Reading the Tea Leaves: A Comparative Analysis of Threat Intelligence

Vector Guo Li[1], *Matthew Dunn*[2], *Paul Pearce*[4], *Damon McCoy*[3],
Geoffrey M. Voelker[1], *Stefan Savage*[1], *Kirill Levchenko*[5]

[1] *University of California, San Diego* [2] *Northeastern University* [3] *New York University*
[4] *Georgia Institute of Technology* [5] *University of Illinois Urbana-Champaign*

Abstract

The term "threat intelligence" has swiftly become a staple buzzword in the computer security industry. The entirely reasonable premise is that, by compiling up-to-date information about known threats (i.e., IP addresses, domain names, file hashes, etc.), recipients of such information may be able to better defend their systems from future attacks. Thus, today a wide array of public and commercial sources distribute threat intelligence data feeds to support this purpose. However, our understanding of this data, its characterization and the extent to which it can meaningfully support its intended uses, is still quite limited. In this paper, we address these gaps by formally defining a set of metrics for characterizing threat intelligence data feeds and using these measures to systematically characterize a broad range of public and commercial sources. Further, we ground our quantitative assessments using external measurements to qualitatively investigate issues of coverage and accuracy. Unfortunately, our measurement results suggest that there are significant limitations and challenges in using existing threat intelligence data for its purported goals.

1 Introduction

Computer security is an inherently adversarial discipline in which each "side" seeks to exploit the assumptions and limitations of the other. Attackers rely on exploiting knowledge of vulnerabilities, configuration errors or operational lapses in order to penetrate targeted systems, while defenders in turn seek to improve their resistance to such attacks by better understanding the nature of contemporary threats and the technical fingerprints left by attacker's craft. Invariably, this means that attackers are driven to innovate and diversify while defenders, in response, must continually monitor for such changes and update their operational security practices accordingly. This dynamic is present in virtually every aspect of the operational security landscape, from anti-virus signatures to the configuration of firewalls and intrusion detection systems to incident response and triage. Common to all such reifications, however, is the process of monitoring for new data on attacker behavior and using that data to update defenses and security practices. Indeed, the extent to which a defender is able to gather and analyze such data effectively defines a de facto window of vulnerability—the time during which an organization is less effective in addressing attacks due to ignorance of current attacker behaviors.

This abstract problem has given rise to a concrete demand for contemporary threat data sources that are frequently collectively referred to as *threat intelligence* (**TI**). By far the most common form of such data are so-called *indicators of compromise:* simple observable behaviors that signal that a host or network may be compromised. These include both network indicators such as IP addresses (e.g., addresses known to launch particular attacks or host command-and-control sites, etc.) and file hashes (e.g., indicating a file or executable known to be associated with a particular variety of malware). The presence of such indicators is a symptom that alerts an organization to a problem, and part of an organization's defenses might reasonably include monitoring its assets for such indicators to detect and mitigate potential compromises as they occur.

While each organization naturally collects a certain amount of threat intelligence data on its own (e.g., the attacks they repel, the e-mail spam they filter, etc.) any single entity has a limited footprint and few are instrumented to carefully segregate crisp signals of attacks from the range of ambiguity found in normal production network and system logs. Thus, it is now commonly accepted that threat intelligence data procurement is a specialized activity whereby third-party firms, and/or collections of public groups, employ a range of monitoring techniques to aggregate, filter and curate quality information about current threats. Indeed, the promised operational value of threat intelligence has created a thriving (multi-billion dollar) market [43]. Established security companies with roots in anti-virus software or network intrusion detection now offer threat intelligence for sale, while some vendors specialize in threat intelligence exclusively, often promising coverage of more sophisticated threats than conventional sources.

Unfortunately, in spite of this tremendous promise, there has been little empirical assessment of threat intelligence data

or even a consensus about what such an evaluation would entail. Thus, consumers of TI products have limited means to compare offerings or to factor the cost of such products into any model of the benefit to operational security that might be offered.

This issue motivates our work to provide a grounded, empirical footing for addressing such questions. In particular, this paper makes the following contributions:

- We introduce a set of basic *threat intelligence metrics* and describe a methodology for measuring them, notably: **Volume**, **Differential Contribution**, **Exclusive Contribution**, **Latency**, **Coverage** and **Accuracy**.
- We analyze 47 distinct IP address TI sources covering six categories of threats and 8 distinct malware file hash TI sources, and report their metrics.
- We demonstrate techniques to evaluate the accuracy and coverage of certain categories of TI sources.
- We conduct the analyses in two different time periods two years apart, and demonstrate the strong consistency between the findings.

From our analysis, we find that while a few TI data sources show significant overlap, most do not. This result is consistent with the hypothesis advanced by [42] that different kinds of monitoring infrastructure will capture different kinds of attacks, but we have demonstrated it in a much broader context. We also reveal that underlying this issue are broader limitations of TI sources in terms of coverage (most indicators are unique) and accuracy (false positives may limit how such data can be used operationally). Finally, we present a longitudinal analysis suggesting that these findings are consistent over time.

2 Overview

The threat intelligence data collected for our study was obtained by subscribing to and pulling from numerous public and private intelligence sources. These sources ranged from simple blacklists of bad IPs/domains and file hashes, to rich threat intelligence exchanges with well labeled and structured data. We call each item (e.g., IP address or file hash) an *indicator* (after *indicator of compromise*, the industry term for such data items).

In this section we enumerate our threat intelligence sources, describe each source's structure and how we collected it, and then define our measurement metrics for empirically measuring these sources. When the source of the data is public, or when we have an explicit agreement to identify the provider, we have done so. However, in other cases, the data was provided on the condition of anonymity and we restrict ourself to describing the nature of the provider, but not their identity. All of our private data providers were appraised of the nature of our research, its goals and the methodology that we planned to employ.

2.1 Data Set and Collection

We use several sources of TI data for our analysis:

Facebook ThreatExchange (FB) [17]. This is a closed-community platform that allows hundreds of companies and organizations to share and interact with various types of labeled threat data. As part of an agreement with Facebook, we collected all its data that it shared broadly. In subsequent analyses, sources with prefix "FB" indicate a unique contributor on the Facebook ThreatExchange.

Paid Feed Aggregator (PA). This is a commercial paid threat intelligence data aggregation platform. It contains data collected from over a hundred other threat intelligence sources, public or private, together with its own threat data. In subsequent analyses all data sources with prefix "PA" are from unique data sources originating from this aggregator.

Paid IP Reputation Service. This commercial service provides an hourly-updated blacklist of known bad IP addresses across different attack categories.

Public Blacklists and Reputation Feeds. We collected indicators from public blacklists and reputation data sources, including well-known sources such as AlienVault [3], Badips [5], Abuse.ch [1] and Packetmail [28].

Threat Intelligence indicators include different types of data, such as IP address, malicious file hash, Domain, URL, etc. In this paper, we focus our analysis on sources that provide IP addresses and file hashes, as they are the most prevalent data types in our collection.

We collect data from all sources on an hourly basis. However, both the Facebook ThreatExchange and the Paid Feed Aggregator change their members and contributions over time, creating irregular collection periods for several of the sub-data sources. Similarly, public threat feeds had varying degrees of reliability, resulting in collection gaps. In this paper, we use the time window from **December 1, 2017** to **July 20, 2018** for most of the analyses, as we have the largest number of active sources during this period. We eliminated duplicates sources (e.g., sources we collected individually and also found in the Paid Aggregator) and sub-sources (a source that is a branch of another source). We further break IP sources into separate categories and treat them as individual feeds, as shown in Section 3. This filtering leaves us with 47 IP feeds and 8 malware file hash feeds.

The ways each TI source collects data varies, and in some cases the methodology is unknown. For example, Packetmail IPs and Paid IP Reputation collect threat data themselves via honeypots, analyzing malware, etc. Other sources, such as Badips or the Facebook ThreatExchange, collect their indicators from general users or organizations—e.g., entities may be attacked and submit the indicators to these threat intelligence services. These services then aggregate the data and report it to their subscribers. Through this level of aggregation the precise collection methodologies and data providence can be lost.

2.2 Data Source Structure

TI sources in our corpus structure and present data in different ways. Part of the challenge in producing cross-dataset metrics is normalizing both the structure of the data as well as its *meaning*. A major structural difference that influences our analysis occurs between data sources that provide data in *snapshots* and data sources that provide *events*.

Snapshot. Snapshot feeds provide periodic snapshots of a set of indicators. More formally, a snapshot is a set of indicators that is a function of time. It defines, for a given point in time, the set of indicators that are members of the data source. Snapshot feeds imply *state*: at any given time, there is a set of indicators that are *in* the feed. A typical snapshot source is a published list of IPs periodically updated by its maintainer. For example, a list of command-and-control IP addresses for a botnet may be published as a snapshot feed subject to periodic updates.

All feeds of file hashes are snapshots and are *monotonic* in the sense that indicators are only added, not removed, from the feed. Hashes are a proxy for the file content, which does not change (malicious file content will not change to benign in the future).

Event. In contrast, event feeds report newly discovered indicators. More formally, an event source is a set of indicators that is a function of a time *interval.* For a given time interval, the source provides a set of indicators that were seen or discovered in that time interval. Subscribers of these feeds query data by asking for new indicators added in a recent time window. For example, a user might, once a day, request the set of indicators that appeared in the last 24 hours.

This structural difference is a major challenge when evaluating feeds comparatively. We need to normalize the difference to make a fair comparison, especially for IP feeds. From a **TI** consumer's perspective, an *event* feed does not indicate when an indicator will expire, so it is up to the consumer to act on the age of indicators. Put another way, the expiration dates of indicators are decided by how users query the feed: if a user asks for the indicators seen in the last 30 days when quering data, then there is an implicit 30-day valid time window for these indicators.

In this paper, we choose a 30-day valid period for all the indicators we collected from event feeds—the same valid period used in several snapshot feeds, and also a common query window option offered by event feeds. We then convert these event feeds into snapshot feeds and evaluate all of them in a unified fashion.

2.3 Threat Intelligence Metrics

The aim of this work is to develop *threat intelligence metrics* that allow a **TI** consumer to compare threat intelligence sources and reason about their fitness for a particular purpose. To this end, we propose six concrete metrics: *Volume*, *Differential contribution*, *Exclusive contribution*, *Latency*, *Accuracy* and *Coverage*.

✧ **Volume.** We define the *volume* of a feed to be the total number of indicators appearing in a feed over the measurement interval. Volume is the simplest **TI** metric and has an established history in prior work [21,23,24,30,35,36,42]. It is also useful to study the daily *rate* of a feed, which quantifies the amount of data appearing in a feed on a daily basis.

