Blacklist Ecosystem Analysis

Spanning Jan 2012 to Jun 2014

Leigh Metcalf
CERT Division, Software Engineering Institute
Carnegie Mellon University
lbmetcalf@cert.org

Jonathan M. Spring
CERT Division, Software Engineering Institute
Carnegie Mellon University
jspring@cert.org

ABSTRACT

Motivation: We compare the contents of 86 Internet blacklists to provide a view of the whole ecosystem of blocking network touch points and blacklists. We aim to formalize and evaluate practitioner tacit knowledge of the fatigue of playing "whack-a-mole" against resilient adversary resources.

Method: Lists are compared to lists of the same data type (domain name or IP address). Different phases of the study use different comparisons. Comparisons include how many lists an indicator is unique to; list sizes; expanded list characterization and intersection; pairwise intersections of all lists; and *following*, a statistical test we define to determine if one list adds elements shortly after another.

Results: Based on a synthesis of multiple methods, domain-name-based indicators are unique to one list 96.16% to 97.37% of the time. IP-address-based indicators are unique to one list 82.46% to 95.24% of the time.

Discussion: There is little overlap between blacklists. Though there are exceptions, the intersection between lists remains low even after expanding each list to a larger neighborhood of related indicators. Few lists consistently provide content before other lists if there is intersection. These results suggest that each blacklist describes a distinct sort of malicious activity and that even merging all lists there is no global ground truth to acquire. Practical insights include (1) network defenders are advised to obtain and evaluate as many lists as practical, (2) "whack-a-mole" is inevitable due to list dynamics, barring a strategic change, an (3) academics comparing their results to one or a few blacklists to test accuracy are advised to reconsider this validation technique.

WICS 2015 Oct 12, 2015, Denver, CO, USA

ACM ISBN 978-1-4503-3822-6/15/10.

DOI: 2808128.2808129

Categories and Subject Descriptors

K.6.5 [**Computing Milieux**]: Management of Computing and Information Systems—*Security and Protection*

Keywords

Blacklists; information sharing; security models; network security

1. MOTIVATION

Blacklists, also known as block lists, threat intelligence feeds, or threat data feeds, are lists of indicators used to deny access to certain parts of the Internet by network defenders. Most organizations connected to the Internet employ a blacklist of some kind to detect and prevent unwanted communications with adversaries. Between public lists, closed-group information exchanges, and proprietary services, there are hundreds of such blacklists available today.

A open challenge for security personnel is to create high-fidelity intelligence from the data that is shared via blacklists. The blacklists and their providers vary widely. Security personnel must evaluate and prioritize which lists to implement and use. We take a step back from the blacklist evaluation cycle to examine the whole blacklist ecosystem. Blacklist ecosystem analysis cannot evaluate individual list effectiveness. However, quantifiable properties of the ecosystem of blacklists, the interrelationships among lists, and the features of all lists must inform any strategy for deriving high-fidelity intelligence.

Prior work on effective information sharing and collaborative security has focused on areas such as the design of the data sharing system [2, 15], creating repositories [14, 16], or social analysis of security personnel [24]. A strategic analysis of the blacklist ecosystem is a study of how to make sense of what is currently shared and what the entities should share. Prior work has touched on blacklist effectiveness and suggested that blacklists are incomplete [18]. Prior blacklist overlap work has been limited to seven days and nine lists [27], which we consider a tactical analysis rather than a strategic one. A broader, strategic assessment of the value of blacklist data is needed to contextualize and frame the challenge of generating high-fidelity intelligence.

There is little public information about how blacklist producers create their lists. This secrecy is justified because disclosure of the precise procedure of generating the lists likely lets the adversaries avoid detection. However, this secrecy does not benefit the operational analyst who must decide which lists to apply on which network access control points and is often left making semi-educated guesses about the providence and usefulness of a list in a particular situation. We previously identified this interaction between the (list) architect, user, and adversary as requiring further study [22], and the blacklist ecosystem helps to inform that broader effort.

From an operational point of view, the question is quite practical. Network defenders need to know which lists they should use to defend their networks. Evaluating individual lists is not generally possible because there is no global ground truth about maliciousness. Ecosystem-wide views of blacklist interaction are informative for the practitioner. If no lists overlap and few mimic one another, then the strategy would appear to be to acquire all lists, since they all contain unique value.

Blacklist interrelation affects the information security evaluation and baseline creation as well. Academic and industry papers often rate performance of a particular task according to its agreement with some blacklist or lists. If all lists were equal or generation methods open, this method would be acceptable. However, because each list is different and largely non-overlapping, the ability to alter results by the choice of list leaves the evaluation process open to manipulation, since an author can choose the list that offers the best agreement.

Prior work suggests that blacklists of domains and IP addresses are untenable as sole defensive measures since the cost of malicious infrastructure is driven down by economic competition [20]. Blacklist ecosystem analysis sheds some light on the accuracy of this model. We can find evidence about this model in the blacklist ecosystem. The model predicts lots of malicious domains; it also predicts that no matter how much effort is spent on blacklisting, it will never catch everything. These elements appear to hold true.

Blacklist ecosystem analysis is one aspect of a larger body of work to quantify strategic cybersecurity issues. The blacklist ecosystem is intimately related to the low cost of domains and infrastructure to adversaries [20], the poor state of repair of consumer devices connected to the Internet that permits abuse [7], the global nature of adversarial capability for information technology [23], the challenges of modeling the interaction between the user and the adversary [22], and the challenges of designing effective and instructive observations in information security [8].

