Quality Evaluation of Cyber Threat Intelligence Feeds

Harm Griffioen, Tim Booij and Christian Doerr
TU Delft, Cyber Security Group
Delft, Netherlands
h.j.griffioen@tudelft.nl,t.m.booij@student.tudelft.nl,c.doerr@tudelft.nl

ABSTRACT

In order to mount an effective defense, information about likely adversaries, as well as their techniques, tactics and procedures is needed. This so-called *cyber threat intelligence* helps an organization to better understand its threat profile. Next to this understanding, specialized feeds of indicators about these threats downloaded into a firewall or intrusion detection system allow for a timely reaction to emerging threats.

These feeds however only provide an actual benefit if they are of high quality. In other words, if they provide relevant, complete information in a timely manner. Incorrect and incomplete information may even cause harm, for example if it leads an organization to block legitimate clients or if the information is too unspecific and results in an excessive amount of collateral damage.

In this paper, we evaluate the quality of 17 open source cyber threat intelligence feeds over a period of 14 months, and 7 additional feeds over 7 months. Our analysis shows that the majority of indicators are active for at least 20 days before they are listed. Additionally, we have found that many list have biases towards certain countries. Finally, we also show that blocking listed IP addresses can yield large amounts of collateral damage.

1 INTRODUCTION

In order to effectively protect a system, one needs information. This includes information about possible attackers, their capabilities, their commonly used tactics and techniques as well as feasible countermeasures. This information can be gathered by a company itself, or obtained from providers specialized in providing this information. At a first glance, the idea of gaining a head start and an upper hand to combat the activities of malicious adversaries seems like an oxymoron. After all, if an intelligence provider distributes information about the activities, used tools or domains and IP addresses from which attacks are being carried out to a wide public, adversaries would immediate recognize having been uncovered and correspondingly change their activities, thereby negating the benefits a defender would have from this information in the first place. The use of cyber threat intelligence is thus a race against the clock, where published information is prone to soon lose its value. This means that the distribution of information about adversarial activities as soon as possible is key. While it is in principle possible for organizations to assemble and grow such a body of knowledge, most shy away from the associated costs and complexity in building such *cyber threat intelligence* (CTI) themselves, but rather turn to commercial and open source threat intelligence providers for help.

ACNS '20, October 19-22, 2020, Rome, Italy
.

In the very recent past, this shift to intelligence-driven defense has led to the emergence of a plethora of companies providing CTI feeds, often at a heavy price tag. Threat intelligence feeds range from complex sitreps, sector analysis and trend reports in essay formats, to machine-readable lists of indicators of compromise such as traffic signatures or malicious IP addresses that organizations can download and install into their firewall, thereby catering to the entire spectrum of organizational cyber security maturity levels. In addition to such for-profit services, a variety of open source alternatives have sprung up, which are provided by security companies as a marketing instrument or driven by a community effort of aggregating information across defenders. While commercial feeds may add unique results for example from internal forensic investigation, many commercial providers have been known to repackage, curate and resell other (open source) lists [5].

By definition, cyber threat intelligence is however a highly perishable good, since as soon as it is discovered and distributed to clients, also adversaries will know that one of their tools or assets has been discovered and would try to replace this "burnt" artifact as soon as possible. Thus, in order to be effective, threat intelligence has to be timely, but also highly accurate. With inaccurate listings, automatic download and application of indicator information could lead to undesired effects such as blocking traffic from benign clients. This collateral damage may overall do more harm than good, which leads to the question how effective these feeds actually are.

In this paper, we aim to answer the question *How effective are open source Cyber Threat Intelligence feeds, and how can we measure their quality?* To do this, we create suitable metrics to evaluate the quality of 24 open source cyber threat intelligence feeds, and estimate the utility and risk each of these services provides to an organization. With our work, we make the following three contributions:

- We introduce a taxonomy to evaluate the quality of cyber threat intelligence feeds aiming to assess the utility the user may gain from such a feed.
- We evaluate the indicators reported on 24 open source threat intelligence feeds across four dimensions, and benchmark using NetFlow data and zone transfers the timeliness, sensitivity, originality and impact of these feeds.
- We empirically analyze the impact, a listing of an indicator on an intelligence feed has, on its activity thereafter. This allows us to evaluate the adoption of these feeds in practice and estimate whether a feed is in practice able to "save" clients and networks from future harm.

The remainder of this paper is structured as follows: Section 2 discusses existing work into cyber threat intelligence and its evaluation. In Section 3, we develop evaluation criteria for a quality

assessment of cyber threat intelligence that will be used within this paper. Section 4 describes open source intelligence feeds collected for the analysis, while Section 5 describes their utility in terms of relevance, timeliness, completeness and accuracy. Section 6 evaluates the adoption of these sources in practice and the benefit they bring to networks. Section 7 summarizes our findings and concludes our work.

2 RELATED WORK

Cyber threat intelligence feeds are widely used in industry, and are relied on as a useful tool to mitigate attacks. Despite major commercial interest in these feeds, initial surveys indicate that the quality of these feeds might not be as high as one would like. In a study in 2015 [3], authors state that most intelligence was not specific enough. Additionally, 66% of respondents state that the information is not timely.

On the same note, Tounsi et al. [12] states that there are still many limitations when it comes to threat intelligence. One of the limitations is that there is too much information, with 250 million indicators per day. Another finding in this paper, is that threat intelligence available to enterprises is often out of date due to the short lifespan, and therefore not always useful. Limitations of blocklists are also apparent in [9], in which the authors show the incompleteness of the evaluated blacklists. To further measure the shortcomings, several works have been focused on empirical evaluation of these feeds [7, 10, 11], as well as test suites with specific goals in testing blocklists [8].