Rationale: To a first approximation, volume captures how much information a feed provides to the consumer. For a feed without false positives (see *accuracy* below), and if every indicator has equal value to the consumer, we would prefer a feed of greater volume to a feed of lesser volume. Of course, indicators do not all have the same value to consumers: knowing the IP address of a host probing the entire Internet for decades-old vulnerabilities is less useful than the address of a scanner targeting organizations in your sector looking to exploit zero-day vulnerabilities.

✧ **Differential contribution.** The *differential contribution* of one feed with respect to another is the number of indicators in the first that are not in the second during the same measurement period. We define differential contribution relative to the size of the first feed, so that the differential contribution of feed A with respect to feed B is $\text{Diff}_{A,B} = |A \setminus B|/|A|$. Thus, $\text{Diff}_{A,B} = 1$ indicates that the two feeds have no elements in common, and $\text{Diff}_{A,B} = 0$ indicates that every indicator in A also appears in B. It is sometimes useful to consider the complement of differential contribution, namely the normalized *intersection* of A in B, given by $\text{Int}_{A,B} = |A \cap B|/|A| = 1 - \text{Diff}_{A,B}$.

Rationale: For a consumer, it is often useful to know how many *additional* indicators a feed offers relative to one or more feeds that the consumer has already. Thus, if a consumer already has feed A and is considering paying for feed B, then $\text{Diff}_{A,B}$ indicates how many new indicators feed A will provide.

✧ **Exclusive contribution.** The *exclusive contribution* of a feed with respect to a set of other feeds is the proportion of indicators unique to a feed, that is, the proportion of indicators that occur in the feed but no others. Formally, the exclusive contribution of feed A is defined as $\text{Uniq}_{A,B} = |A \setminus \bigcup_{B \neq A} B|/|A|$. Thus, $\text{Uniq}_{A,B} = 0$ means that every element of feed A appears in some other feeds, while $\text{Uniq}_{A,B} = 1$ means no element of A appears in any other feed.

Rationale: Like differential contribution, exclusive contribution tells a **TI** consumer how much of a feed is different. However, exclusive contribution compares a feed to all other feeds available for comparison, while differential contribution compares a feed to just another feed. From a **TI** consumer's perspective, exclusive contribution is a general measure of a feed's unique value.

✧ **Latency.** For an indicator that occurs in two or more feeds, its *latency* in a feed is the elapsed time between its first appearance in any feed and its appearance in the feed in question. In the feed where an indicator first appeared, its latency is zero. For all other feeds, the latency indicates how much later the

same indicators appears in those feeds. Taster's Choice [30] referred to latency as *relative first appearance time*. (We find the term *latency* to be more succinct without loss of clarity.) Since latency is defined for one indicator, for a feed it makes sense to consider statistics of the distribution of indicator latencies, such as the median indicator latency.

Rationale: Latency characterizes how quickly a feed includes new threats: the sooner a feed includes a threat, the more effective it is at helping consumers protect their systems. Indeed, several studies report on the impact of feed latency on its effectiveness at thwarting spam [10, 32].

The metrics above are defined without regard for the *meaning* of the indicators in a feed. We can calculate the volume of a single feed or the differential contribution of one feed with respect to another regardless of what the feed purports to contain. While these metrics are easy to compute, they do little to tell us about the fitness of a feed for a particular purpose. For this, we need to consider the meaning or purpose of the feed data, as advertised by the feed provider. We define the following two metrics.

✧ **Accuracy.** The *accuracy* of a feed is the proportion of indicators in a feed that are correctly included in the feed. Feed accuracy is analogous to *precision* in Information Retrieval. This metric presumes that the description of the feed is well-defined and describes a set of elements that should be in the feed given perfect knowledge. In practice, we have neither perfect knowledge nor a perfect description of what a feed should contain. In some cases, however, we can construct a set A^- of elements that should definitely not be in a feed A. Then $\text{Acc}_A \leq |A \setminus A^-|/|A|$.

Rationale: The accuracy metric tells a TI consumer how many false positives to expect when using a feed, and, therefore, dictates how a feed can be used. For example, if a consumer automatically blocks all traffic to IP addresses appearing in a feed, then false positives may cause disruption in an enterprise by blocking traffic to legitimate sites. On the other hand, consumers may tolerate some false positives if a feed is only used to gain additional insight during an investigation.

✧ **Coverage.** The *coverage* of a feed is the proportion of the intended indicators contained in a feed. Feed coverage is analogous to *recall* in Information Retrieval. Like accuracy, coverage presumes that the description of the feed is sufficient to determine which elements should be in a feed, given perfect knowledge. In some cases, it is possible to construct a set A^+ of elements that should be in a feed. We can then upper-bound the coverage $\text{Cov}_A \leq |A|/|A^+|$.

Rationale: For a feed consumer who aims to obtain complete protection from a specific kind of threat, coverage is a measure of how much protection a feed will provide. For example, an organization that wants to protect itself from a particular botnet will want to maximize its coverage of that botnet's command-and-control servers or infection vectors.

In the following two sections, we use these metrics to evaluate two types of TI: IP address feeds and file hash feeds.

3 IP Threat Intelligence

One of the most common forms of TI are feeds of IP addresses considered malicious, suspicious, or otherwise untrustworthy. This type of threat intelligence dates back at least to the early spam and intrusion detection blacklists, many of which are still active today such as SpamhausSBL [40], CBL [8] and SORBS [39]. Here, we apply the metrics described above to quantify the differences between 47 different IP address TI feeds.

3.1 Feed Categorization

IP address TI feeds have different meanings, and, therefore, purposes. To meaningfully compare feeds to each other, we first group feeds into *categories* of feeds whose indicators have the same intended meaning. Unfortunately, there is no standard or widely accepted taxonomy of IP TI feeds. To group feeds into semantic categories, we use metadata associated with the feed as well as descriptions of the feed provided by the producer, as described below.

Metadata. Some feeds provide category information with each indicator as metadata. More specifically, all of the Paid Aggregator feeds, Alienvault IP Reputation and Paid IP Reputation include this category metadata. In this case, we use its pre-assigned category in the feed. Facebook ThreatExchange feeds do not include category information in the metadata, but instead provide a descriptive phrase with each indicator. We then derive its category based on the description.

Feed description. For feeds without metadata, we rely on online descriptions of each feed, where available, to determine its semantic category. For example, the website of feed Nothink SSH [27] describes that the feed reports brute-force login attempts on its corresponding honeypot, which indicates the feed belongs to brute-force category.

We grouped our IP feeds into categories derived from the information above. In this work, we analyze six of the most prominent categories:

- **Scan:** Hosts doing port or vulnerability scans.
- **Brute-force:** Hosts making brute force login attempts.
- **Malware:** Malware C&C and distribution servers.
- **Exploit:** Hosts trying to remotely exploit vulnerabilities.
- **Botnet:** Compromised hosts belonging to a botnet.
- **Spam:** Hosts that sent spam or should not originate email.

Table 1 lists the feeds, grouped by category, used in the rest of this section. The symbols ○ and △ before the feed name indicate whether the feed is a snapshot feed or an event feed, respectively (see Section 2.2). All data was collected during our measurement period, **December 1st, 2017** to **July 20th, 2018**. Note that a few feeds, like Paid IP Reputation, appear in multiple categories. In these feeds, indicators are associated with different categories via attached metadata. We split these feeds into multiple virtual feeds each containing indicators belonging to the same category.

3.2 Volume

Volume is one of the oldest and simplest TI metrics representing how informative each data source is. Table 1 shows the total number of unique IP addresses collected from each feed during the measurement period, under column *Volume*. Feeds are listed in order of decreasing volume, grouped by category. The numbers we show are after the removal of invalid entries identified by the sources themselves. Column *Avg. Rate* shows the average number of new IPs we received per day, and *Avg. Size* lists the average daily working set size of each feed, that is, the average size of the snapshot.

✦ **Finding:** Feeds vary dramatically in volume. Within every category, big feeds can contain orders of magnitude more data than small feeds. For example, in the scan category, we saw over 361,004 unique IP addresses in DShield IPs but only 1,572 unique addresses in PA Analyst in the same time period. Clearly, volume is a major differentiator for feeds.

Average daily rate represents the amount of new indicators collected from a feed each day. Some feeds may have large volume but low daily rates, like Feodo IP Blacklist in the malware category. This means most indicators we get from that feed are old data present in the feed before our measurement started. On the other hand, the average rate of a feed could be greater than the volume would suggest, like Nothink SSH in the brute-force category. This is due to the fact that indicators can be added and removed multiple times in a feed. In general, IP indicators tend to be added in a feed only once: 37 among 47 IP feeds have over 80% of their indicators appearing only once, and 30 of them have this rate over 90%. One reason is that some snapshot feeds maintain a valid period for each indicator, as we found in all *PA* feeds where the expiration date of each indicator is explicitly recorded. When the same indicator is discovered again by a feed before its expiration time, the feed will just extend its expiration date, so this occurrence will not be captured if we simply subtract the old data from the newly collected data to derive what is added on a day. For event feeds and snapshot feeds in *PA* where we can precisely track every occurrence of each indicator, we further examed data occurrence frequency and still found that the vast majority of IPs in feeds only occurred once—an observation that relates to the dynamics of cyber threats themselves.

Nothink SSH, as we mentioned above, is a notable exception. It has over 64% of its indicators appearing 7 times in our data set. After investigating, we found that this feed posts all its previous data at the end of every month, behavior very likely due to the feed provider instead of the underlying threats.

The working set size defines the daily average amount of indicators users need to store in their system to use a feed (the storage cost of using a feed). The average working set size is largely decided by the valid period length of the indicators, controlled either by the feed (snapshot feeds) or the user (event feeds). The longer the valid period is, the larger the working set will be. Different snapshoot feeds have different choices for this valid period: PA AlienVault IPs in the scan category sets a 90-day valid period for every indicator added to the feed, while PA Abuse.ch Ransomware uses a 30-day period. Although we do not know the data expiration mechanism used by snapshot feeds other than *PA* feeds, as there is no related information recorded, we can still roughly estimate this by checking the *durations* of their indicators—the time

Table 1. IP TI feeds used in the study. A ○ denotes a *snapshot feed* and △ indicates an *event feed* (Section 2.2). *Volume* is the total number of IPs collected during our measurement period. *Exclusive* is the exclusive contribution of each feed (Section 3.4). *Avg. Rate* is the number of average daily new IPs added in the feed (Section 3.6), and *Avg. Size* is the average working set size of each feed (Section 3.2).