This paper unifies and expands on a series of white papers. Detailed results and extensive tables that do not fit within this paper are available in the white papers [10, 11].

2. METHOD

Basic results include reverse counts, list size measurements, and pairwise intersections. Notable results reported here include which lists appear to be following other lists. Methods for these processes are described in this section.

List acquisition occurred in two phases. Phase one includes 25 blacklists and seven whitelists collected from January 1, 2012 through March 31, 2013. Phase two stops whitelist collection but adds new blacklists up to a total of 85, and spans March 16, 2013 to June 30, 2014. Lists are selected to cover a variety of target behaviors and geographic areas (as purported by the list owners), such as botnet command and control, spam email senders, phishing senders, identifiers within email message bodies, scanning, and malicious download locations. List acquisition covers a consecutive date range of 30 months with some core methods common to both phases. Phases one and two employ some analyses that are run only on each phase, as the results of phase one inform our research questions.

List acquisition has potential inconsistencies. For example, our list acquisition was not constant. Lists were acquired at certain time points, and each list could not be acquired at exactly the same time. This asynchrony makes determining who listed what first difficult; therefore, we worked in units of days when determining “at the same time” and treated anything on the same day as equivalent. In some cases, list providers limited downloads to once per day; whereas others encouraged two or three daily downloads. If an indicator was listed only in between downloads, it would not be observed. We judged that these inconsistencies are not relevant to the granularity at which we are comparing the lists.

Comparison across such large time windows has certain potential pitfalls, especially for IP addresses based on how they are used on the Internet. Over time, IP addresses are reassigned and reused due to features such as NAT, DHCP, BGP, and IP address stewardship or assignment changes from the regional Internet registries (RIRs).

We expect that these mechanisms have a real impact on measurement over more than one year. All of these technical features have the effect of apparently and erroneously increasing the intersection between lists. The increase in intersection is because the same identifier is used by multiple machines, and the lists may be detecting activity from a machine for each identifier it has. Alternatively, if an identifier is shared by multiple machines, two lists may detect distinct behavior from distinct machines, but appear to intersect because those machines share an identifier. These impacts generally serve to make the reverse count analysis an upper bound for how much intersection there is between lists. We account for the effect of this overestimation analytically in Section **??**.

Core analyses run in both phases are reverse counts and list counts (Sections 2.1 and 2.2). Since the results across phases agree, the results from phase-one analysis are considered likely to hold during phase two, and phase-two analysis is used to explore further questions.

Methods unique to phase one include characterizations of the lists by identifier structure and Internet structure data such as passive DNS and BGP. The function of these methods was to determine whether lists were detecting elements in related neighborhoods or on related infrastructure while not detecting rote identical elements.

Methods unique to phase two include exhaustive pairwise intersections and following analysis. Phase two also added 60 new lists to the analysis to expand the reach of the results.

2.1 Reverse Counts

The method used for counting how many indicators are unique to one list, two lists, three lists, etc., is straightforward. Each comparable indicator (i.e., all the IP addresses) is tagged with how many lists contained it. The number of lists per indicator is counted; call it n. The reported result is the number of indicators on n lists for $n = 1$ up to the maximum n observed.

2.2 List Counts

List counts are the total number of unique indicators observed on the list at any time during the observation period. Each list is given an anonymized numeric identifier and labeled either LI for a list of IP addresses or LD for a list of domain names. This naming convention is used wherever lists must be referred to individually. Each list's identifier is the same throughout the report.

2.3 IP-Based Characterization

To characterize the IP address content of lists, we calculate three functions related to the autonomous system number (ASN) responsible for each IP address and intersections between blacklists and white lists. ASN assignment information is derived from open source data [13, 12] using the open-source SiLK toolsuite [3, 25].

"ASN Counts" results display the number of unique ASNs represented by IPs on the blacklist and the total number of IPs that all of those ASNs represent.

"Top 5 Countries by ASN" results take the ASN counts one step further. Each ASN is associated with the country in which the company that owns the ASN is registered. This does not necessarily represent the geolocation of the IP addresses, but it does indicate which countries are in legal control of the host companies. "UNK" represents unknown.

For "ASN Intersections by Count" and "ASN Intersections by Percentage" results, we intersect not by IP address but over the set of ASNs associated with any IP address on each blacklist during phase one. Therefore, if each list has at least one IP address that is owned by the same ASN, that ASN is counted as shared between the list. It does not mean the IP addresses are necessarily the same. Otherwise, intersections are calculated using standard set operations. The percentage reported is the percentage of the smaller list.

2.4 Domain-Based Characterization

There are many different measurements by which to contextualize and cluster list contents during phase one that are unique to domain-based lists. These approaches include labeling with results from Google's Safe Browsing corpus, clustering by TLD, passive DNS analysis to determine which domains are active, and clustering by name server of active domains.

The intersection of each list with Google's Safe Browsing corpus of malicious URLs [5] must be determined using methods unlike the other intersection methods. Google does not make the clear-text URLs available. Rather, Google provides the hashes of entries in a custom data structure and a client for checking URLs against that hash database. We stored all the hashes listed as malicious by Google Safe Browsing from October 1, 2011 through April 30, 2013. To determine if a domain was known by Google, we check to see if the domain tests positive using the provided tool referencing the stored hash lists. Due to the nature of the Safe Browsing data format, we cannot accurately state how many domains are on the Safe Browsing blacklist, thus Safe Browsing results cannot be further compared to the other lists.