In 2014, Kührer et al. published a paper in which multiple blacklists are empirically analyzed [7]. In this paper, the authors identify for a number of domain lists the active, parked and sinkholed domains. The analysis in the paper gives insight into domain blacklists, measuring their accuracy, completeness and estimating the timeliness of the blacklists. As the goal of the study was not to create metrics to generalize the evaluation of cyber threat intelligence feeds, no metrics have been created and evaluated in this topic.

A Defcon presentation by Pinto and Maxwell [2] aims to measure the effectiveness of threat intelligence feeds in two dimensions. In this presentation, the authors show evaluations for the scope and accuracy of these feeds. Their research has been further complemented by Pawlińnski and Kompanek [1], which state that there are eight criteria in which the quality of a threat intelligence feed can be measured. These metrics are however not evaluated on a large number of CTI feeds, and some of these metrics are hard to evaluate automatically.

In this paper, we propose four quality metrics for CTI feeds that can be automatically analyzed, and analyze 24 different open source CTI feeds using these metrics.

3 QUALITY CRITERIA FOR THREAT INTELLIGENCE

In this section, we will describe four different criteria, which we will use in the following to evaluate the quality of open source threat intelligence feeds. As discussed in the related work, Pawlińnski and Kompanek [1] have proposed at an industry forum a taxonomy to benchmark threat intelligence along the dimensions of (a) relevance, (b) accuracy, (c) completeness, (d) timeliness, and (e) ingestibility. We find this classification however problematic, as several of the criteria are entangled: for example, in the machine learning and pattern recognition domains, relevance is usually measured by precision and recall. In other words, how many of the selected items in a dataset are correctly identified, and how many of the relevant items are found in the dataset, respectively. Recall however also partially assesses similar aspects as completeness, so quantification results would contain some degree of correlation. Along the same lines, accuracy is also widely used concept in machine learning, and in binary classification measures the ratio of true results to all examined data, or $\frac{TP+TN}{TP+TN+FP+FN}$. While threat intelligence is a classification task, classifying activity as either malicious or non-malicious, threat intelligence *feeds* are not classification tasks, but should mainly contain information from one label. Therefore, a binary accuracy characterization does not work well due to imbalance of the data present in the feeds. For these reasons, we propose complementary metrics to measure the quality of these feeds.

A Taxonomy for CTI Quality

In order to evaluate the quality of cyber threat intelligence, we therefore propose a set of four metrics: timeliness, sensitivity, originality and impact, which we will describe in further detail in the following:

(1) *Timeliness.* The goal of subscribing to a threat intelligence feed is to obtain early warning of some emergent malicious activity, so that infections in the local area network can be stopped in time before significant losses are incurred. Hence, the earlier indicators such as IP addresses or domain names are flagged, the higher the utility of the feed is to the subscriber, and in turn we can also conclude the better the quality of the provided information. One essential quality criteria of a threat intelligence feed is thus the timeliness of the information posted, in other words how soon a domain or IP address is included in such lists after it has started malicious activities. A high timeliness will minimize the amount of damage that could be incurred as part of a compromise, as it shortens the time window during which hosts may be under adversarial control and the time an adversary may for example exfiltrate data or abuse the infected client.

(2) *Sensitivity.* In order to be included into a feed, the threat intelligence provider has to observe some malicious activity in the first place. This is typically done using a variety of sensors, recording network traffic patterns, DNS lookups, as well as for example based on the forensic analysis of malware samples. If a particular malware instance, C&C server, or maliciously acting host shows only low, sporadic activity, there is a high likelihood that it would not be seen by a provider and thereby go by unmitigated until the problem grows above a certain threshold.
With sensitivity, we therefore assess what volume is necessary for the intelligence provider to take notice of a malicious activity, in other words what is the average and typical minimum threshold at which detection will take place. In addition to quantifying the overall per-feed threshold, we can also measure the sensitivity of a threat intelligence feed with respect to a geographical focus: if a provider predominantly

has sensors in a specific region, detection will be biased against threats emerging or deployed in this particular area, while comparatively insensitive towards threats originating outside of the measurement coverage. As Internet threats by definition operate worldwide, heavy geographical biases therefore introduce a significant risk of getting hit unprepared.

(3) *Originality.* In practice, an organization would likely subscribe to several threat intelligence feeds, as CTI providers often specialize towards a particular type of threats. We also see this behavior in threat intelligence providers themselves, who – as we have said earlier – are also often aggregating, curating and repacking other sources to be marketed as their own service. An essential metric of a cyber threat intelligence feed is therefore originality, in other words the amount of information that is unique to this particular source and that could not be obtained otherwise.
While originality measure the contribution made by one specific feed, it can also be used as a metric to quantify an ecosystem of intelligence feeds as a whole. Consider a number of k feeds which all report malware C&C servers. If all indicators provided by these feeds are highly unique, in other words there is no or only limited overlap between them, this also means that even their union provides only an insufficient peak at the population of C&C servers. We can thus say that in case of high ecosystem-wide originality each feed only draws for samples from a large problem space, and in these cases the set of intelligence feeds is unsuited to provide sufficient defense against this particular type of threat.

