Cloud Computing," https://www.calyptix.com/research-2/top-5-risks-of-cloud-computing/.

[23] "Top 500 Web Hosts," https://www.hostingkingdom.com/top-500.

[24] "Top Cloud Data Security Risks, Threats, And Concerns," https://blog.panoply.io/top-cloud-security-threats-risks-and-concerns.

[25] "Top Cloud Security Risks Every Company Faces," https://www.whizlabs.com/blog/cloud-security-risks/.

[26] "Total size of the public cloud computing market from 2008 to 2020," https://tinyurl.com/y5g4xbaq.

[27] "Wikipedia: Cloud computing providers," https://en.wikipedia.org/wiki/Category:Cloud_computing_providers/, Retrieved on October 5, 2019.

[28] I. Ahmad and H. Bakht, "Security Challenges from Abuse of Cloud Service Threat," *International Journal of Computing and Digital Systems*, vol. 8, no. 01, pp. 19–31, 2019.

[29] S. Alrwais, X. Liao, X. Mi, P. Wang, X. Wang, F. Qian, R. Beyah, and D. McCoy, "Under the shadow of sunshine: Understanding and detecting bulletproof hosting on legitimate service provider networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 805–823.

[30] R. Barnett, "Web application defender's field report: Account takeover campaigns spotlight," June 2016, https://tinyurl.com/ztqq64m.

[31] K. Benson, A. Dainotti, K. C. Claffy, and E. Aben, "Gaining Insight into AS-level Outages Through Analysis of Internet Background Radiation," in *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2013, pp. 447–452.

[32] Dhanalakshmi, "The latest cloud hosting service to serve malware," September 2018, https://www.zscaler.com/blogs/research/latest-cloud-hosting-service-serve-malware.

[33] F. Doelitzscher, M. Knahl, C. Reich, and N. Clarke, "Anomaly Detection in IaaS Clouds," in *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, vol. 1, 2013, pp. 387–394.

[34] T. Emmons, "Do DDoS Attacks Originate From Cloud Service Providers?" May 2019, https://tinyurl.com/y6zama4a.

[35] S. GDI Content Partners, "Scamalytics Release High Risk ISPs For March 2020," https://www.globaldatinginsights.com/content-partners/scamalytics/scamalytics-release-high-risk-isps-for-march-2020/.

[36] IPinfo.io, "Hosted Domains by ASNs Report," 2020, https://ipinfo.io/hosting.

[37] M. Jonker, A. King, J. Krupp, C. Rossow, A. Sperotto, and A. Dainotti, "Millions of targets under attack: a macroscopic characterization of the DoS ecosystem," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 100–113.

[38] S. Lab, "BLAG: Blacklist Aggregator," 2020, https://steel.isi.edu/members/sivaram/BLAG/.

[39] X. Liao, S. Alrwais, K. Yuan, L. Xing, X. Wang, S. Hao, and R. Beyah, "Lurking Malice in the Cloud: Understanding and Detecting Cloud Repository as a Malicious Service," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1541–1552.

[40] ——, "Cloud repository as a malicious service: challenge, identification and implication," *Cybersecurity*, vol. 1, no. 1, p. 14, 2018.

[41] X. Liao, C. Liu, D. McCoy, E. Shi, S. Hao, and R. Beyah, "Characterizing long-tail SEO spam on cloud web hosting services," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 321–332.

[42] J. Lindemann, "Towards Abuse Detection and Prevention in IaaS Cloud Computing," in *2015 10th International Conference on Availability, Reliability and Security*, 2015, pp. 211–217.

[43] R. Miao, R. Potharaju, M. Yu, and N. Jain, "The Dark Menace: Characterizing Network-based Attacks in the Cloud," in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 169–182.

[44] Qurium, "Who are we?" https://www.qurium.org/gdpr/.

[45] ——, "Azerbaijan and the FineProxy DIY DDoS Service (Region40 / QualityNetwork)," 2018, https://www.qurium.org/alerts/azerbaijan/azerbaijan-and-the-region40-ddos-service/.

[46] ——, "FineProxy Used to Launch DDoS Attack Against Site Critical of Azerbaijani State Oil Company's Leader," 2018, https://tinyurl.com/y2t5fqdm.

[47] S. Ramanathan, J. Mirkovic, and M. Yu, "BLAG: Improving the Accuracy of Blacklists," in *Proceedings of NDSS*, 2020.

[48] S. Torabi, E. Bou-Harb, C. Assi, E. B. Karbab, A. Boukhtouta, and M. Debbabi, "Inferring and Investigating IoT-Generated Scanning Campaigns Targeting A Large Network Telescope," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2020.

[49] N. Vlajic, M. Chowdhury, and M. Litoiu, "IP Spoofing In and Out of the Public Cloud: From Policy to Practice," *Computers*, vol. 8, no. 4, p. 81, 2019.

[50] M. D. Will Glazier, “Automation Attacks at Scale - Credential Exploitation,” *Grehack : 2017*, 2017, https://grehack.fr/data/2017/slides/GreHack17_Automation_Attacks_at_Scale_paper.pdf.

[51] B. Z. H. Zhao, M. Ikram, H. J. Asghar, M. A. Kaafar, A. Chaabane, and K. Thilakarathna, “A Decade of Mal-Activity Reporting: A Retrospective Analysis of Internet Malicious Activity Blacklists,” in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, pp. 193–205.