The "Name Servers and Domains" results report the number of active domains per blacklist. To be active in the DNS, a domain name must have a name server. We use passive DNS data for phase one to determine the number and percentage of domains on the blacklist that had valid name server (NS) records. We calculate the number of unique name servers for all domains, the average number of name servers a given domain uses, and the number of domains on the list usually associated with any given name server.

The "Top 5 Name Servers" results extract one item from the process for the name servers and domains: the five name servers that served the largest number of domains on each blacklist.

2.5 Expanded List Intersection

To extend the analysis, we attempted to find any latent links between the lists. We did this by expanding the lists to other identifiers that could be considered immediately related to those on the list. The analysis collected about each expanded list was similar to that described in the core analyses.

We use a commercial passive DNS source as the data for the expansion and follow our indicator expansion method [21]. Every phase-one list was expanded one step. For lists that contained IPs, the IP addresses were expanded to domains ("ID"); for lists that contained domains, the domains were expanded to IP addresses ("DI").

The reason for this expansion is to gather information about whether lists are monitoring similar resources but reporting more specific results. One IP address may commonly host 10,000 domains. In the rare case they are all malicious, one list may contain half, and another list may contain the other half, making it appear that the lists are disjoint. However, after one "DI" expansion, each list would contain the same, single IP, indicating that these lists are more related than initially believed.

The method we use is the most permissive so that we do not miss any relationship. By permissive we mean the expansion is not filtered to remove false positives, as it would normally be when pivoting for operational CND. Thus, the intersections provided by this method are an overestimate.

We also compare the expanded "DI" lists to the phase-one lists containing IP addresses. The result is some measure of similarity between the domain-based lists and the IP-based lists. The lists were compared in pairs of two, one list from the original IP-based lists, and one list from the "DI" expanded lists. The five pairs with the largest rote intersection cardinality and largest intersection based on the percentage of the smaller list are reported for these comparisons.

Phase-one lists that contain both domains and IPs are compared to themselves. In this case, the domains from such lists were extracted, a "DI" expansion was performed, and the IP addresses derived from expansion were compared to the IP addresses on the same list. The percentage of IPs derived from the expansion that also appeared on the original list is reported.

2.6 Time Series

This method examines the timing differences between when different phase-one lists published certain identifiers. The results contain five data elements, for a single pair of lists in the relevant time frame. For each list pairing, we report

- rote intersection size
- percentage relative to each of the two lists
- list A and how many identifiers it published before list B
- list B and how many identifiers it published before list A
- how many identifiers were published on the same day

The values are calculated by finding the intersection between the two lists for the relevant time period (2012 or 2013Q1) and comparing the date that each list published each identifier that they share in common. Due to our collection infrastructure, the variable rate at which lists publish data and analysis practicalities, the extent of our granularity for the comparison was one calendar day. Each identifier was considered to be published on a date if the UTC time of publication to the list fell on that date.

2.7 Pairwise Intersection Counts

Each possible pairing of phase-two lists is generated and the cardinality of the intersection between the two sets is reported. With 18 domain-name-based lists, there are $\binom{18}{2}$ or 162 pairings. With 67 IP-address-based lists there, are $\binom{67}{2}$ or 2244 pairings.

2.8 Following

We define *following* to be if two blacklists contain similar content during the same time period where one list consistently lists elements earlier. Overlap could be due to similar search strategies or outright copying. More formally, two lists are not following if any intersection is essentially random, with as many elements discovered first by list 1 as by list 2. To test for following, we performed a one-sample t-test on phase-two lists that had intersections of greater than 1000 elements. We test whether the average difference between shared element discovery times is 0, because this is true if a list finds as many elements earlier as later. If we can reject this null hypothesis of a 0 mean, we have reason to believe that one list is following the other. We calculate this determination on the granularity of one calendar day, not per second, due to our coarse collection schedule.

The t-test is calculated as follows. For each shared element between the lists, a time delta t_Δ is calculated as $t_\Delta = t_1 - t_2$, where t_1 and t_2 are the times list 1 and list 2 published the element, respectively. Over all shared elements, this difference creates a list of deltas t_Δ^1 through t_Δ^n, where n is the number of shared elements; call this set T_Δ. The t-test is set to test that the mean of T_Δ is 0, so we set $\mu_0 = 0$. We calculate $\bar{x}$ as the mean of T_Δ and s as the standard deviation of T_Δ. The value of the t-test for each list pairing is calculated as in Equation 1:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{1}$$

The p-value is calculated by the standard single-value, two-tailed t-test based on the degrees of freedom $n-1$. The result is the probability p that the experimental results are observed by chance even though the null hypothesis is true if we repeated the same experiment. There is only one blacklist ecosystem, so we must test certainty this way rather than repeating the measurement. We discuss what it means for the null hypothesis to be false ($\bar{x} \neq 0$) in Section **??**.

A summary goal is to report on the number of indicators involved in a nonzero-mean relationship between two lists. We are unaware of a precedent for what should be considered a reasonable p-value in such science of security work. Initially, we tested a p-value of 0.01. At this value, we failed to reject the null hypothesis for 2 of 21 domain-name-based intersections and 54 of 859 IP-address-based intersections (i.e., most results were significant).