(4) *Impact.* When an organization applies the information obtained from the threat intelligence feeds, this should lead to a mitigation of a particular threat, as connections to and from a malicious host are suppressed and no command & control activity or an initial infection should happen anymore. Based this positive impact, an application of the threat information can also have negative consequences, especially if the information is not specific enough or contains false positives.
The former is particularly of concern if feeds only provide IP address information, such as the IP address a command & control server is currently hosted at. While in times of domain-generation algorithms (DGAs) indicators such as domain names have an extremely short lifetime, in many circumstances an actor will not host malicious infrastructure on a dedicated machine, but rather employ the services of commercial vendors as this offer much higher flexibility and incurs no loss (expect for the forfeiture of prepaid service) such as the seizure of own hardware. This however also means that at particular IP address that is flagged as malicious other services may be present which are then also blocked as collateral damage.
Our metric impact measures the consequences to an organization if the information from a threat intelligence feed is applied, for example by blocking IP addresses in the firewall. This can have both positive and negative consequences, and we care whether all of the malicious activity will be suppressed given the feed's data, and whether it *only* covers malicious activity or the application will also cause harm to benign services. For example, if a malware communicates with its C&C server using 10 IP addresses, the blockage is only really successful and useful if all 10 addresses are included in the feed as otherwise the activity simply continues using an alternative channel, and only these 10 addresses are blocked.

4 DATASETS

Goal of this paper is to evaluate the quality of cyber threat intelligence feeds, which we will do based on the criteria described in the previous section. For this purpose, we have monitored a total of 24 open source feeds which blacklist domain names as well as IP addresses based on detected malicious activities, annotated into major categories such as botnet C&C server activity, usage as a phishing domains etc. These feeds were continuously monitored over a period of 7 months from August 1, 2018 until February 28, 2019, and when available also all historical records back until January 1, 2018. This yielded a total of 1,383,040 indicators that we are going to use for this evaluation.

For our analysis, we monitored 17 threat intelligence feeds over a period of 14 months, and 7 feeds over a period of 7 months. In table 1, we will briefly enumerate each of the feeds included in this analysis.

In order to evaluate the available cyber threat intelligence feeds with respect to timeliness, accuracy, completeness and relevance, we make use of two auxiliary datasets:

- **Active domain crawls** Based on zone transfers on registered domains from ICANN and national domain registries, we have crawled approximately 277 million unique domains across 1151 generic and country code top level domains on a daily basis. This data shows which IP address was connected to which domain at any given day.
- **NetFlows of a tier 1 operator** To detect whether an IP address is actually receiving traffic or not, and to investigate the response of networks to a blacklisting, we leverage NetFlow data collected at the backbone of a tier 1 network operator. These NetFlows were recorded at each of the operator's core routers at a sampling rate of 8192:1 and thus allowed the reconstruction of activity towards specific IP addresses. We provided a list of IP addresses flagged as malicious by the threat intelligence feeds, and received an anonymized list of IP addresses that connected to the suspicious targets. In order to preserve the privacy of the customers, the IP addresses of the clients were anonymized by the ISP using the technique described in [6] and obfuscated at the level of autonomous systems. This allowed to quantify the activity of malicious endpoints without learning anything of about identity of the actual users.

4.1 Anonymization

In order to preserve the privacy of users in the NetFlow dataset, the IP addresses of senders and receivers are randomized to mask their identity. While for NetFlow datasets only a deterministic,

random one-to-one mapping of original to anonymized IP addresses is necessary to match outgoing requests with returning answers, in such blind randomization the relationship information of networks is lost. Thus, it is not possible to preserve information locality information such as a C&C activity realized by several hosts in the same /24 subnet, as these hosts would be scattered across the entire IPv4 space.

In this paper, we use the method introduced by Xu et al. [13], who introduce a random one-to-one mapping while preserving network information. If we represent network addresses in a binary tree which each bit of the IP address when read left to right will result in a transition to the left or right subtree under a node, an IP block under a shared prefix will be expressed as an entire subtree under one specific node. Consider the example in 1(a), all IP addresses in the prefix P_1 start with the digits "00' in their address, while IP addresses in the adjacent address block begin with "01". Under each leaf node – which are marked in grey – are then all IP addresses associated with this particular IP allocation.

In Xu et al.'s "Cryptography-based Prefix-preserving Anonymization" (Crypto-PAn) [13] scheme, the bit value of every non-leaf node is flipped randomly. This means that if two IP addresses shared a k-bit prefix, also the anonymized IP addresses will share an identical, but now randomized k-bit prefix. Within each netblock, IP addresses can now be scrambled without loosing information about the logical coherence of the addresses to one provider, and prefix-preserving anonymization comes in handy for the evaluation of threat intelligence feeds as related activity is often located in adjacent IP addresses or subnet blocks as we will show later. The randomness in Crypto-PAn is drawn from the AES block cipher, and a short encryption key is thus sufficient to provide an effective IP randomization function. The authors prove in [13] that this scheme delivers semantic security.

The anonymization of NetFlows was done on site of the Tier1 operator using a secret key chosen by the operator, so that only obfuscated data was analyzed within the context of this project, thus preserving the identity of Internet users. In order to match the information on malicious activity from the threat intelligence feeds to the traffic patterns in the NetFlow dataset, the operator additionally provided us with a lookup table of the malicious IP addresses from the feeds to the anonymized counterpart in the dataset to enable the analysis presented in the remainder of this paper.

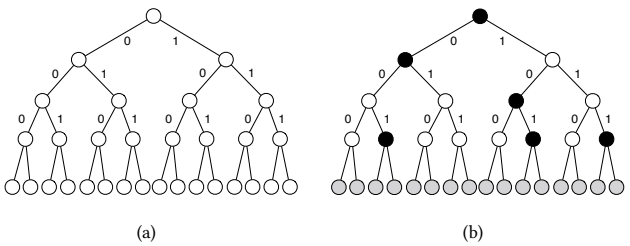

Figure 1: Prefix-preserving randomization after Xu et al. [13]

5 QUALITY EVALUATION OF FEEDS

Based on the criteria introduced in section 3, in this section we discuss the results of the quality evaluation of the 24 tested cyber threat intelligence feeds. The following subsections will first review their performance in terms of timeliness, sensitivity, originality and impact, before in section 6 we will in further detail analyze the question of their overall utility and adoption in practice.