Feed	*Volume*	*Exclusive*	*Avg. Rate*	*Avg. Size*
Scan Feeds				
○ PA AlienVault IPs[1]	425,967	48.6%	1,359	128,821
△ DShield IPs	361,004	31.1%	1,556	69,526
○ PA Packetmail ramnode	258,719	62.0%	870	78,974
△ Packetmail IPs	246,920	48.6%	942	29,751
○ Paid IP Reputation	204,491	75.6%	1,362	8,756
○ PA Lab Scan	169,078	63.1%	869	9,775
○ PA Snort BlockList	19,085	96.3%	56	4,000
△ FB Aggregator$_1$	6,066	71.3%	24	693
○ PA Analyst	1,572	34.5%	6.3	462
Botnet Feeds				
○ PA Analyst	180,034	99.0%	697	54,800
○ PA CI Army	103,281	97.1%	332	30,388
○ Paid IP Reputation	77,600	99.9%	567	4,278
○ PA Botscout IPs	23,805	93.8%	81	7,180
○ PA VoIP Blacklist	10,712	88.0%	40	3,633
○ PA Compromised IPs	7,679	87.0%	21	2,392
○ PA Blocklist Bots	4,179	80.7%	16	1,160
○ PA Project Honeypot	2,600	86.5%	8.5	812
Brute-force Feeds				
△ Badips SSH	542,167	84.1%	2,379	86,677
△ Badips Badbots	91,553	70.8%	559	17,577
○ Paid IP Reputation	89,671	52.8%	483	3,705
○ PA Brute-Force	41,394	92.1%	138	14,540
△ Badips Username Notfound	37,198	54.2%	179	3662.8
△ Haley SSH	31,115	43.6%	40	1,224
△ FB Aggregator$_2$	22,398	77.3%	74	2,086
△ Nothink SSH	20,325	62.7%	224	12,577
△ Dangerrulez Brute	10,142	4.88%	37	1,102
Malware Feeds				
○ Paid IP Reputation	234,470	99.1%	1,113	22,569
△ FB Malicious IPs	30,728	99.9%	129	3,873
○ Feodo IP Blacklist	1,440	47.7%	1.3	1,159
○ PA Lab Malware	1,184	84.6%	3.5	366
△ Malc0de IP Blacklist	865	61.0%	2.9	86.6
○ PA Bambenek C2 IPs	785	92.1%	3.4	97.9
○ PA SSL Malware IPs	676	53.9%	2.9	84.0
○ PA Analyst	492	79.8%	2.1	149
○ PA Abuse.ch Ransomware	256	7.03%	1.6	117
○ PA Mal-Traffic-Anal	251	60.5%	0.9	72
○ Zeus IP Blacklist	185	49.1%	0.5	101
Exploit Feeds				
△ Badips HTTP	305,020	97.6%	1,592	22,644
△ Badips FTP	285,329	97.5%	1,313	27,601
△ Badips DNS	46,813	99.3%	231	4,758
△ Badips RFI	3,642	91.4%	16	104
△ Badips SQL	737	79.5%	4.4	99.2
Spam Feeds				
○ Paid IP Reputation	543,583	99.9%	3,280	6,551
△ Badips Postfix	328,258	90.5%	842	27,951
△ Badips Spam	302,105	89.3%	1,454	30,197
○ PA Botscout IPs	14,514	89.3%	49	4,390
○ Alienvault IP Reputation	11,292	96.6%	48	1,328

[1]This feed is aggregated by *PA* from Alienvault OTX, the Alienvault IP Reputation is the public reputation feed we collected from AlienVault directly. They are different feeds.

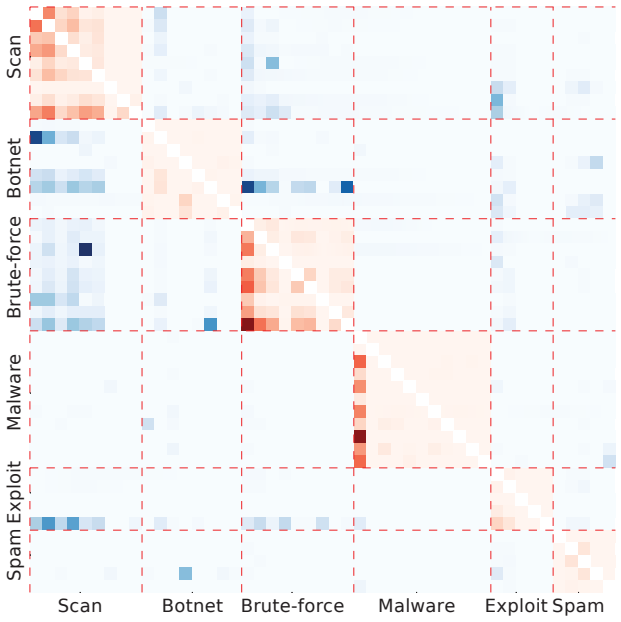


Figure 1. Feed intersection for all IP feeds. Each row/column represents a feed, shown in the same order as Table 1. Darker (more saturated) colors indicate greater intersection.

between an indicator being added and being removed. Four Paid IP Reputation feeds have more than 85% of durations shorter than 10 days, while the one in the malware category has more than 40% that span longer than 20 days. Feodo IP Blacklist has over 99% of its indicators valid for our entire measurement period, while over 70% of durations in the Zeus IP Blacklist are less than 6 days. We did not observe a clear pattern regarding how each snapshot feed handles the expiration of indicators.

3.3 Differential Contribution and Intersection

The differential contribution metric measures the number of indicators in one feed that are not in another. Equivalently, we can consider the intersection of two feeds, which is the number of elements in one feed that are present in the other, normalized by the size of the first: $|A \cap B|/|A|$. Figure 1 shows the intersection relationship of all feeds in the study. Each cell in the matrix represents the number of elements in both feeds, normalized by the size of the feed spanning the rows on the table. That is, A, in the expression above, ranges over rows, and B over columns of the matrix. Darker (more saturated) colors indicate greater intersection. Comparisons of feeds within a category are shaded red and comparisons of feeds between different categories are shaded blue. Note that the matrix is asymmetric, because, in general, $|A \cap B|/|A| \neq |A \cap B|/|B|$. Elements of the matrix are in the same order as in Table 1.

✦ **Finding:** Feeds in scan and brute-force categories have higher pairwise intersections: Half of the pairwise intersection rates in two categories are greater than 5%. The scan category has 29 out of 72 pairs (excluding self comparisons) with an intersection rate larger than 10%, and the same case occurred in 19 out of 72 pairs in the brute-force category.

On the other side, feeds in the botnet, exploit, malware and spam category do not share much data between each other: all 4 categories have more than three-quarters of pairwise intersection rates less than 1%. A few big feeds in these categories can share a significant amount of data with some small feeds in the same category—a characteristic that appears as a dark vertical line within its category in Figure 1. Paid IP Reputation in the malware category, for example, shares over 30% of 6 other malware feeds. But the intersections among the vast majority of feeds in these 4 categories are low. This finding is consistent with prior work [26, 42], but we provide a more comprehensive view regarding different categories.

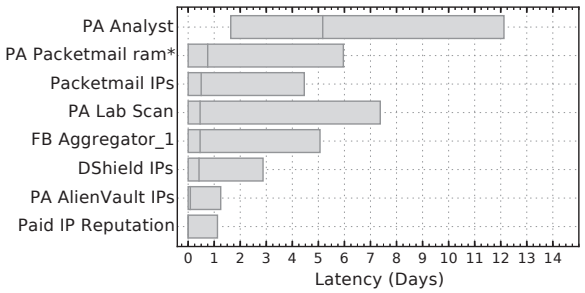
Figure 1 also shows the relation between feeds across different categories. We can clearly see a relation between scan and brute-force feeds: multiple scan feeds have non-trivial intersection with feeds in the brute-force category. In fact, 23.1% of all 760,263 brute-force IPs we collected are also included by scan feeds in our dataset. There are also three botnet feeds—PA CI Army, PA VoIP Blacklist and PA Compromised IPs—that have over 10% of its data shared with multiple feeds in the scan category.

3.4 Exclusive Contribution

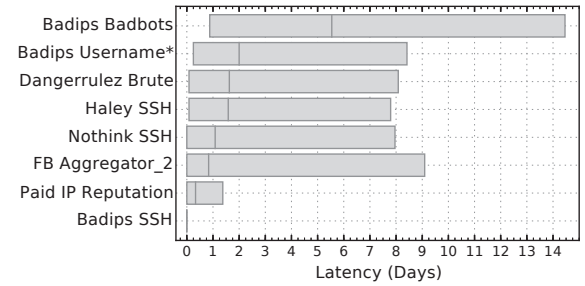
Exclusive contribution represents the number of indicators in a feed that are in no other feeds. We calculate each feed's exclusive contribution among all the feeds in the same category, emphasizing their uniqueness regarding the scope of data they claim to report. Each feed's exclusive contribution is presented in Table 1 in column *Exclusive*, calculated based on its volume.

✦ **Finding:** As we already observed in Section 3.3, botnet, exploit and spam feeds have relatively low pairwise intersections. Consequently, the feeds in these four categories have high exclusive contribution rates in general: the median exclusive contribution rates of these four categories are 90.9%, 97.5% and 90.5%, respectively. The malware category has a low median exclusive rate, since multiple small feeds have non-trivial intersection with the largest feed Paid IP Reputation, but the two largest feeds in malware both have a exclusive rate over 99%. Scan and brute-force feeds have more intersection within its category, and their exclusive rates are lower: 62.0% median rate in scan and 62.7% in brute-force, and the top two largest feeds in both categories have an exclusive rate below 85%.

If we assume a process where a feed is more likely to have popular elements, then smaller feeds would be subsumed by larger feeds. Yet, for some small feeds like Malc0de IP Blacklist in the malware and PA Project Honeypot in the botnet categories, even though they are several orders of magnitude smaller than the largest feeds in their categories, a significant



(a) Latency distribution in scan feeds



(b) Latency distribution in brute-force feeds

Figure 2. Distribution of indicators' latency in scan and brute-force feeds. Each box shows the latency distribution of shared IPs in the feed calculated in hours from 25 percentile to 75 percentile, with the middle line indicating the median. ("Badips Username*" here is the abbreviation for feed name Badips Username Notfound; "PA Packetmail Ram*" for PA Packetmail Ramnode)

proportion of their indicators is still unique to the feed. When we aggregate the data in each category, 73% of all scan feed indicators are unique to a single feed and 88% of brute force feed indicators are unique to one feed. For other categories, over 97% of elements in the category are unique to a single feed. This result agrees with previous work that most data in threat intelligence feeds is unique [26, 42].