However, after inspecting the results we fear this choice of P risked a high type I error (α). When summarizing the results, we set a more aggressive p-value for certainty that the mean was nonzero: 2.2×10^{-16}. Different fields of inquiry tend to tune p-values from the de facto 0.01 to the practicalities of their field, for example particle physics customarily uses a p-value on the order of 10^{-7} [17, p. 9]. We chose the most aggressive threshold that R can report. However, since we cannot re-run the test this year (there is only one blacklist ecosystem), the results should be considered as exploratory analysis rather than a formal hypothesis test.

We only considered pairwise intersections with more than 1,000 elements to ensure that the sample was robust and to help control for anomalous small intersections. The indicators from any pairwise intersection that pass this test have some non-random relationship. Each pairwise intersection provides indicators; we report on the total unique indicators involved in any such potential following relationship by reporting the cardinality of the union of the set of indicators involved in any pairwise intersection passing this test.

3. RESULTS

The results presented in this section are more concise than the results from the 2013 report. This conciseness is partly because the results are largely compatible with prior results and so do not need to be repeated. Furthermore, since the number of lists analyzed increased to 85, we cannot report as many detailed results and need to focus more on summarizing the results in meaningful ways.

For example, we checked to see if any of the blacklisted IP addresses were known sinkhole IP addresses. This information would essentially invalidate the indicator as an indicator of malicious activity, since sinkholes are operated by network defenders who clean up and collect intelligence on threats. Only one list out of 67, LI_3, contained any sinkhole IP addresses and that list contained only 10.

All the reported results are meant to inform the extent of uniqueness of black lists. The reverse counts indicate how frequently indicators appeared on multiple lists. List counts give a sense of the variety of lists involved. Phase one results (Section 2.3 through Section 2.6) are merely highlighted; the full 280 pages of results are available for further analysis [10]. Pairwise intersections provide a more-detailed look at how large the intersection is between each pair of lists, demonstrating that a few lists overlap quite a lot. The analysis of "following" attempts to quantify these pairwise interactions to determine whether there is a reliable cause or predictable ordering of which list produces an indicator first, or if the two lists just happen to be listing the same indicators essentially randomly.

3.1 Reverse Counts

Since reverse counts are a core analysis, we present sets of results for phases one and two.

3.1.1 Phase One (01/01/2012 to 31/03/2013)

Reverse counts for phase one are broken up into domains and IP addresses and CY2012 and 2013Q1 (January 1 to March 31, 2013).

Domain-based lists for phase one demonstrate little variability based on the duration of the time frame, results in Table 1 and Table 2. The intersection during 2013Q1 is higher by 1 percentage point. It is not clear if this result is a real trend; the phase-two results for domains are directly between these two results, so it is likely not a real upward trend.

Table 1: Reverse Counts – 2012 Domains

# Lists	Count	Percentage
1	13,680,233	96.6043%
2	314,570	2.2214%
3	159,025	1.1230%
4	3,933	0.0278%
5	3,186	0.0225%
6	156	0.0011%
7	5	0.0000%

These two time frame lengths show somewhat different behavior for IP addresses. Table 3 shows that 66% of IPs were unique to one blacklist when the measurement period was a year, but during the shorter period of Table 4 74% were unique to one list.

Table 2: Reverse Counts – 2013Q1 Domains

# Lists	Count	Percentage
1	7,516,207	95.6611%
2	258,820	3.2941%
3	67,426	0.8582%
4	13,870	0.1765%
5	790	0.0101%
6	7	0.0001%

Table 3: Reverse Counts – 2012 IPs

# Lists	Count	Percentage
1	42,799,153	66.0519%
2	11,736,822	18.1134%
3	5,216,018	8.0499%
4	2,582,977	3.9863%
5	1,341,547	2.0704%
6	646,923	0.9984%
7	285,919	0.4413%
8	129,623	0.2000%
9	54,218	0.0837%
10	2,825	0.0044%
11	150	0.0002%
12	29	0.0000%
13	4	0.0000%

3.1.2 Phase Two (16/03/2013 to 30/06/2014)

For domain names, 30,784,571 total unique indicators were observed during the 15-month observation period. There were 29,602,108 indicators observed on exactly one list. There were 1,182,463 domain names observed on multiple lists, or 3.84% of all observed domain-name indicators. Of the indicators that appeared on multiple lists, 780,162 indicators appeared on exactly two lists, or 66% of the indicators that appeared more than once. Table 5 displays the complete results for how often domain-name indicators appeared on multiple lists.

For IP addresses, 121,921,509 total unique IP address indicators were observed during the 15-month observation period. There were 100,532,890 indicators observed on exactly one list. There were 21,388,619 IP address indicators observed on more than one list, or 17.54%, with almost half of those (10,412,833) occurring on exactly two lists. Table 6 displays the complete results for how often IP-address indicators appeared on multiple lists.