5.1 Timeliness

In this section, we are assessing the timeliness of cyber threat intelligence feeds based on the amount of traffic a particular destination has received, prior and after it was included in the analyzed feeds.

To visualize the process, figure 2 depicts connections within the Tier1 network to seven example destination IP addresses between July 2018 and January 2019 that were in the second half of 2018 flagged as malicious. For each day, we aggregated flows from distinct clients towards each destination, the size of each circle shows in logarithmic scale the total number of recorded flows. Note that the IP addresses are anonymized as discussed in section 4: while the anonymization protocol matched the feed indicators to the IP addresses, the shown IP addresses are randomized at the level of prefixes. Thus, no conclusion can be taken about the concrete IP addresses at hand or their location in the world.

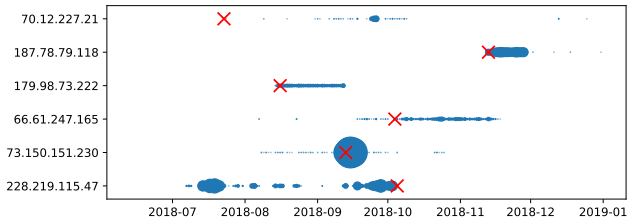

Figure 2: Scatter plot of netflow activity. The size of the line shows the amount of traffic observed. Crosses denote when the IP address was blacklisted.

As we see in the graph, we find that activity on IP addresses and their appearance in intelligence feeds frequently diverges significantly in practice. The first three IP addresses in the figure are examples of a very timely detection – the IP addresses are reported as soon as the first activity arises, and in the first case even months before significant botnet traffic appears towards this C&C server. Not every intelligence report is however as successful. In the fourth and fifth case, the IP addresses are active for several weeks prior to reporting, and in case of 73.150.151.230 it is only marked as malicious after a significant traffic volume emerges. An even worse outcome is shown just below in case of 228.219.115.47: while after the including of the IP address in the threat feeds activity abruptly stops, the IP address had been active for almost 3 months prior, and been engaged in thousands of connections with clients.

While we could simply mark the onset of a significant number of flows to a flagged IP as the beginning of the illicit activity, this procedure would result in overestimations for example when IP addressed used legitimately before are reallocated by a provider to a customer that starts to abuse them. We can however identify the

Table 1: List of evaluated open source feeds

TI Feed	Automated	Period	Amount of IPs
Badips	Yes	14 months	95
Bambenek	Yes	14 months	1,796
Blocklist.de	Hybrid	14 months	944,622
BotScout Bot List	No	14 months	1,564
Botvrij	No	14 months	95
BruteForceBlocker	No	14 months	4,663
CI Army IP	Hybrid	14 months	181,439
CINSscore	Hybrid	14 months	250,153
Charles the Haleys	No	7 months	38,999
Cruzit	No	7 months	49,911
Danger.rulez	No	7 months	3,099
Dshield	No	14 months	106
Emerging Threats	No	14 months	10,464
Greensnow	No	14 months	116,748
MalwareConfig	Yes	14 months	19
Malwaredomainlist	Yes	14 months	1,011
Myip	No	7 months	55,936
Nothink	Yes	7 months	42
Phishtank	Yes	14 months	2,708
Ransomwaretracker	Hybrid	14 months	383
Rutgers	Yes	14 months	112,898
Talos	Hybrid	7 months	2683
Tech. Blogs and Reports	Yes	14 months	6,151
Zeustracker	Yes	7 months	112

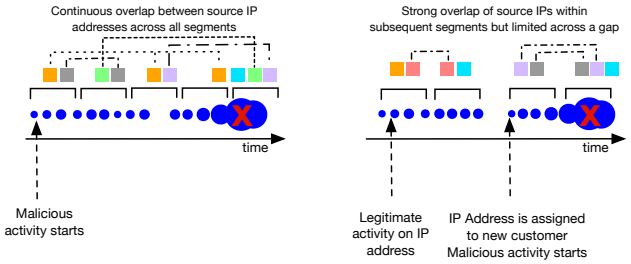

Figure 3: By comparing the set of IP addresses that are connecting to an IP address flagged as malicious, we can approximate the start of the malicious activity.

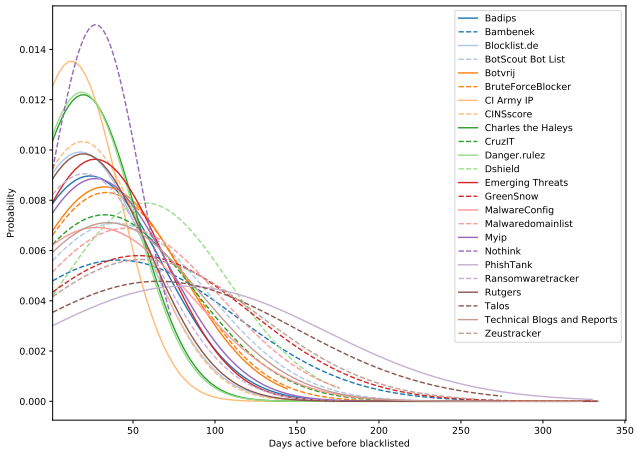

Figure 4: PDF of the times it takes a list to blacklist a host after it became active. From the plot can be seen that hosts are routinely active for multiple weeks before a destination is marked as malicious by threat intelligence feeds.