3.5 Latency

Feed latency measures how quickly a feed reports new threat indicators. The sooner a feed can report potential threats, the more valuable it is for consumers. The absolute latency of an indicator in a feed is the time from the beginning of the corresponding event until when the indicator shows up in the feed. However, it is difficult to know the actual time when an event begins from the threat intelligence data. Instead, we measure the *relative latency*, which is the delay of an indicator in one feed to be the time between its appearance in that feed and the first seen among all the feeds.

Relative latency can only be calculated for indicators that occur in at least two feeds. As discussed in Section 3.4, the number of common indicators in the botnet, malware, exploit and spam feeds is very low (fewer than 3% of elements occur in more than one feed). Relative latency calculated for these feeds is less meaningful. For this analysis, therefore, we focus on scan and brute-force feeds.

Another issue is the time sensitivity of IP threats. An event that originated from an IP address, like scanning activity or a brute-force attack, will not last forever. If one scan feed reports an IP address today and another feed reports the same IP three months later, it would make little sense to consider them as one scanning event and label the second occurrence as being three months late. Unfortunately, there is no easy way we can clearly distinguish events from each other. Here we use a one-month window to restrict an event, assuming that the same attack from one source will not last for more than 30 days; although arbitrary, it provides a reasonably conservative threshold, and experimenting with other thresholds produced similar overall results. More specifically, we calculate relative latency by tracking the first occurrence of IPs in all feeds in a category, then recording the latency of the following occurrences while excluding ones that occur after 30 days. By just using the first appearance of each IP as the base, we avoid the uncertainty caused by multiple occurrence of indicators and different valid periods used among feeds.

Figures 2a and 2b show the relative latency distribution among feeds in the scan and brute-force categories, in hours. We focus on just those feeds that have over 10% of their data shared with others to ensure the analysis can represent the latency distribution of the overall feed. There is one feed in each category (PA Snort BlockList in scan and PA Brute-Force in brute-force) that is excluded from the figure.

✦ **Finding:** From the distribution boxes we can see that Paid IP Reputation in scan and Badips SSH in brute-force are the fastest feeds in their category, as they have the lowest median and 75th percentile latencies. On the other hand, PA Analyst in scan and Badips Badbots in brute-force are the slowest feeds. Figure 2a shows that all scan feeds except one have their 25th percentile latency equal to 0, indicating these feeds, across different sizes, all reported a significant portion of their shared data first. A similar case also happens in the brute-force category.

One may reasonably ask whether large feeds report data sooner than small feeds. The result shows that this is not always the case. FB Aggregator$_1$ is the second smallest feed in our scan category, yet it is no slower than several other feeds which have over 10 times of its daily rate. Badips Badbots, on the other hand, has the second largest rate in brute-force category, but it is slower than all the other feeds in the brute-force category. Feeds that are small in volume can still report a lot of their data first.

Another factor that could affect latency is whether feeds copy data from each other. For example, 93% of Dangerrulez Brute also appears in Badips SSH. If this is the case, we expect Dangerrulez Brute will be faster than Badips SSH on reporting their shared data. However, we compared the relative latency between just two feeds and found Badips SSH reported 88% of their shared indicators first. We further conducted this pairwise latency comparison between all feeds

Table 2. IP TI feeds accuracy overview. *Unrt* is fraction of unroutable addresses in each feed (Section 3.6). *Alexa Top* is the number of IPs intersected with top Alexa domain IP addresses, and *CDNs* is the number of IPs intersected with top CDN provider IP addresses.

Feed	*Added*	*Unrt*	*Alexa*	*CDNs*
Scan Feeds				
PA AlienVault IPs	313,175	0.0%	1	0
DShield IPs	339,805	0.03%	68	62
PA Packetmail ramnode	200,568	<0.01%	0	0
Packetmail IPs	211,081	0.0%	0	0
Paid IP Reputation	200,915	1.65%	6	21
PA Lab Scan	169,037	<0.01%	0	0
PA Snort BlockList	12,957	0.42%	1	0
FB Aggregator$_1$	5,601	0.0%	0	0
PA Analyst	1,451	0.41%	0	0
Botnet Feeds				
PA Analyst	180,034	<0.01%	0	0
PA CI Army	76,125	<0.01%	0	0
Paid IP Reputation	73,710	1.66%	6	74
PA Botscout IPs	18,638	0.09%	1	0
PA VoIP Blacklist	9,290	0.32%	0	0
PA Compromised IPs	4,883	0.0%	0	0
PA Blocklist Bots	3,594	0.0%	0	0
PA Project Honeypot	1,947	0.0%	0	0
Brute-force Feeds				
Badips SSH	456,605	0.19%	217	1
Badips Badbots	91,553	1.04%	46	1,251
Paid IP Reputation	87,524	0.03%	0	10
PA Brute-Force	31,555	0.0%	0	0
Badips Username Notfound	37,198	0.53%	4	0
Haley SSH	8,784	0.03%	0	0
FB Aggregator$_2$	17,779	0.0%	0	0
Nothink SSH	20,325	1.51%	2	0
Dangerrulez Brute	8,247	0.0%	0	0
Malware Feeds				
Paid IP Reputation	217,073	0.13%	291	3,489
FB Malicious IPs	29,840	2.14%	2	0
Feodo IP Blacklist	296	0.0%	0	0
PA Lab Malware	806	2.85%	0	0
Malc0de IP Blacklist	668	0.0%	8	11
PA Bambenek C2 IPs	777	9.13%	0	0
PA SSL Malware IPs	674	0.0%	0	0
PA Analyst	486	0.0%	0	0
PA Abuse.ch Ransomware	256	3.12%	0	0
PA Mal-Traffic-Anal	193	0.51%	0	0
Zeus IP Blacklist	67	0.0%	1	0
Exploit Feeds				
Badips HTTP	305,020	0.67%	16	2,590
Badips FTP	285,329	1.33%	14	2
Badips DNS	46,813	0.50%	119	244
Badips RFI	3,642	2.22%	0	0
Badips SQL	737	1.89%	0	1
Spam Feeds				
Paid IP Reputation	543,546	78.7%	1	0
Badips Spam	302,105	0.02%	19	0
Badips Postfix	193,674	1.29%	18	1
PA Botscout IPs	11,358	0.06%	0	0
Alienvault IP Reputation	10,414	0.07%	63	1,040

in scan, brute-force and malware (since Paid IP Reputation shares non-trivial amount of data with a few small feeds in the malware category), and did not see a clear latency advantage between any two feeds. Note that this observation does *not* prove there is no data copying, since the shared data between two feeds might partially come from copying and partially from the feeds' own data collection. Furthermore, our latency analysis is at a one-hour granularity.

3.6 Accuracy

Accuracy measures the rate of false positives in a feed. A false positive is an indicator that data is labeled with a category to which it does not belong. For example, an IP address found in a scan feed that has not conducted any Internet scanning is one such false positive. As well, even if a given IP is in fact associated with malicious activity, if it is not unambiguously actionable (e.g., Google's DNS at 8.8.8.8 is used by malicious and benign software alike) then for many use cases it must also be treated as a false positive. False positives are problematic for a variety of reasons, but particularly because they can have adverse operational consequences. For example, one might reasonably desire to block all new network connections to and from IP addresses reported as hosting malicious activity (indeed, this use is one of the promises of threat intelligence). False positives in such feeds, though, could lead to blocking legitimate connections as well. Thus, the degree of accuracy for a feed may preclude certain use cases.

Unfortunately, determining which IPs belong in a feed and which do not can be extremely challenging. In fact, at any reasonable scale, we are unaware of any method for unambiguously and comprehensively establishing "ground truth" on this matter. Instead, in this section we report on a proxy for accuracy that provides a conservative assessment of this question. To wit, we assemble a *whitelist* of IP addresses that either should not reasonably be included in a feed, or that, if included, would cause significant disruption. We argue that the presence of such IPs in a feed are clearly false positives and thus define an upper bound on a feed's accuracy. We populate our list from three sources: unroutable IPs, IPs associated with top Alexa domains, and IPs of major content distribution networks (CDNs).

Unroutable IPs. Unroutable IPs are IP addresses that were not BGP-routable *when they first appeared* in a feed, as established by contemporaneous data in the RouteViews service [44]. While such IPs could have appeared in the source address field of a packet (i.e., due to address spoofing), it would not be possible to complete a TCP handshake. Feeds that imply that such an interaction took place should not include such IPs. For example, feeds in the Brute-force category imply that the IPs they contain were involved in brute-force login attempts, but this could not have taken place if the IPs are not routable. While including unroutable addresses in a feed is not, in itself, a problem, their inclusion suggests a quality control issue with the feed, casting shade on the validity of other indicators in the feed.

To allow for some delays in the feed, we check if an IP was routable at any time in the seven days prior to its first appearance in a feed, and if it had, we do not count it as unroutable. Table 2, column *Unrt*, shows the fraction of IP indicators that were not routable at any time in the seven days prior to appearing in the feed. This analysis is only conducted for the IPs that are added after our measurement started. The number of such IPs is shown in column *Added*,

and the unroutable fraction shown in *Unrt* is with respect to this number.

Alexa. Blocking access to popular Internet sites or triggering alarms any time such sites are accessed would be disruptive to an enterprise. For our analysis, we periodically collected the Alexa top 25 thousand domains (3–4 times a month) over the course of the measurement period [2]. To address the challenge that such lists can have significant churn [33], we restrict our whitelist to hold the *intersection* of all these top 25K lists (i.e., domains that were in the top 25K every time we polled Alexa over our 8-month measurement period), which left us with 12,009 domains. We then queried DNS for the A records, NS records and MX records of each domain, and collected the corresponding IP addresses. In total, we collected 42,436 IP addresses associated with these domains. We compute the intersection of these IPs with TI feeds and show the results in column *Alexa* in Table 2.

CDNs. CDN providers serve hundreds of thousands of sites. Although these CDN services can (and are) abused to conduct malicious activities [9], their IP addresses are not actionable. Because these are fundamentally shared services, blocking such IP addresses will also disrupt access to benign sites served by these IPs. We collected the IP ranges used by 5 popular CDN providers: AWS CloudFron [12], Cloudflare [11], Fastly [18], EdgeCast [16] and MaxCDN [25]. We then check how many IPs in TI feeds fall into these ranges. Column *CDNs* in Table 2 shows the result.

✦ **Finding:** Among the 47 feeds in the table, 33 feeds have at least one unroutable IP, and for 13 of them, over 1% of the addresses they contain are unrouteable. Notably, the Paid IP Reputation feed in the spam category has an unroutable rate over 78%. Although it is not documented, a likely explanation is that this feed may include unroutable IPs intentionally, as this is a known practice among certain spam feeds. For example, the Spamhaus DROP List [41] includes IP address ranges known to be owned or operated by malicious actors, whether currently advertised or not. Thus, for feeds that explicitly do include unroutable IPs, their presence in the feeds should not necessarily be interpreted as a problem with quality control.