3.2 List Counts

The size of the lists surveyed varies widely. Since phase one lists are a subset of the phase two lists, we report the sizes of the lists only during phase two (due to space constraints). Some lists have over ten million indicators, some have less than a thousand, and most are in between. The list names are anonymized and given a random identifier; LD indicates a list of domains, whereas LI indicates a list of IP addresses. Results are based on the number of unique identi-

Table 4: Reverse Counts – 2013Q1 IPs

# Lists	Count	Percentage
1	17,123,159	74.3910%
2	3,502,662	15.2172%
3	1,266,362	5.5017%
4	543,050	2.3593%
5	272,854	1.1854%
6	147,747	0.6419%
7	80,819	0.3511%
8	49,435	0.2148%
9	30,786	0.1337%
10	875	0.0038%
11	30	0.0001%
12	3	0.0000%

Table 5: Reverse count of the number of times each phase-two domain is on domain-based blacklists. $(30,784,571$ **total domains on 18 lists over 15 months.)**

# Lists	Count	Ratio
1	29602108	0.96158910
2	780162	0.02534263
3	163768	0.00531981
4	94065	0.00305559
5	67677	0.00219841
6	41195	0.00133817
7	21702	0.00070496
8	9401	0.00030538
9	3420	0.00011109
10	920	0.00002989
11	138	0.00000448
12	14	0.00000045
13	1	0.00000003

Table 6: Reverse count of the number of times each phase-two IP address is on IP-address-based blacklists, max was 1 IP on 38 lists. $(121,921,509$ **total IP addresses on 67 lists over 15 months.)**

# Lists	Count	Ratio
1	100532890	0.82457058
2	10412833	0.08540604
3	3699338	0.03034196
4	2153492	0.01766294
5	1407801	0.01154678
6	986683	0.00809277
7	716422	0.00587609
8	531285	0.00435760
9	392986	0.00322327
10	288769	0.00236848
11	211412	0.00173400
12	153286	0.00125725
13	111568	0.00091508
14	81692	0.00067004
15	60492	0.00049616
16	45576	0.00037381
17	33681	0.00027625
18	25552	0.00020958
19	19157	0.00015713
20	14568	0.00011949
21	11246	0.00009224
22	8514	0.00006983
23	6662	0.00005464
24	5309	0.00004354
25	3990	0.00003273
26	2798	0.00002295
27	1674	0.00001373
28	995	0.00000816
29	429	0.00000352
30+	409	0.00000335

fiers observed over the 15-month observation period, regardless of how long the identifier was on the list. Table 7 provides the sizes of all lists of domain-name-based indicators. Table 8 provides the sizes of all lists of IP-address-based indicators.

3.3 IP-Based Characterization

Characterizing the IP addresses on the blacklists by ASN demonstrates some overlap. The three lists with the most ASN overlap have 96% of all their represented ASNs in common. However, even at such a coarse granularity, the commonality does not hold; many pairs of lists do not have many ASNs in common. The largest ASN overlap between 15 IP-based is 19% of ASNs represented by the lists. This result is larger than the number of rote IP addresses shared, but active ASNs are far fewer and larger than individual IPs.

3.4 Domain-Based Characterization

Table 9 summarizes passive DNS activity for 2012. The variability in percentage of list active is more likely a spurious effect of list collection. However, the relation of number of name servers and domains reflects how reliant adversaries are on centralized DNS infrastructure. Since few domains use many name servers, it appears adversaries are not heavily reliant on single-name servers. The highly common name servers are mainly commercial servers for registries, which serve enough domains that many happen to be malicious.

Google Safe Browsing had a low intersection rate with almost all lists including the ID expansion of the IP-address based lists. There were two exceptions to this trend: two domain-based lists had 41% and 22% of their domains, respectively, known to Safe Browsing in 2013 and 35% and 29%, respectively, in 2012.

3.5 Expanded List Intersection

Some lists follow similar malicious resources differently; however, it is rare. There is only one clear instance of a pair of lists that do not overlap much rote but do overlap after expansion. These are list_12 and list_03 in the white paper;

Table 7: Unique domains per phase-two list.

List	Unique Entries	List	Unique Entries
LD_1	411871	LD_10	251044
LD_2	24103937	LD_11	2802602
LD_3	55110	LD_12	1442233
LD_4	83884	LD_13	173
LD_5	73351	LD_14	2738773
LD_6	47790	LD_15	61424
LD_7	67025	LD_16	2559
LD_8	3498	LD_17	178632
LD_9	499358	LD_18	61088

see Table 127 [10]. The IP addresses of the two lists do not overlap much in 2013, but after indicator expansion to domains, there is over 99% overlap between the domains.

On the whole, when expanding the IP-address-based lists to all domains hosted on the IPs in the list, they retain their distinctness. The ID expansion of the 2012 lists still has 77% of domains unique to one of 11 lists, out of 211 million domains, as displayed in Table 10.

3.6 Time Series

The relatively low base intersection rate makes measuring which list is first less certain. There are some phase-one lists that do appear to be consistently faster than those lists that do intersect with them. Feed_04 is the singular example of this. Otherwise, it's not generally clear that one list is consistently earlier or later than another just by taking the average of the number of indicators that appear on one or the other first. This uncertain result leads to our decision to determine "following" more rigorously in phase-two analysis.

3.7 Pairwise Intersections

The results for the pairwise intersections of all lists is quite long. The full results can be found in the appendix of our 2014 white paper [11]. The lists are anonymized following the same pattern as described in Section 3.2.

3.8 Following

Our "following" test fails to reject the null hypothesis if the temporal intersection features between lists appears dependent on the lists' interaction. This interaction may be due to any variable that is influencing one list to consistently contain an indicator before another; our test is agnostic of the cause of the temporally linked interaction.

The total number of unique domain names in the set of lists involved in following interactions is 809,394, or 68.45% of the 1,182,463 indicators that appeared on multiple lists. There were 17 pairwise intersections of domain-name-based lists that contributed to this total, out of 21 total pairwise list intersections with more than 1,000 elements.