starting point and transition phases such as IP churn or hacked hosts by comparing the sets of client IP addresses that connect to a flagged destination. Intuitively, we exploit here that a server running as a reporting point for ransomware or as a command & control instance for bots or would be contacted in regular intervals by infected clients [4], while regular website would attract a more diverse group of visitors that would make a connection at unspecific times than the same set of clients coming back in the similar regular intervals. In the left part of figure 3, there exists a large overlap of the client IP addresses making contact for the entire period beginning the start of traffic and the IP being reported malicious. In the right part, we can identify a break in this pattern, with a consecutive sequence of windows showing significant overlap until being marked, while earlier activity shows only insignificant overlap with this period. When doing this analysis for the IP addresses flagged as malicious by the feeds, we can thereby obtain a conservative lower rather than a loose upper bound for the start of the malicious activity.

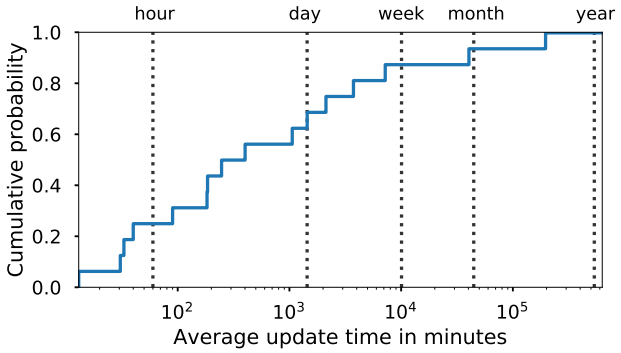

Figure 5: CDF of update frequency of the evaluated lists. Two thirds of the threat intelligence feeds are updated a least once a day, one thirds even includes new indicators in hourly intervals.

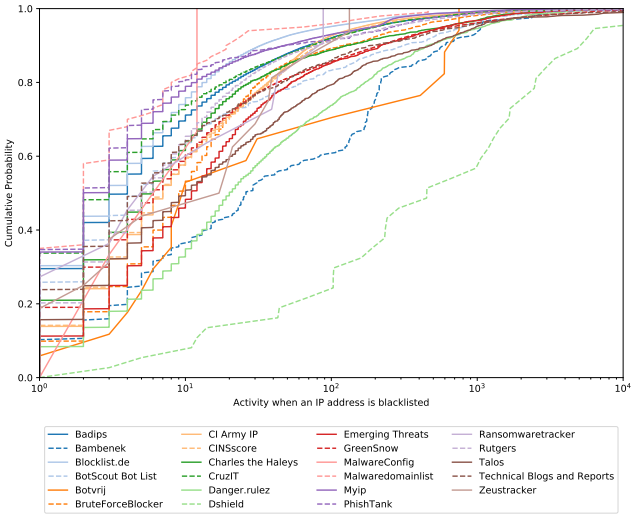

Figure 6: Cumulative density function of the minimum activity before an IP addresses is included in a threat intelligence feed.

We conducted the analysis of timeliness for all 1,383,040 indicators across the 24 threat intelligence feeds and counted the number of consecutive days IP addresses have received activity from clients before they were included in a particular list. Based on this analysis, we find that the first examples of successful indication in figure 2 are the exception rather than the norm, surprisingly we find that it takes on average 21 days before indicators are included in a list.

Figure 4 splits this analysis out, where each curve shows in a probability density function the number of active days until listed by an individual provider. As can be seen in the graph, a handful of feeds are clearly leading the pack, with the response time of CI Army being about 50% better than the overall average. In the bulk of the feeds, we find surprisingly high homogeneity and overall slow inclusion of malicious sources into the feeds, with turn around times of on average approximately one month. Even lists commonly praised by practitioners as "high quality" or "industry standards", such as the widely used *Emerging Threats* score surprisingly average in this respect. At the lower end of the scale, we have already observed activity for on average 65 to 80 days, before the slowest to respond feeds – Talos and PhishTank – include these IP address in their reports as malicious.

This lag between the emergence of malicious activity and the inclusion in the threat intelligence feeds might be due to a slow update frequency. To investigate this hypothesis, we analyzed the inter-arrival time when information was included across the 24 feeds, figure 5 shows a cumulative density function of the time in between list updates. As we can see from the graph, information is pushed at a very high frequency to the portfolio of lists, in one third of the cases updates occur at least hourly, while approximately two thirds of items are updated at a granularity of at least once per day. Thus, it is not the processing of the lists where reporting latency occurs, but during the selection and preparation of indicators.

5.2 Sensitivity

As we have already seen above, there seems to be a significant deviation between feeds in how soon indicators are included after the first sign of network activity. Figure 6 lists a cumulative density function of the number of connections we observe before an address is included in a particular intelligence feed as malicious. While the majority of lists is surprisingly homogeneous in their sensivity – we see that the bulk of them triggers with 50% likelihood if at least 6*8192 - 10*8192 flows are recorded –, also here many drastic outliers emerge. Dshield's sensitivity is across the board 1.5 - 2 orders of magnitude lower than the rest, here indicators are almost exclusively listed only when they show major activity. When we refer back to figure 4, we notice this feed to also perform sub-average with respect to timeliness. Other feeds that are also not very sensitive, like Danger.rulez, have better timeliness.