We further checked feeds for the presence of any "reserved IPs" which, as documented in RFC 8190, are not globally routable (e.g., private address ranges, test networks, loopback and multicast). Indeed, 12 feeds reported at least one reserved IP, including four of the Paid IP Reputation feeds (excepting the spam category), six of the Badips feeds, and the FB Malicious IPs and DShield IPs feeds. Worse, the Paid IP Reputation feeds together reported over 100 reserved IPs. Since such addresses should never appear on a public network, reporting such IPs indicates that a feed provider fails to incorporate some basic sanity checks on its data.

There are 21 feeds that include IPs from top Alexa domains, as shown in column *Alexa* in Table 2. Among these IPs there are 533 A records, 333 IPs of MX records and 63 IPs of NS records. The overlapped IPs include multiple instances from notable domains. For example, the IP addresses of www.github.com are included by Malc0de IP Blacklist. Paid IP Reputation in the malware category contains the IP address for www.dropbox.com. Alienvault IP Reputation contains the MX record of groupon.com, and Badips SSH also contains the IP addresses of popular websites such as www.bing.com.

Most of the feeds we evaluated do not contain IPs in CDN ranges, yet there are a few (including multiple Paid IP Reputation feeds, Badips feeds and Alienvault IP Reputation) that have significant intersection with CDN IPs. Alienvault IP Reputation and Badips feeds primarily intersect with Cloudflare CDN, while most of the overlap in the Paid IP Reputation malware category overlaps with AWS CloudFront.

Overall, the rate of false positives in a feed is not strongly correlated with its volume. Moreover, certain classes of false positives (e.g., the presence of Top Alex IPs or CDN IPs) seem to be byproducts of how distinct feeds are collected (e.g., Badips feeds tend to contain such IPs, irrespective of volume). Unsurprisingly, we also could find not correlation between a feed's latency and its accuracy.

3.7 Coverage

The coverage metric provides a quantitative measure of how well a feed captures the intended threat. A feed with perfect coverage would include all indicators that belong in a category. Unfortunately, as discussed above, there is no systematic way for evaluating the exact accuracy or coverage of a feed since it is unrealistic to obtain ground truth of all threat activities on the Internet.

However, there are some large-scale threat activities that are well-collected and well-studied. One example is Internet scanning. Researchers have long been using "Internet telescopes" to observe and measure network scanning activities [6, 15, 29]. With a large telescope and well-defined scan filtering logic, one can obtain a comprehensive view of global scanning activities on the Internet.

To this end, we collected three months of traffic (from January 1st to March 31st 2018) using the UCSD network telescope [38], which monitors a largely quiescent /8 network comprising over 16 million IP addresses. We then used the default parameters of the Bro IDS [7] to identify likely scanning traffic, namely flows in which the same source IP address is used to contact 25 unique destination IP addresses on the same destination port/protocol within 5 minutes. Given the large number of addressed being monitored, any indiscriminate scanner observed by TI feeds will likely also be seen in our data. Indeed, by intersecting against this telescope data we are able to partially quantify the coverage of each TI scanning feed.

The scanners we collected from the telescope consist of 20,674,149 IP addresses. The total number of IPs in all the scan feeds during this period is 425,286, which covers only 1.7% (363,799 shared IPs) of all the telescope scan IPs. On the

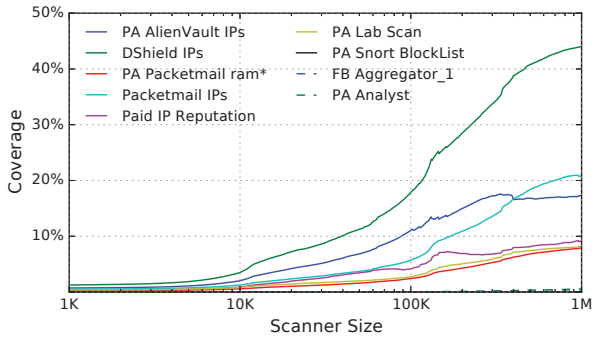


Figure 3. The coverage of each feed on different sizes of scanners. **Y** axis is the proportion of scanners of a given size or larger that are covered by each feed.

other hand, telescope scanners intersect with 85% of all IPs in scan feeds. When looking at each feed, PA AlienVault IPs, DShield IPs Packetmail IPs, PA Lab Scan and PA Packetmail ramnode all have over 85% of their data intersected with telescope scanners; the other four, though, have less than 65% of their data shared (and the rate for PA Snort BlockList is only 8%).

To further understand how well each scan feed detects scanning activities, we measure how different sizes of scanners in the telescope are covered by each feed. Here, *scanner size* means how many IPs a scanner has scanned in the telescope within a day. Figure 3 shows the coverage rate of each feed over different sizes of scanners, ranging from 1,000 to 1 million. (There are 7,212,218 scanners from the telescope whose sizes are over 1K, 271,888 that are over 100K and 17,579 are over 1 million.)

✦ **Finding:** The union of all the scan IPs in the feeds covers less than 2% of the scanners collected by the telescope. Even if we only look at the scanners with sizes larger than 10,000, the overall coverage is still around 10%, suggesting the coverage capability of scan feeds is very limited. The graph shows that, as the scanner size increases, the coverage of each feed over the datasets also increases, and large feeds cover more percent of telescope scanners than small feeds. This trend aligns with the intuition that scan feeds tend to capture more extensive scanners.

It is surprising that the small scan feeds in our collection have a smaller percentage of their IPs shared with telescope scanners. This contradicts the idea that small feeds would contain a larger percentage of extensive scanners (that would most likely also be observed by the telescope).

4 File Hash Threat Intelligence

File hashes in a threat intelligence feed are indicators for malicious files. It is one of the most lightweight ways to mark files as suspicious. One can incorporate this data to block malicious downloads, malicious email attachments, and malware. Likewise, file hashes can be used to whitelist applications and these feeds can be used to ensure malicious files do not appear in a customer's whitelist. In this section we present our analysis on eight file hash feeds, also collected from December 1st, 2017 to July 20th, 2018. We use the same metrics defined in Section 2.3.

The file hash feeds we collected use a range of different hash functions to specify malicious files, including MD5, SHA1, SHA256 and SHA512 (and some feeds provided values for multiple different hash functions to support interoperability). Since most indicators in our dataset are MD5s, we have normalized to this representation by using other feeds and the VirusTotal service to identify hash aliases for known malicious files (i.e., which MD5 corresponds to a particualr SHA256 value).

4.1 Volume

File hashes, unlike IP threat data, are not transient—a file does not change from malicious to benign—and thus a far simpler volume analysis is appropriate. We report volume as the number of new hashes that are added to each feed during our measurement period.

As seen in Table 3, we examine each feed's volume and average daily rate. Like IP feeds, file hash feeds also vary dramatically in volume. The majority of the hashes are concentrated in three feeds: FB Malware, PA Malware Indicators, and PA Analyst, which also exhibit the highest daily rates. The other feeds are multiple order of magnitude smaller comparatively.

4.2 Intersection and Exclusive Contribution

As we mentioned earlier, to conduct intersection and exclusive analysis of file hash feeds, we need to convert indicators into the same hash type. Here we convert non-MD5 hashes into MD5s, using either metadata in the indicator itself (i.e., if it reports values for multiple hash functions) or by querying the source hash from VirusTotal [45] which reports the full suite of hashes for all files in its dataset. However, for a small fraction of hashes we are unable to find aliases to conver them to the MD5 representation and must exclude them from the analysis in this section. This filtering is reflected in Table 3, in which the Volume column represents the number of unique hashes found in each feed and the Converted column is the subset that we have been able to normalize to a MD5 representation.

✦ **Finding:** The intersections between hash feeds are minimal, even among the feeds that have multiple orders of magnitude differences in size. Across all feeds, only PA Analyst has relatively high intersections: PA Analyst shares 27% of PA OSINT's MD5s and 13% of PA Twitter Emotet's MD5s. PA Malware Indicators has a small intersection also with these two feeds. All other intersections are around or less than 1%. Consequently, the vast majority of MD5s are unique to one feed, as recorded in column *Exclusive* in Table 3. The "lowest" exclusivity belongs to PA Twitter Emotet and PA OSINT (still

Table 3. File hash feeds overview. The second column group presents feed volume, average daily rate, the number of converted MD5s (Section 4.2) and exclusive proportion. *Not in VT* is fraction of hashes that are not found in VirusTotal, *Not det.* the fraction of hashes that are found in VirusTotal but are not labeled as malicious by any products, and *Detected* the fraction that are found in VirusTotal and are labeled malicious by at least one product. Column *Not in SD* shows the fraction of hashes in a feed that are not in Shadowserver Bin Check. *In NSRL* and *In AppInfo* show the absolute number of hashes found in Shadowserver (Section 4.3). *Exclusive* is based on the MD5-normalized hashes counted under *Converted*. All the other percentages in the table are based on *Volume*.

Feed	Volume	Avg. Rate	Converted	Exclusive	Not in VT	Not det.	Detected	Not in SD	In NSRL	In AppInfo
FB Malware	944,257	4,070	944,257	>99.99%	37.41%	50.50%	12.09%	99.89%	442	706
PA Malware Indicators	39,702	171	39,702	98.73%	0.02%	0.04%	99.94%	>99.99%	2	0
PA Analyst	38,586	166	37,665	97.97%	4.26%	2.82%	92.92%	99.95%	8	19
PA Twitter Emotet	1,031	4.44	960	77.29%	11.74%	0.78%	87.49%	99.81%	0	2
PA OSINT	829	3.57	783	71.65%	19.06%	0.84%	80.10%	99.88%	1	0
PA Sandbox	298	1.28	115	95.65%	72.81%	0.34%	26.85%	100%	0	0
PA Abuse.ch	267	1.15	3	100%	98.88%	0.75%	0.37%	100%	0	0
PA Zeus Tracker	17	0.07	17	100%	88.24%	5.88%	5.88%	100%	0	0

77.29% and 71.65%, respectively). All other feeds showcase an over 95% exclusive percentage, demonstrating that most file hash feeds are distinct from each other.

Due to the different sources of malware between feeds, a low intersection is to be expected in some cases. For example, PA Twitter Emotet and PA Zeus Tracker should have no intersection, since they are tracking different malware strains. The other, more general feeds could expect some overlap, but mostly exhibit little to no intersection. Considering the sheer volume of the FB Malware feed, one might expect it would encapsulate many of the smaller feeds or at least parts of them. This is not the case, however, as FB Malware has a negligible intersection with all other feeds.