The total number of unique IP addresses in a set that failed the hypothesis test of a zero mean for the pairwise intersection is 5,803,501, or 27.13% of the 21,388,619 indicators that appeared on multiple lists. There were 648 pairwise intersections of IP-address-based lists that contributed to this total, out of 859 total pairwise list intersections with more than 1,000 elements.

Table 8: Unique IP addresses per phase-two list.

List	Unique Entries	List	Unique Entries
LI_1	22250	LI_35	32612
LI_2	62884574	LI_36	8565
LI_3	3738277	LI_37	13463
LI_4	863	LI_38	32294176
LI_5	72644	LI_39	2093
LI_6	16024	LI_40	359251
LI_7	18878208	LI_41	351799
LI_8	10378	LI_42	3552898
LI_9	615914	LI_43	522814
LI_10	5858	LI_44	171776
LI_11	51309	LI_45	776793
LI_12	3024492	LI_46	444116
LI_13	551965	LI_47	246350
LI_14	134890	LI_48	11145061
LI_15	2355	LI_49	9638563
LI_16	3462	LI_50	4309163
LI_17	6795	LI_51	689524
LI_18	60403	LI_52	703105
LI_19	4432	LI_53	4200727
LI_20	10975	LI_54	2342
LI_21	5738359	LI_55	58097
LI_22	160605	LI_56	25068
LI_23	1142022	LI_57	4201662
LI_24	2702	LI_58	4514
LI_25	119353	LI_59	1752202
LI_26	40051	LI_60	53189
LI_27	1448865	LI_61	1261
LI_28	597228	LI_62	25418
LI_29	58707	LI_63	255558
LI_30	3794	LI_64	4418
LI_31	1746662	LI_65	8048
LI_32	10756	LI_66	4027
LI_33	3705188	LI_67	3955
LI_34	44729		

4. DISCUSSION

There are many common blacklists that describe indicators of malicious activity for the Internet. These lists generally do not intersect. Therefore, it appears that these lists do not converge on one set of malicious indicators. There are not obvious subsets of convergent lists either, such as might be explained by communities of lists tracking similar kinds of malicious behavior, such as phishing senders. For comprehensive detection, it is better to consider all the lists together than to rely on an intersection.

This result of relatively small intersection is consistent with recent results about overlap among open-source cyber-intel indicators [26]. It is also consistent with independent

Table 9: Passive DNS activity of domains on blacklists, 2012. N per D is namer servers per domain, D per N is domains per name server.

# Active	% of List	# of NS	N per D	D per N
2,398,180	86.7231	733,140	3.4176	11.1795
307,472	86.8342	207,608	4.4252	6.5539
265,080	64.5182	207,137	4.6222	5.9152
187,800	96.8426	130,131	3.6899	5.3251
200,320	93.9292	121,610	3.7458	6.1703
13,867	91.1104	17,182	4.5257	3.6525
2,345	7.6916	5,541	4.8162	2.0383
2,033	0.0181	4,431	4.4929	2.0614
447	12.1965	1,927	5.7584	1.3358
19	3.9583	71	5.0000	1.3380

Table 10: Reverse count of domains unique to one expanded ID list source, (out of $211,852,820$ **total domains and 11 lists, 2012).**

# Lists	Count	Percentage
1	163,502,488	77.1774%
2	31,006,423	14.6358%
3	5,433,562	2.5648%
4	3,681,132	1.7376%
5	2,973,937	1.4038%
6	1,966,187	0.9281%
7	2,123,193	1.0022%
8	1,127,653	0.5323%
9	33,737	0.0159%
10	3,728	0.0018%
11	780	0.0004%

results from larger studies by industry [1, p. 9] and academics [9]. Due to the long duration of our study, the conservative assumptions we make, and the independent corroboration by three other sources, we are highly confident this is a genuine characteristic of the global blacklist ecosystem.

It is important to accurately measure the extent of the existing overlap. Features of regular Internet operation complicate this calculation. Although IP address movement and reassignment can be accurately estimated for the Internet as a whole, we cannot reliably estimate the probability that any single IP address was reassigned. These mechanisms inflate the amount of intersection by some factor. Such an effect does not compromise our conclusions because the relevant aspect of our conclusion is how little intersection there is between the lists. The unknown intersection inflation factor means our naive measurement is actually only an upper limit on the intersection. Even making the conservative assumption to ignore this reassignment inflation, the conclusion is still strong. Nevertheless, we attempt to use the phase-two analysis to estimate reassignment inflation and thus get a better idea of the true operational intersection between lists.

The number of IP addresses unique to one list drops markedly in phase two, despite adding lists. We believe the pressure on adversaries to avoid black lists has driven them to use new infrastructure faster during this time frame. During 2012 into 2013 is when the phenomena of crimeware-as-a-service [19] and exploit-as-a-service [6] were first reported, so this timing aligns with our observations.

Competition among list vendors contributes to genuine intersection among lists. If an indicator is on two lists and one list followed the other then the lists genuinely overlap on that indicator. The following test indicates only that there is some relationary factor we have not accounted for. Random factors such as inflationary Internet features, like DHCP and NAT, should not usually cause this "following" behavior. We consider intersection due to "following" relationships as the lower bound on the genuine intersection among lists because it definitely excludes inflationary features.

The naive reverse counts do not account for any inflationary Internet features. Our "following" test is likely too strict and undercounts the duplicative results from lists because of the low p-value used in the test and the artificial limit of testing only intersections with at least 1,000 indicators. Therefore, we believe the genuine result is somewhere in the range created by the two methods.