Sensitivity is however not only dependent on the types of sensors a threat intelligence provider utilizes, but also where these sensors are located. Threats emerging a specific geographic areas might hence be under- or overestimated, leading to an overall bias in sensitivty. As there is no ground truth on where in the world threats are actually located, we can only do a relative evaluation on the position of the listed indicators – based on their IP prefix information – for each individual intelligence feed. Figure 7 shows this relative geographical distribution of indicator per feed, which clearly reveal major differences in reporting between the providers and likely the location of their sensor infrastructure. For instance, more than 40% of all reports made by Bambenek are located in the United States, whereas more than 40% of reports on MalwareConfig originate from Turkey and around 40% of the data provided by GreenSnow relates to the China.

These biases become even more apparent when we normalize the IP addresses reported as malicious by the number of IP addresses allocated within that region. Assuming that malicious activity is not strongly concentrated within individual countries, we thus obtain a normalized geographical reporting as shown in figure 8, this shows that for example CI Army and CINSscore are heavily leaning towards reports from Turkmenistan, which is nearly entirely absent in the reports from all the other threat intelligence providers.

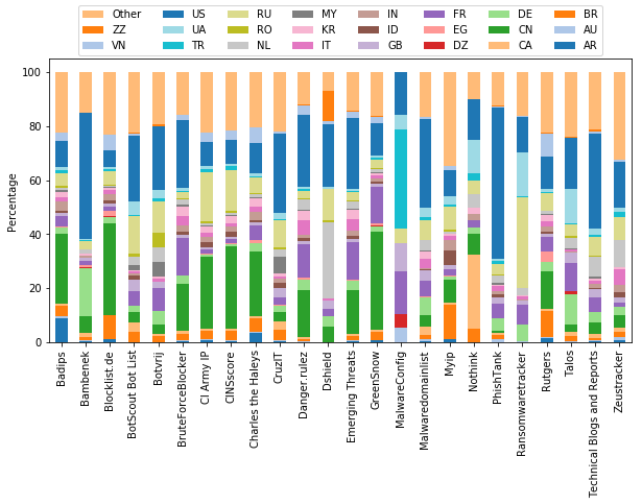

Figure 7: Geographical distribution of indicators per list.

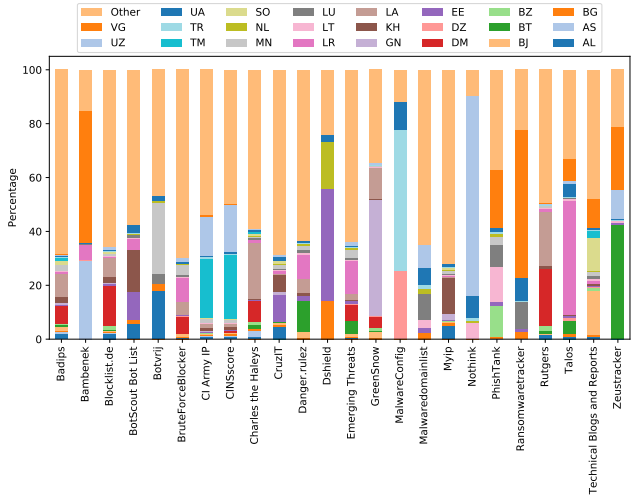

Figure 8: Relative geographical distribution of indicators, normalized by total number of IP addresses in a country.

On a positive note, the distribution of reports by Emerging Threats, Badips, Blocklist.de, BruteForceBlocker, CruzIT and Myip show no clear geographical preference, and which seems to lend to the conclusion that their measurement infrastructure is sufficiently diverse.

5.3 Originality

To investigate the uniqueness of the provided information, we traced for each of the 1.38 million indicators when it was first emerging on a particular list and whether individual indicators were afterwards also included on other lists. Besides the result of independent original research, such reuse might also indicate that a particular feed would important the data provided by others. Figure 9 shows the later reuse of indicator information across lists, where

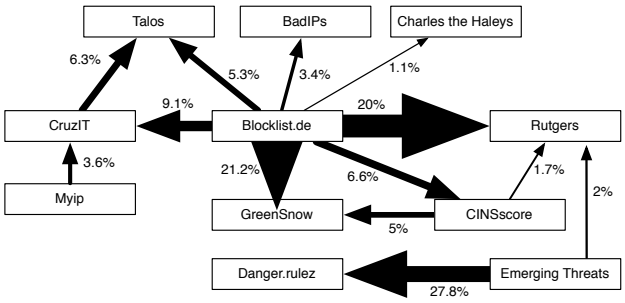

Figure 9: Indicator reuse across feeds occurs only sporadically, with the expection of two threat intelligence feeds.

an arrow indicates that information first originated at the source of the arrow and was later included in the list its points to. The thickness of the arrow and the corresponding label corresponds with the percentage of information on the receiving list, that earlier appeared somewhere else. For readability, only flows where more than 5% potentially originated from a different list are shown.

While commercial threat intelligence providers often only consolidate and curate information as discussed above, we see also some repackaging – although at highly varying degrees – in case of open source feeds. The indicators first appearing on the feed Blocklist.de routinely appear on other lists, and in two cases a fifth of all indicators are shared with Blocklist.de where they appear earlier. Similarly, a quarter of the indicators on Danger.rulez previously appeared on Emerging Threats, however at a global scale such repacking is comparatively seldom and only 11 out of the 24 lists showed such relationships at all. Across the entire dataset, indicators reappear comparatively seldom, in total only 85,906 out of the total 1.38 million entries were also listed on another feed (6.2%).

5.4 Impact

In order to evaluate the level of potential collateral damage, we resolved the IPv4 and IPv6 A records of 277 million domain name across across 1151 top-level domain zones that we received from the TLD operators on a daily basis. For every threat intelligence feed, we then analyze how many domain names were pointing to a particular IP address on the day it was marked as malicious, as all of these domain names would no longer be resolvable if a customer would apply the ruleset provided by the threat intelligence provider in for example a firewall. Figure 10 shows the cumulative density function of the number of domain names resolving to the indicated IP addresses by threat feed.