Due to the lack of intersection among the feeds, we omit the latency analysis of the hash feeds, as there is simply not enough intersecting data to conclude which feeds perform better with regards to latency.

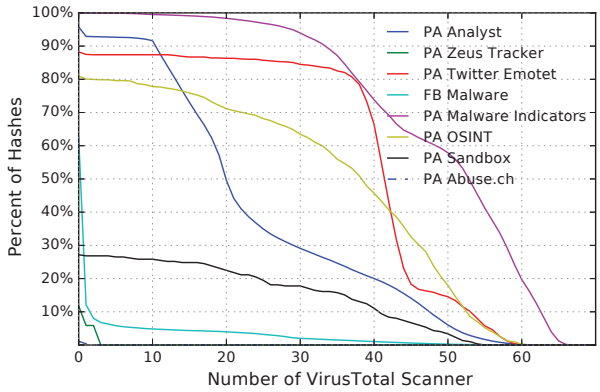


Figure 4. VirusTotal detection distribution. Each point means the proportion of indicators (Y value) in a feed that is detected by *over* X number of AV scanners in VirusTotal.

4.3 Accuracy

Assessing the accuracy of file hash feeds presents a problem: there is no universal ground truth to determine if a file is malicious or benign. Thus, to gauge the accuracy of the feeds, we use two metrics: a check for malicious hashes against VirusTotal, and a check for benign hashes against Shadowserver's Bin Check service. Note that all the percentages discussed below are based on the *Volume* of each feed.

4.3.1 VirusTotal

VirusTotal is a service that is often used when analyzing malware to get a base of information about a suspected file. Anyone can upload a file to be scanned. Upon submission, these files will be scanned by more than 70 antivirus scanners, which creates a report on how many antivirus scanners mark it malicious, among other information. In this analysis, we query VirusTotal for the hashes in each file hash feed and then inspect the percent of hashes that are marked as malicious and how many AV scanners have recorded them. Due to the high volume of the FB Malware feed and the query rate limit of VirusTotal, we randomly sampled 80,000 hashes from the feed for this analysis.

Table 3 shows a breakdown of the base detection rates for each feed from VirusTotal. As the PA feeds decrease in volume, the rates at which they are found in VirusTotal also decreases. The larger PA feeds have a much higher detection rate than their smaller counterparts. On the other hand, FB Malware only has 37% of its data detected by antivirus scanners and 50% in VirusTotal with no detection despite being the largest feed. This could indicate that FB Malware focuses on threats that specifically target Facebook and that are not as relevant to most VirusTotal users, such as malicious browser extensions [14, 20, 22]. This might undermine the limited coverage of VirusTotal as an oracle to detect targeted threats that are not of broader interest.

To further understand how the scanners in VirusTotal report the feed's data, we plot a graph of what percentage of hashes in each feed are detected by how many VirusTotal scanners. As seen in Figure 4, four feeds have more than 50% of their samples detected by over 20 scanners. PA Malware Indicators and PA Twitter Emotet did not experience a large detection drop before 35 scanners, indicating that most indicators in the

two feeds are popular malicious files recognized by many AV vendors. While PA Sandbox has a large percent of its hashes not presented in VirusTotal, over 70% of its samples that are detected are marked by over 20 AV scanners, showcasing a high confidence detection.

4.3.2 Shadowserver

To more fully gauge the accuracy of the file hash feeds, we also examined how each feed measured against Shadowserver's Bin Check Service [34]. The service checks file hashes against NIST's National Software Registry List (NSRL) in addition to Shadowserver's own repository of known software. Table 3 details how each feed compares with Shadowserver's Bin Check service.

It might be expected that there would be no hash found with Shadowserver's Bin Check service, but it is not the case. Some of the samples from the feeds that appear in Shadowserver are well known binaries such as versions of Microsoft Office products, Window's Service Packs, calc.exe, etc. In the event malware injects itself into a running process, it remains plausible that some of these well-known binaries find their way into TI feeds from users wrongly attributing maliciousness. While FB Malware has over one thousand hashes in Shadowserver, this is not a widespread issue, as all feeds have <1% of their hashes contained within Shadowserver's Bin Check service. This showcases that while there are a few exceptions, the feeds mostly do not contain well-known, benign files.

✦ **Finding:** Each PA feed has a negligible rate of occurrence within Shadowserver regardless of their VirusTotal detection, showing they do not contain generic false positives. Larger feeds exhibit high VirusTotal detection rates except for FB Malware, while small feeds have relatively low detection rates. This suggests that small hash feeds might focus more on specific malicious files that are not widely known. FB Malware has a low VirusTotal occurrence despite its size and has over one thousand hashes in Shadowserver, but its overall low percentage of hashes within Shadowserver indicates that it does not contain many known files and might have threats not typically recognized by VirusTotal's scanners.

5 Longitudinal Comparison

In addition to the measurement period considered so far (December 1, 2017 to July 20, 2018), we also analyzed data from the same IP feeds from January 1, 2016 to August 31, 2016. These two measurement periods, 23 months apart, allow us to measure how these IP feeds have changed in two years. Table 4 summarizes the differences between these two measurement periods. In the table, *2018* represents the current measurement period and *2016* the period January 1, 2016 to August 31, 2016.

Volume. As shown in Table 4, feed volume has definitely changed after two years. Among 43 IP feeds that overlap both time periods, 21 have a higher daily rate compared with 2 years ago, 15 feeds have a lower rate, and 7 feeds do not change substantially (the difference is below 20%). Volume can change dramatically over time, such as PA AlienVault IPs in the scan category which is 13 times larger than before. On the other hand, a feed like PA Blocklist Bots is now over 90% smaller.

Intersection and Exclusive Contribution. Despite the volume differences, the intersection statistics between feeds are largely the same across two years, with feeds in scan and brute-force having high pairwise intersections and feeds in other categories being mostly unique. Certain specific pairwise relations also did not change. For example, Badips SSH still shared over 90% of data in Dangerrulez Brute back in

Table 4. Data changes in IP feeds compared against the ones in 2016, *Avg. Rate* shows the percentage of daily rate changed over the old feeds. The two columns under *Unrt* show the unroutable rates of feeds in 2016 and 2018 separately. The two columns under *CDN* present the number of IPs fall in CDN IP ranges in old and new data.

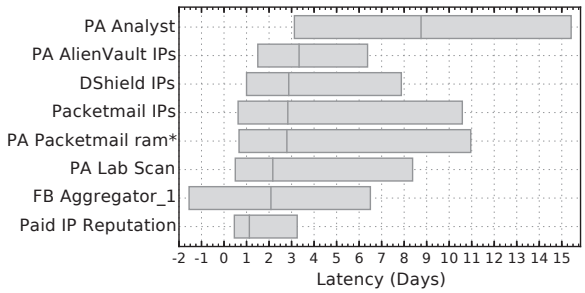
		Unroutable		*CDN*	
Feed	*Avg. Rate*	2016	2018	2016	2018
Scan Feeds					
PA AlienVault IPs	+1,347%	0.0%	0.0%	0	0
PA Packetmail ram*	+733%	<0.01%	<0.01%	0	0
Packetmail IPs	+135%	0.0%	0.0%	0	0
Paid IP Reputation	−57%	8.73%	1.65%	910	21
PA Lab Scan	−1%	0.0%	<0.01%	0	0
PA Snort BlockList	−97%	<0.01%	0.42%	1	0
FB Aggregator$_1$	+332%	0.0%	0.0%	6	0
PA Analyst	−44%	0.0%	0.41%	0	0
Botnet Feeds					
PA CI Army	+114%	<0.01%	<0.01%	0	0
Paid IP Reputation	−39%	0.63%	1.66%	15	74
PA Botscout IPs	+1%	0.01%	0.09%	1	0
PA VoIP Blacklist	+252%	0.0%	0.32%	0	0
PA Compromised IPs	−36%	0.10%	0.0%	0	0
PA Blocklist Bots	−95%	0.0%	0.0%	0	0
PA Project Honeypot	+63%	0.0%	0.0%	0	0
Brute-force Feeds					
Badips SSH	+30%	0.07%	0.19%	0	1
Badips Badbots	+1,732%	0.0%	1.04%	187	1,251
Paid IP Reputation	−62%	6.55%	0.03%	335	10
PA Brute-Force	−72%	0.0%	0.0%	0	0
Badips Username*	+3,040%	0.0%	0.53%	0	0
Haley SSH	+428%	0.04%	0.03%	0	0
FB Aggregator$_2$	+387%	0.12%	0.0%	0	0
Nothink SSH	+886%	0.56%	1.51%	0	0
Dangerrulez Brute	+0%	0.0%	0.0%	1	0
Malware Feeds					
Paid IP Reputation	−36%	0.18%	0.13%	15265	3,489
FB Malicious IPs	−77%	6.81%	2.14%	264	0
Feodo IP Blacklist	+0%	0.0%	0.0%	0	0
Malc0de IP Blacklist	−9%	0.0%	0.0%	132	11
PA Bambenek C2 IPs	+79%	0.0%	9.13%	0	0
PA SSL Malware IPs	−34%	0.0%	0.0%	0	0
PA Analyst	−93%	0.34%	0.0%	0	0
PA Abuse.ch*	−99%	0.49%	3.12%	0	0
PA Mal-Traffic-Anal	−53%	0.0%	0.51%	0	0
Zeus IP Blacklist	−66%	0.0%	0.0%	6	0
Exploit Feeds					
Badips HTTP	+326%	0.30%	0.67%	436	2,590
Badips FTP	+556%	0.01%	1.33%	0	2
Badips DNS	+9,525%	0.17%	0.50%	7	244
Badips RFI	+226%	0.0%	2.22%	0	0
Spam Feeds					
Paid IP Reputation	+133%	59.3%	78.7%	0	0
Badips Spam	+12,767%	0.0%	0.02%	0	0
Badips Postfix	−53%	<0.01%	1.29%	0	1
PA Botscout IPs	+18%	0.0%	0.06%	0	0
AlienVault IP Rep	+8%	0.57%	0.07%	479	1,040

2016, and Paid IP Reputation in malware was still the only feed that has a non-trivial intersection with multiple small feeds. Again, most data was exclusive to each feed two years ago: Across all six categories more than 90% of the indicators are not shared between feeds.

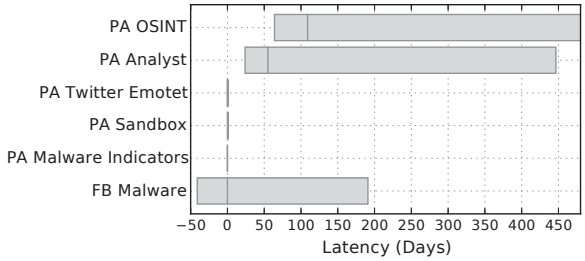
Latency. The latency relationship between feeds was also similar: timely feeds today were also timely two years ago, and the same with tardy feeds.