An indicator provides unique CND value unless that indicator duplicates information on another list, where duplication cannot simply be due to indicator churn or Internet architecture changes. The range of unique value to CND from an indicator on domain-name-based lists is narrower than that for IP-address-based lists, but both ranges indicate highly unique indicators. Phase-two domain-name-based indicators do not provide unique value to CND between 2.63% and 3.84% of the time. That is, between 96.16% and 97.37% of domain-name-based indicators are uniquely provided by a single source. Phase-two IP-address-based indicators do not provide unique value to CND between 4.76% and 17.54% of the time. This wider range for IP-address-based lists is expected because there are fewer IP addresses than domain names, and because IP addresses are more commonly reused. As is evident from these two ranges, the large majority of the time, any list's indicator will provide unique information and value to CND.

Thus far, we have discussed the facts of what we observe about the blacklist ecosystem. The natural next question is why is the ecosystem this way. Our current explanation is that each list, or perhaps a pair of related lists, describes and follows a specific type of malicious behavior with a specific detection strategy from a particular sensor vantage [4]. These three axes of variation provide a very large and sparsely populated space for the blacklists in the ecosystem. Like ecological niches, they are incentivized to spread out to areas of little competition.

Further, one must hold two of these three aspects fixed to perform an intelligible comparison between two lists. Otherwise which aspect causes the variation is not deducible. This is an internal validity challenge in experiment design on engineered mechanisms [8]. In practice, it is not possible to hold two of these three aspects fixed. Therefore, it is not practical to rigorously compare one list to another list. Since comparison and benchmarking is not practicable, there is no thorough or convenient way to evaluate the performance of any of these lists. Any baseline is properly understood as just another blacklist, with different detection strategy and sensor vantage used to create the baseline. Each list is therefore currently best understood as a one-of-a-kind authority on the particular type of activity it detects.

This problem is especially acute for academic researchers attempting to prove their method is accurate by comparing their results to known lists. Most lists do not intersect, so what a researcher considers to be a "good" rate of intersection to prove a research method accurate may be meaningless. Further, it is important to consider which lists are used as benchmarks, since so few common public lists intersect.

A further difficulty with this situation is that there is no ready taxonomy or terminology for describing precisely what activity a malicious actor is performing. Attempts to categorize a list as following a particular malicious activity will run into terminology and communication issues between researchers. The best way to determine what malicious activity a list is following is to know what detection algorithm the list uses; however, list detection algorithms are justifiably almost never shared. This leaves both the academic and operational cybersecurity community with few resources to evaluate efficacy.

The CND lesson is that any one list, or any ten lists, cannot provide a comprehensive description of all malicious indicators. Every list the defender can obtain and use will probably continue to provide new, non-overlapping defense to the network. Though the defender must evaluate the quality of new identifiers, any new list can provide useful identifiers of malicious activity not already contained in the defender's list. This lesson implies that CND organizations should share indicators as quickly as possible with as many trusted partners as possible. Quick sharing will not eliminate the threat; however, it will drive the adversaries to use new infrastructure for every attack rather than being able to reuse infrastructure to attack multiple organizations.

A CND analyst or architect can also conclude that blacklists are insufficient for adequate network defense. If blocking is so fragile, it is too easy to avoid. Other established methods of CND should be prioritized and put into production as appropriate, such as gray lists, behavioral analysis, web proxy content analysis, criminal penalties, speed bumps, and organization-specific white lists. Such measures will go much further to eliminate the threat, especially if used in conjunction with quickly shared blacklists [20].

These blacklist results likewise challenge threat intelligence analysts. Existing blacklists should be used to examine new threats with caution. Investigations certainly cannot rely only on blacklists for the detection of ongoing activity. Reputation and context of larger units of the Internet become increasingly important to get a better idea of what behavior is suspicious. For this task, processes such as intelligent indicator expansion are useful [21].

5. FUTURE WORK

We hypothesize that the dynamics of blacklist ecosystems generalize to any situation in which two conditions hold: defenders are reacting to threats primarily by blocking previously identified indicators, and the attack vector is the result of a digital economy (thus the marginal cost of the vector asymptotes to zero). For example, we expect that similar non-overlap occurs in anti-virus signatures, which are essentially blacklists for files.

There have been some efforts to determine the completeness of blacklists [9]. It is not clear whether the methods used are simply comparisons to yet another blacklist, and thus whether they are trustworthy. Intuitively, if so few lists overlap, it is unlikely that the set of them is complete. However, more work in data analysis and statistics, such as extending the capture-recapture population estimation techniques, needs to be completed before we can be sure about the completeness of the blacklist ecosystem.

Acknowledgements

Copyright 2015 ACM

This material is based upon work funded and supported by Department of Homeland Security under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.

This material has been approved for public release and unlimited distribution.

Carnegie Mellon®, CERT® and CERT Coordination Center® are registered marks of Carnegie Mellon University.

6. REFERENCES

[1] 2015 data breach investigations report (DBIR). Tech. rep., Verizon, 2015.

[2] BURGER, E. W., GOODMAN, M. D., KAMPANAKIS, P., AND ZHU, K. A. Taxonomy model for cyber threat intelligence information exchange technologies. In *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security* (2014), ACM, pp. 51–60.

[3] CERT/NETSA AT CARNEGIE MELLON UNIVERSITY. CERT/CC Route Views Project Page. [Accessed: Feb 13, 2014].