As we can see in the graph, there are drastic differences in the amount of collateral damage between feeds. A homogeneous set of feeds – among them BruteForceBlocker, Talos, CruzIT, CI Army IP, and Rutgers – are comparatively targeted, more than half of their entries are not affected any other domain names, while the 80% most targeted indicators affect less than 6 other domain names if applying an IP-based block. This is somewhat logical for a list that focuses on bruteforcing, which is typically not happening from servers that host websites or are operated by a shared web hoster,

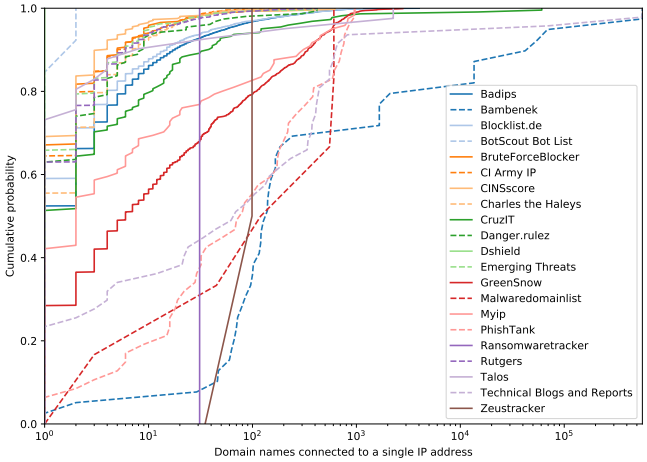

Figure 10: Cumulative probability function of the domain names associated with the IP addresses indicated as malicious per threat intelligence feed.

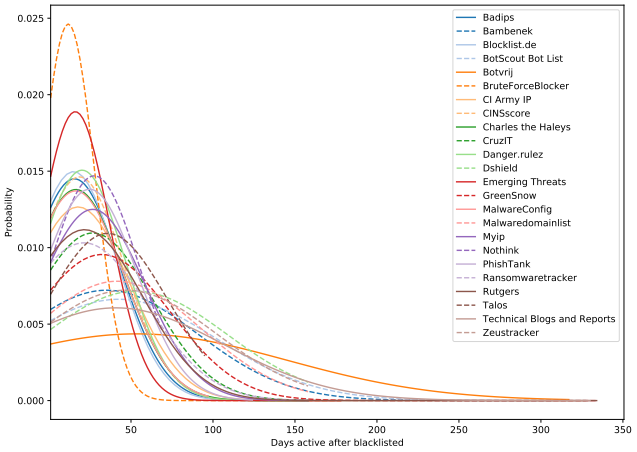

Figure 11: Continuation of activity to flagged destinations after listing in a feed.

this is however not the case for the information included on CruzIT, CI Army IP or Rutgers, which include IP addresses used to attack or probe certain networks. For Talos we know that it is curated by Cisco, which makes it likely that this is the reason for the low amount of collateral damage.

This is however not true for all of the feeds. In case of Bambenek, only 16% of the best performing indicators will block less than 50 live domains hosted at these websites, where the 50% worst performing indicators even affect 100 or more domains as collateral damage. While some of the blocked domains may certainly also contain malicious activity, some of the instances included large shared hosters, in one case with more than 900,000 domains pointing to the blacklisted IP address. While such issues could be explained due to automatic collection of indicators, we also found a surprisingly poor track record in case of "Technical Blogs and Reports", a curated list of human analyst reports. This indicates that human-made feeds are not necessarily better, or at least suggests that feeds do not filter records that could potentially be harmful to normal system operation.

6 DISCUSSION

After an evaluation of each of the 24 feeds across the four dimensions timeliness, sensitivity, originality and impact, we will take a step back in this section and evaluate the ecosytem of intelligence feeds as a whole. Specifically we will investigate how widely these feeds are adopted by network owners and operators in practice, and review the issue of the surprisingly high level of originality for the ecosystem as a whole.

6.1 Adoption of Intelligence Feeds

As cyber threat intelligence feeds are meant to alert and empower network owners to block malicious activity, their application should lead to a reduction in network traffic towards the hosts flagged as malicious. This observation provides us with an angle to investigate the adoption of threat intelligence feeds worldwide. After all, as soon as an indicator has appeared on a list we would expect a significant drop of activity – if not the absence of requests – from subscribers.

When networks apply the information provided in cyber threat intelligence feeds at scale, we should ideally see the activity from infected clients to malicious destinations drop and eventually die out. As we have however seen above, intelligence feeds are not universally adopted but have a specific regional footprint and there exists only a marginal overlap between lists; thus, even if a network would subscribe to and apply the information from every single intelligence feed we cannot expect all activity to immediately seize.

Figure 11 shows a probability density function of how long activity towards a particular destination continued after an indicator was listed on a particular intelligence feed. As we see from the graph, the inclusion of indicators on for example BruteForceBlocker seems to be an effective deterrent, as on average activity stops within 2 weeks time. Other lists such as Botvrij are less successful: more than 50% of all hosts reported as malicious by this feed continue their activity for at least 79 days, with an extremely long trail, thus a listing on this block list seems to have almost no impact on the criminal activity itself. Like in case of timeliness until detection (see section 5.1), the threat intelligence feeds are also surprisingly homogeneous with respect to the continuation of activities, and we can clearly see in figure 11 two main clusters, with activity termination peaking around 20 days after listing and 60 days after reporting.