Accuracy. Feeds have more unroutable IPs now than before as shown in Table 4: In 2016, 22 of the 43 IP feeds had at least 1 unroutable IP; four feeds had unroutable rates over 1%. When checking the intersection with popular CDNs, the feeds that contain IPs in CDN ranges two years ago are also the ones that have these IPs today.

Shared indicators 2016–2018. We compared the data we collected from each feed in the two time periods, and found that 30 out of 43 feeds in 2018 intersect with their data from two years ago, and 9 feeds have an intersection rate over 10%. Three feeds in malware category, namely Feodo IP Blacklist, PA Abuse.ch Ransomware and Zeus IP Blacklist, have over 40% of their data shared with the past feed, meaning a large percent of C&C indicators two years ago are still identified by the feeds as threats today. Feeds in the botnet category, however, are very distinct from the past, with all feeds having no intersection with the past except Paid IP Reputation.



(a) Latency distribution in scan feeds relative to the Internet telescope



(b) Latency distribution in file hash feeds relative to VirusTotal

Figure 5. Distribution of indicators' latency in scan and file hash feeds. Note that the scan feeds' distribution are calculated in hour granularity while the file hash feeds' distribution are calculated in day granularity.

6 Absolute Latency

We defined our latency metric in this paper as relative latency between TI sources, since it is easy to compute and allows consumers to compare feeds to each other on this aspect. However, it is also critical to know about the absolute latency distribution of indicators. Absolute latency represents how fast a feed can actually report a threat, which directly decides the effectiveness of the data when used in a pro-active way. As we already discussed in Section 3.5, absolute latency is hard to measure, as we do not have ground truth of the underlying threat.

In Section 3.7, we used an Internet telescope as our approximation for ground truth to measure the coverage of scan feeds. In Section 4.3, we used VirusTotal as an oracle to measure the accuracy of file hash feeds. Although these sources are not real ground truth and it is unclear how far away they are, these large and well-managed sources can help us, to a certain extent, profile the performance of TI feeds. In this section, we use these two sources again to approximate the absolute latency of indicators in scan IP feeds and malicious file hash feeds.

More specifically, we measure the latency of IPs in scan feeds relative to the first occurrence time of the same IP in the scanners collected from the telescope. Considering the massive size of the telescope, it should presumably detect scanners much sooner after the scanning event actually happened. We measure latency of file hashes relative to the `first_seen` timestamps queried from VirusTotal. The `first_seen` timestamp represents the time when the corresponding file is first uploaded to VirusTotal. VirusTotal is a very popular service and it is a convention for many security experts to upload new malware samples to VirusTotal once they discovered them. Therefore, this timestamp roughly entails when the security community first noticed the malicious file and can be a good approximation for absolute latency.

Figure 5 show the latency distribution of each feed, using the same plotting convention as in Section 3.5. Some feeds are not shown in the figure as there are too little data points in those feeds to reason about distribution.

✦ **Finding:** Comparing Figure 5a to Figure 2a, we can see that the median latency of feeds are all larger. This is consistent with our assumption that a large sensor tends to receive indiscriminate scanners sooner. Scan feeds' median lantecy are one to three days relative to the Internet telescope, except PA Analyst, whose median latency is almost nine days. The order of median latency between feeds changed compared with Figure 2a, but since the original relative median latencies among scan feeds are very close, the new order here is more likely to be statistics variances. Also, note that although the PA AlienVault IPs seems much slower than it is in Figure 2a, its 75 percentile latency is still the second smallest one.

On the other hand, the latency distributions of hash feeds vary more dramatically. PA Malware Indicators, PA Sandbox and PA Twitter Emotet are almost as fast as VirusTotal: all three feeds have 25 percentile and median latency equal to

zero. PA OSINT and PA Analyst are comparatively much slower, and PA OSINT even has a 75 percentile latency of 1680 days. This might be because of the heterogeneous nature of malware feeds. The figure also shows that feed volumes do not imply their latency, as PA Analyst and FB Malware are much slower than the small hash feeds.

Figure 5 demonstrates that the Internet telescope and VirusTotal are indeed good approximations for absolute latency measurement, as most indicators in TI feeds are observed relatively later. However, every scan feed has over 2% of its indicators detected earlier than the telescope did. FB Aggregator$_1$ and DShield IPs even have over 10% of their indicators observed earlier. There is also a similar case in file hash feeds. This aligns with our observation in Section 3.5 that small feeds can still report a non-trivial amount of their data first. Another interesting observation is that both Facebook feeds, FB Aggregator$_1$ and FB Malware, have a large percent of their data observed earlier than the telescope or VirusTotal. This again suggests that Facebook (and its threat intelligence partners) might face more targeted threats, so those threats will be first observed by Facebook.

7 Discussion

7.1 Metrics Usage

Threat intelligence has many different potential uses. For example, analysts may consume threat data interactively during manual incident investigations, or may use it to automate the detection of suspicious activity and/or blacklisting. When not itself determinative, such information may also be used to *enrich* other data sources, informing investigations or aiding in automatic algorithmic interventions. We have introduced a set of basic threat intelligence metrics—volume, intersection, unique contribution, latency, coverage and accuracy—that can inform and quantify each of those uses. Depending on a number of factors, such as the intended use case and the cost of false positives and negatives, some of these metrics will become more or less important when evaluating a TI source. For example, a feed with poor accuracy but high coverage might be ideal when an analyst is using a TI source interactively during manually incident investigations (since in this case, the analyst, as a domain expert, can provide additional filtering of false positives). Similarly, latency might not be a critical metric in a retrospective use case (e.g., post-discovery breach investigation). However, if an organization is looking for a TI source where the IPs are intended to be added to a firewall's blacklist then accuracy and latency should likely be weighted over coverage, assuming that blocking benign activity is more costly.

Another common real-world scenario is that a company has a limited budget to purchase TI sources and has a specific set of threats (i.e., botnet, brute-force) they are focused on mitigating. In such cases, the metrics we have described can be used directly in evaluating TI options, biasing towards sources that maximize coverage of the most relevant threats while limiting intersection.

7.2 Data Labeling

Threat intelligence IP data carries different meanings. To properly use this data, it is critical to know what the indicators actually mean: whether they are Internet scanners, members of a botnet or malicious actors who had attacked other places before. We have attempted to group feeds by their intended meaning in our analysis.

However, this category information, which primarily comes from TI sources themselves, is not always available. Feeds such as Alienvault IP Reputation and Facebook Threat Exchange sources contain a significant number of indicators labeled "Malicious" or "Suspicious." The meanings of these indicators are unclear, making it difficult for consumers to decide how to use the data and the possible consequences.

For feeds that provide category information, it is sometimes too broad to be meaningful. For example, multiple feeds in our collection simply label their indicators as "Scanner." Network scanning can represent port scanning (by sending SYN packets), or a vulnerability scan (by probing host for known vulnerabilities). The ambiguity here, as a result of ad-hoc data labeling, again poses challenges for security experts when using TI data.

Recently, standard TI formats have been proposed and developed, notably IODEF [19], CybOX [13] and STIX [37], that try to standardize the threat intelligence presentation and sharing. But these standards focus largely on the data format. There is room to improve these standards by designing a standard *semantics* for threat intelligence data.

7.3 Limitations

There are several questions that our study does not address. We attempted to collect data from a diverse set of sources, including public feeds, commercial feeds and industrial exchange feeds, but it is inherently not comprehensive. There are some prohibitively expensive or publication-restricted data sources that are not available to us. More specialized measurement work should be done in the future to further analyze the performance of these expensive and exclusive data sources.

A second limitation is our visibility into how different companies use threat intelligence operationally. For a company, perhaps the most useful kind of metric measures how a threat intelligence source affects its main performance indicators as well as its exposure to risk. Such metrics would require a deep integration into security workflows at enterprises to measure the operation effect of decisions made using threat intelligence. This would allow CIOs and CSOs to better understand exactly what a particular threat intelligence product contributes to a company. As researchers, we do not use TI operationally. A better understanding of operational needs would help refine our metrics to maximize their utility for operations-driven consumers.

The third limitation is the lack of ground truth, a limitation shared by all the similar measurement work. It is simply very difficult to obtain the full picture of a certain category of threat, making it very challenging to precisely determine accuracy and coverage of feeds. In this study, we used data from an Internet telescope and VirusTotal as a close approximation. There are also a handful of cases where a security incident has been comprehensively studied by researchers, such as the Mirai study [4], and such efforts can be used to evaluate certain types of TI data. But such studies are few in number. One alternative is to try to establish the ground truth for a specific network. For example, a company can record all the network traffic going in and out of its own network, and identify security incidents either through its IDS system or manual forensic analysis. Then it can evaluate the accuracy and coverage of a TI feed under the context of its own network. This can provide a customized view of TI feeds.

8 Related Work

Several studies have examined the effectiveness of blacklist-based threat intelligence [23, 31, 32, 35, 36]. Ramachandran *et al.* [32] showed that spam blacklists are both incomplete (missing 35% of the source IPs of spam emails captured in two spam traps), and slow in responding (20% of the spammers remain unlisted after 30 days). Sinha *et al.* [36] further confirmed this result by showing that four major spam blacklists have very high false negative rates, and analyzed the possible causes of the low coverage. Sheng *et al.* [35] studied the effectiveness of phishing blacklists, showing the lists are slow in reacting to highly transient phishing campaigns. These studies focused on specific types of threat intelligence sources, and only evaluated their operational performance rather than producing empirical evaluation metrics for threat intelligence data sources.

Other studies have analyzed the general attributes of threat intelligence data. Pitsillidis *et al.* [30] studied the characteristics of spam domain feeds, showing different perspectives of spam feeds, and demonstrated that different feeds are suitable for answering different questions. Thomas *et al.* [42] constructed their own threat intelligence by aggregating the abuse traffic received from six Google services, showing a lack of intersection and correlation among these different sources. While focusing on broader threat intelligence uses, these studies did not focus on generalizable threat metrics that can be extended beyond the work.

Little work exists that defines a general measurement methodology to examine threat intelligence across a broad set of types and categories. Metcalf *et al.* [26] collected and measured IP and domain blacklists from multiple sources, but only focused on volume and intersection analysis. In contrast, we formally define a set of threat intelligence metrics and conduct a broad and comprehensive study over a rich variety of threat intelligence data. We conducted our measurement from the perspective of consumers of TI data to offer guidance on choosing between different sources. Our study also demonstrated the limitation of threat intelligence more thoroughly, providing comprehensive characteristics of cyber threat intelligence that no work had addressed previously.