[4] COLLINS, M. Using vantage to manage complex sensor networks. In *FloCon 2015, 11th Annual* (Portland, OR, January 2015), Software Engineering Institute, Carnegie Mellon University.

[5] GOOGLE. Google Safe Browsing FAQ. http://code.google.com/apis/safebrowsing/safebrowsing_faq.html, November 2, 2011.

[6] GRIER, C., BALLARD, L., CABALLERO, J., CHACHRA, N., DIETRICH, C. J., LEVCHENKO, K., MAVROMMATIS, P., MCCOY, D., NAPPA, A., PITSILLIDIS, A., PROVOS, N., RAFIQUE, M. Z., RAJAB, M. A., ROSSOW, C., THOMAS, K., PAXSON, V., SAVAGE, S., AND VOELKER, G. M. Manufacturing compromise: The emergence of exploit-as-a-service. In *Proceedings of the 2012 ACM Conference on Computer and Communications*

Security (Raleigh, North Carolina, USA, 2012), CCS '12, ACM, pp. 821–832.

[7] HALLENBECK, C., KING, C., SPRING, J. M., AND VIXIE, P. Abuse of customer premise equipment and recommended actions. In *Black Hat USA 2014* (Las Vegas, Nevada, Aug 7, 2014), UBM.

[8] HATLEBACK, E., AND SPRING, J. M. Exploring a mechanistic approach to experimentation in computing. *Philosophy & Technology 27*, 3 (2014), 441–459.

[9] KÜHRER, M., ROSSOW, C., AND HOLZ, T. Paint it black: Evaluating the effectiveness of malware blacklists. Tech. Rep. TR-HGI-2014-002, Ruhr-Universität Bochum, Horst Görtz Institute for IT Security, June 2014.

[10] METCALF, L. B., AND SPRING, J. M. Everything you wanted to know about blacklists but were afraid to ask. Tech. Rep. CERTCC-2013-39, Software Engineering Institute, CERT Coordination Center, Pittsburgh, PA, 2013.

[11] METCALF, L. B., AND SPRING, J. M. Blacklist ecosystem analysis update: 2014. Tech. Rep. CERTCC-2014-82, Software Engineering Institute, CERT Coordination Center, Pittsburgh, PA, December 2014.

[12] RIPE NETWORK COORDINATION CENTER. Routing information service (RIS). http://www.ripe.net/data-tools/stats/ris/routing-information-service, January 3, 2012.

[13] ROUTE-VIEWS. University of oregon route views project. http://www.routeviews.org, January 3, 2012.

[14] SCHEPER, C., CANTOR, S., AND MAUGHAN, D. Predict: a trusted framework for sharing data for cyber security research. In *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security* (2011), ACM, pp. 105–106.

[15] SERRANO, O., DANDURAND, L., AND BROWN, S. On the design of a cyber security data sharing system. In *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security* (2014), ACM, pp. 61–69.

[16] SHARMA, V., BARTLETT, G., AND MIRKOVIC, J. Critter: Content-rich traffic trace repository. In *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security* (2014), ACM, pp. 13–20.

[17] SINERVO, P. K. Signal significance in particle physics. In *Advanced statistical techniques in particle physics* (Durham, UK, 2002), M. R. Whalley and L. Lyons, Eds.

[18] SINHA, S., BAILEY, M., AND JAHANIAN, F. Shades of grey: On the effectiveness of reputation-based "blacklists". In *Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on* (2008), IEEE, pp. 57–64.

[19] SOOD, A. K., AND ENBODY, R. J. Crimeware-as-a-service: A survey of commoditized crimeware in the underground market. *International Journal of Critical Infrastructure Protection 6*, 1 (2013), 28–38.

[20] SPRING, J. M. Modeling malicious domain name take-down dynamics: Why eCrime pays. In *IEEE eCrime Researchers Summit* (September 17, 2013), Anti-Phishing Working Group.

[21] SPRING, J. M. A notation for describing the steps in indicator expansion. In *IEEE eCrime Researchers Summit* (September 17, 2013), Anti-Phishing Working Group.

[22] SPRING, J. M. Toward realistic modeling criteria of games in internet security. *Journal of Cyber Security & Information Systems 2*, 2 (2014), 2–11.

[23] SPRING, J. M., KERN, S., AND SUMMERS, A. Global adversarial capability modeling. In *IEEE eCrime Researchers Summit* (Barcelona, May 28, 2015), Anti-Phishing Working Group, pp. 22–42.

[24] SUNDARAMURTHY, S. C., MCHUGH, J., OU, X. S., RAJAGOPALAN, S. R., AND WESCH, M. An anthropological approach to studying csirts. *IEEE Security & Privacy*, 5 (2014), 52–60.

[25] THOMAS, M., METCALF, L., SPRING, J. M., KRYSTOSEK, P., AND PREVOST, K. Silk: A tool suite for unsampled network flow analysis at scale. In *IEEE BigData Congress* (Anchorage, AK, July 2014), IEEE.

[26] TROST, R. Threat intelligence library - a new revolutionary technology to enhance the soc battle rhythm! In *Black Hat USA 2014* (Las Vegas, Nevada, Aug 7, 2014), UBM.

[27] ZHANG, J., CHIVUKULA, A., BAILEY, M., KARIR, M., AND LIU, M. Characterization of blacklists and tainted network traffic. In *Passive and Active Measurement*, M. Roughan and R. Chang, Eds., vol. 7799 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 218–228.