6.2 Do we have enough coverage?

There remain however questions about the quality of the ecosystem of cyber threat intelligence providers as a whole. Although it is desirable for a customer that CTI feeds have a large degree of originality as otherwise a customer would subscribe – and pay for – redundant information, we have seen that the amount of overlap between the entire spectrum of analyzed feeds was actually remarkably low. This on the one hand is commendable as it maximizes value of CTI users, on the other hand it also raises questions

whether the cyber threat intelligence feeds really provide sufficient information to stop malicious activity in their tracks.

As discussed in section 4, the 24 evaluated open source feeds spanned the entire ecosystem of malicious activity from bruteforcing activity, ransomware and other malware, to botnets. As each type of malicious activity was covered by multiple feeds, we actually would expect *some* overlap in reported indicators. The fact that there is almost no overlap between lists of similar scope could be the result of two reasons: first, all individual lists for example targeting botnets or ransomware rely on orthogonal detection methods and are therefore providing complementary information. Second, the lists monitor malicious activity in comparable ways, but the overall volume of malicious activity is so large that effectively each lists only obtains a tiny sample, and such low rate sampling from a large universe would statistically lead to a very low chance for duplicates.

Conceptually, we can see that we are probably dealing with the latter than the former reason, after all methods to for example detect ransomware C&C servers are limited, and most likely all providers would employ off-the-shelf tools such as an analysis of network activity across malware samples or a forensic analysis thereof. This unfortunately drives us to the conclusion that cyber threat intelligence feeds cover much less of malicious activity than we would expect and require, to apply intelligence feeds and confidently expect that with a very high degree of certainty malicious activity will be stopped through these indicators.

7 CONCLUSIONS

Effective protection requires insights into the activities of adversaries, commonly referred to as cyber threat intelligence. In order to protect against evolving threats, networks can subscribe to CTI feeds which list indicators, such as domain names, IP addresses, or hashes, related to malicious activity.

In this paper, we have analyzed 1.38 million indicators provided by 24 open source cyber threat intelligence feeds over a period of 14 months, and analyzed whether the information provided by these lists is timely, original, and estimated how sensitive the detection of the intelligence providers are as well as the positive and negative impacts a utilization of these feeds would have in practice. We find large variations between the performance of these lists, some are providing indicators within a few days while others only report activity months after it has commenced. This variation is surprising as we find all feeds to be relatively homogeneous in sensitivity, in other words the threshold beyond which they pick up undesired activity.

Nearly all of the analyzed lists are able to provide a significant intelligence contribution. Although lists contain a small degree of overlap, these lists are not merely subsets or repackaged versions of each other. This on the one hand is valuable to defenders as each feed does provide benefit with limited redundancy, at the same time the little overlap across all CTI feeds raises the question how much the ecosystem of cyber threat intelligence feeds as a whole really covers, and whether the current feeds will provide defenders with a suitable defense posture if applied.

REFERENCES

[1] [n. d.]. Evaluating Threat Intelligence Feeds. ([n. d.]). Retrieved from: https://www.first.org/resources/papers/munich2016/kompanek-pawlinski-evaluating-threat-ntelligence-feeds.pdf.

[2] [n. d.]. Measuring the IQ of your Threat Intelligence. ([n. d.]). Retrieved from: https://www.slideshare.net/AlexandrePinto10/defcon-22-measuring-the.

[3] [n. d.]. The Second Annual Study on Exchanging Cyber Threat Intelligence: There Has to Be a Better Way. ([n. d.]). Retrieved from: https://www.ponemon.org/blog/the-second-annual-study-on-exchanging-cyber-threat-intelligence-there-has-to-be-a-better-way.

[4] Jasper Abbink and Christian Doerr. 2017. Popularity-Based Detection of Domain Generation Algorithms. In *2nd International Workshop on Malware Analysis*.

[5] EclecticIQ. [n. d.]. Intelligence-powered Defences. ([n. d.]). https://www.eclecticiq.com/dss

[6] Michalis Foukarakis, Demetres Antoniades, Spiros Antonatos, and Evangelos P. Markato. 2007. Flexible and high-performance anonymization of NetFlow records using anontool. In *Third International Conference on Security and Privacy in Communications Networks and the Workshops*.

[7] Marc Kührer, Christian Rossow, and Thorsten Holz. 2014. Paint it black: Evaluating the effectiveness of malware blacklists. In *International Workshop on Recent Advances in Intrusion Detection*. Springer, 1–21.

[8] Adam Oest, Yeganeh Safaei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Kevin Tyers. 2019. PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques Against Browser Phishing Blacklists. In *PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques against Browser Phishing Blacklists*. IEEE, 0.

[9] Anirudh Ramachandran, Nick Feamster, and Santosh Vempala. 2007. Filtering spam with behavioral blacklisting. In *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 342–351.

[10] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong, and Chengshan Zhang. 2009. An empirical analysis of phishing blacklists. In *Sixth Conference on Email and Anti-Spam (CEAS)*. California, USA.

[11] Sushant Sinha, Michael Bailey, and Farnam Jahanian. 2008. Shades of Grey: On the effectiveness of reputation-based "blacklists". In *2008 3rd International Conference on Malicious and Unwanted Software (MALWARE)*. IEEE, 57–64.

[12] Wiem Tounsi and Helmi Rais. 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & security* 72 (2018), 212–233.

[13] Jun Xu, Jinliang Fan, Mostafa Ammar, and Sue B. Moon. 2001. On the Design and Performance of Prefix-preserving IP Traffic Trace Anonymization. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement (IMW '01)*.