9 Conclusion

This paper has focused on the simplest, yet fundamental, metrics about threat intelligence data. Using the proposed metrics, we measured a broad set of TI sources, and reported the characteristics and limitations of TI data. In addition to the individual findings mentioned in each section, here we highlight the high-level lessons we learned from our study:

- TI feeds, far from containing homogeneous samples of some underlying truth, vary tremendously in the kinds of data they capture based on the particularities of their collection approach. Unfortunately, few TI vendors explain the mechanism and methodology by which their data are collected and thus TI consumers must make do with simple labels such as "scan" or "botnet", coupled with inferences about the likely mode of collection. Worse, a significant amount of data does not even have a clear definition of category, and is only labelled as "malicious" or "suspicious", leaving the ambiguity to consumers to decide what action should be taken based on the data.

- There is little evidence that larger feeds contain better data, or even that there are crisp quality distinctions between feeds across different categories or metrics (i.e., that a TI provider whose feed performs well on one metric will perform well on another, or that these rankings will hold across threat categories). How data is collected also does not necessarily imply the feeds' attributes. For example, crowdsourcing-based feeds (e.g., Badips feeds), are not always slower in reporting data than the self-collecting feeds (like Paid IP Reputation).

- Most IP-based TI data sources are collections of singletons (i.e., that each IP address appears in at most one source) and even the higher-correlating data sources frequently have intersection rates of only 10%. Moreover, when comparing with broad sensor data in known categories with broad effect (e.g., random scanning) fewer than 2% of observed scanner addresses appear in most of the data sources we analyzed; indeed, even when focused on the largest and most prolific scanners, coverage is still limited to 10%. There are similar results for file hash-based sources with little overlap among them.

The low intersection and coverage of TI feeds could be the result of several non-exclusive possibilities. First is that the underlying space of indicators (both IP addresses and malicious file hashes) is large and each individual data source can at best sample a small fraction thereof. It is almost certain that this is true to some extent. Second, different collection

methodologies—even for the same threat category—will select for different sub distributions of the underlying ground truth data. Third, this last effect is likely exacerbated by the fact that not all threats are experienced uniformly across the Internet and, thus, different methodologies will skew to either favor or disfavor targeted attacks.

Based on our experience analyzing TI data, we try to provide several recommendations for the security community on this topic moving forward:

- The threat intelligence community should standardize data labeling, with a clear definition of what the data means and how the data is collected. Security experts can then assess whether the data fit their need and the type of action should be taken on this data.
- There are few rules of thumb in selecting among TI feeds, as there is not a clear correlation between different feed properties. Consumers need empirical metrics, such as those we describe, to meaningfully differentiate data sources, and to prioritize certain metrics based on their specific need.
- Blindly using TI data—even if one could afford to acquire many such sources—is unlikely to provide better coverage and is also prone to collateral damage caused by false positives. Customers need to be always aware of these issues when deciding what action should be taken on this data.
- Besides focusing on the TI data itself, future work should investigate the operational uses of threat intelligence in industry, as the true value of TI data can only be understood in operational scenarios. Moreover, the community should explore more potential ways of using the data, which will extend our understanding of threat intelligence and also influence how vendors are curating the data and providing the services.

There are many ways we can use threat intelligence data. It can be used to *enrich* other information (e.g., for investigating potential explanations of a security incident), as a probabilistic canary (i.e., identifying an overall site vulnerability via a single matching indicator may have value even if other attacks of the same kind are not detected) or in providing a useful source of ground truth data for supervised machine learning systems. However, even given such diverse purposes, organizations still need some way to prioritize which TI sources to invest in. Our metrics provide some direction for such choices. For example, an analyst who expects to use TI interactively during incident response would be better served by feeds with higher coverage, but can accommodate poor accuracy, while an organization trying to automatically label malicious instances for training purposes (e.g., brute force attacks) will be better served by the converse. Thus, if there is hope for demonstrating that threat intelligence can materially impact operational security practices, we believe it will be found in these more complex uses cases and that is where future research will be most productive.

10 Acknowledgment

We would like to thank our commercial threat providers who made their data available to us and made this research possible. In particular, we would like to thank Nektarios Leontiadis and the Facebook ThreatExchange for providing the threat data that helped facilitate our study. We are also very grateful to Alberto Dainotti and Alistair King for sharing the UCSD telescope data and helping us with the analysis, Gautam Akiwate for helping us query the domain data, and Matt Jonkman. We are also grateful to Martina Lindorfer, our shepherd, and our anonymous reviewers for their insightful feedback and suggestions. This research is a joint work from multiple institutions, sponsored in part by DHS/AFRL award FA8750-18-2-0087, NSF grants CNS-1237265, CNS-1406041, CNS-1629973, CNS-1705050, and CNS-1717062.

References

[1] Abuse.ch. https://abuse.ch/.

[2] Top Alexa domains. https://www.alexa.com/topsites/.

[3] Alienvault IP reputation. http://reputation.alienvault.com/reputation.data.

[4] ANTONAKAKIS, M., APRIL, T., BAILEY, M., BERNHARD, M., BURSZTEIN, E., COCHRAN, J., DURUMERIC, Z., HALDERMAN, J. A., INVERNIZZI, L., KALLITSIS, M., ET AL. Understanding the mirai botnet. In *USENIX Security Symposium* (2017).

[5] Badips. https://www.badips.com/.

[6] BENSON, K., DAINOTTI, A., SNOEREN, A. C., KALLITSIS, M., ET AL. Leveraging internet background radiation for opportunistic network analysis. In *Proceedings of the 2015 Internet Measurement Conference* (2015), ACM.

[7] The Bro network security monitor. https://www.bro.org/index.html.

[8] Composite Blocking List. https://www.abuseat.org/.

[9] Spreading the disease and selling the cure. https://krebsonsecurity.com/2015/01/spreading-the-disease-and-selling-the-cure/.

[10] CHACHRA, N., MCCOY, D., SAVAGE, S., AND VOELKER, G. M. Empirically Characterizing Domain Abuse and the Revenue Impact of Blacklisting. In *Proceedings of the Workshop on the Economics of Information Security (WEIS)* (State College, PA, 2014).

[11] Cloudflare, fast, global content delivery network. https://www.cloudflare.com/cdn/.

[12] AWS CloudFront, fast, highly secure and programmable content delivery network. https://aws.amazon.com/cloudfront/.

[13] Cyber Observable eXpression. http://cyboxproject.github.io/documentation/.

[14] DeKoven, L. F., Savage, S., Voelker, G. M., and Leontiadis, N. Malicious browser extensions at scale: Bridging the observability gap between web site and browser. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)* (2017), USENIX.

[15] Durumeric, Z., Bailey, M., and Halderman, J. A. An internet-wide view of internet-wide scanning. In *USENIX Security Symposium* (2014).

[16] Edgecast CDN, Verizon digital and media services. https://www.verizondigitalmedia.com/platform/edgecast-cdn/.

[17] Facebook threat exchange. https://developers.facebook.com/programs/threatexchange.

[18] Fastly managed CDN. https://www.fastly.com/products/fastly-managed-cdn.

[19] Incident Object Description Exchange Format. https://tools.ietf.org/html/rfc5070.

[20] Jagpal, N., Dingle, E., Gravel, J.-P., Mavrommatis, P., Provos, N., Rajab, M. A., and Thomas, K. Trends and lessons from three years fighting malicious extensions. In *USENIX Security Symposium* (2015).

[21] Jung, J., and Sit, E. An empirical study of spam traffic and the use of dns black lists. In *Proceedings of the ACM Conference on Internet Measurement* (2004).

[22] Kapravelos, A., Grier, C., Chachra, N., Kruegel, C., Vigna, G., and Paxson, V. Hulk: Eliciting malicious behavior in browser extensions. In *USENIX Security Symposium* (2014), San Diego, CA.

[23] Kührer, M., Rossow, C., and Holz, T. Paint it black: Evaluating the effectiveness of malware blacklists. In *International Workshop on Recent Advances in Intrusion Detection* (2014), Springer.

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

[25] MaxCDN. https://www.maxcdn.com/one/.

[26] Metcalf, L., and Spring, J. M. Blacklist ecosystem analysis: Spanning jan 2012 to jun 2014. In *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security* (2015), ACM.

[27] Nothink honeypot SSH. http://www.nothink.org/honeypot_ssh.php.

[28] Packetmail.net. https://www.packetmail.net/.

[29] Pang, R., Yegneswaran, V., Barford, P., Paxson, V., and Peterson, L. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement* (2004), ACM.

[30] Pitsillidis, A., Kanich, C., Voelker, G. M., Levchenko, K., and Savage, S. Taster's Choice: A Comparative Analysis of Spam Feeds. In *Proceedings of the ACM Internet Measurement Conference* (Boston, MA, Nov. 2012), pp. 427–440.

[31] Ramachandran, A., Feamster, N., Dagon, D., et al. Revealing botnet membership using dnsbl counter-intelligence. *SRUTI 6* (2006).

[32] Ramachandran, A., Feamster, N., and Vempala, S. Filtering spam with behavioral blacklisting. In *Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS)* (2007).

[33] Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S. D., and Vallina-Rodriguez, N. A long way to the top: Significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference* (2018), ACM.

[34] Shadowserver. https://www.shadowserver.org/.

[35] Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., and Zhang, C. An empirical analysis of phishing blacklists. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)* (2009).

[36] Sinha, S., Bailey, M., and Jahanian, F. Shades of grey: On the effectiveness of reputation-based "blacklists". In *2008 3rd International Conference on Malicious and Unwanted Software (MALWARE)*, IEEE.

[37] Structured Threat Information eXpression. https://stixproject.github.io/.

[38] UCSD network telescope. https://www.caida.org/projects/network_telescope/.

[39] The spam and open relay blocking system. http://www.sorbs.net/.

[40] The Spamhaus block list. https://www.spamhaus.org/sbl/.

[41] The Spamhaus Don't Route Or Peer Lists. https://www.spamhaus.org/drop/.

[42] Thomas, K., Amira, R., Ben-Yoash, A., Folger, O., Hardon, A., Berger, A., Bursztein, E., and Bailey, M. The abuse sharing economy: Understanding the limits of threat exchanges. In *International Symposium on Research in Attacks, Intrusions, and Defenses* (2016), Springer.

[43] Threat intelligence market analysis by solution, by services, by deployment, by application and segment forecast, 2018 - 2025. https://www.grandviewresearch.com/industry-analysis/threat-intelligence-market.

[44] University of Oregon route views project. http://www.routeviews.org/routeviews/.

[45] VirusTotal. https://www.virustotal.com/#/home/upload.