Evaluating Third-Party Bad Neighborhood Blacklists for Spam Detection

Giovane C. M. Moura, Anna Sperotto, Ramin Sadre, and Aiko Pras
University of Twente
Centre for Telematics and Information Technology (CTIT)
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
Design and Analysis of Communications Systems (DACS)
Enschede, The Netherlands
Email: {g.c.m.moura, a.sperotto, r.sadre, a.pras}@utwente.nl

***Abstract*—The distribution of malicious hosts over the IP address space is far from being uniform. In fact, malicious hosts tend to be concentrated in certain portions of the IP address space, forming the so-called Bad Neighborhoods. This phenomenon has been previously exploited to filter Spam by means of Bad Neighborhood blacklists. In this paper, we evaluate how much a network administrator can rely upon different Bad Neighborhood blacklists generated by third-party sources to fight Spam.**

One could expect that Bad Neighborhood blacklists generated from different sources contain, to a varying degree, disjoint sets of entries. Therefore, we investigate (i) how specific a blacklist is to its source, and (ii) whether different blacklists can be interchangeably used to protect a target from Spam. We analyze five Bad Neighborhood blacklists generated from real-world measurements and study their effectiveness in protecting three production mail servers from Spam. Our findings lead to several operational considerations on how a network administrator could best benefit from Bad Neighborhood-based Spam filtering.

I. Introduction

The distribution of malicious hosts over the IP address space is far from being uniform. In fact, malicious hosts tend to be concentrated in certain portions of the IP address space [1]–[3]. An explanation for this behavior is that networks are managed differently from each other — and poorly managed networks are more likely to have more vulnerable hosts than well managed networks, which would lead to more compromised machines.

In [4], van Wanrooij and Pras exploited this behavior to build an efficient mail filter, coining the term *Internet Bad Neighborhood*. An Internet Bad Neighborhood (BadHood in the rest of this work) is a netblock or prefix (in CIDR notation [5]) of a certain size to which a certain number of misbehaving hosts belong. The idea is that the probability that a particular IP address behaves badly increases if neighbor IP addresses (i.e., hosts within the same prefix) behave badly. Ultimately, this allows to protect a target from *any* host from a certain BadHood (prefix) – which can be seen as a form of *predicting new attacking sources*, and not only to *reacting* to observed /32 sources – which is main benefit of the BadHood approach, as we have shown in Chapter 2 in one of the authors' PhD dissertation [6].

In a following research work [7], we focused on identifying emerging behaviors in BadHoods and also on refining the BadHood definition to the particular case of Spam. After that, we investigated how to meaningfully aggregate BadHoods using network prefixes from /24 to /8 [8].

For our previous studies, we employed real-world available Internet blacklists. Third-party sources can make their lists publicly available on the Internet, such as the Passive Spam Block List (PSBL) [9]. Typically, these lists contain IP addresses that have been observed conducting malicious activities [10]. Alternatively, lists can be obtained by means of private agreements with operators, e.g., ISPs, who keep track of malicious activities in their networks. In this work, we refer to such blacklists as *raw blacklists*. By aggregating the individual IP addresses in the raw lists to netblocks, *BadHood blacklists* can be generated, providing for each /24 netblock the number of malicious IP addresses.

A raw blacklist can only provide a partial view on the Spam activities in the Internet since it only contains IP addresses that have previously contacted the measurement points of the list provider. Therefore, one could expect that BadHood blacklists generated from such raw lists contain, to a varying degree, disjoint sets of BadHood entries. This raises the research questions we investigate in this paper:

1) How specific is a BadHood blacklist to its measurement points?
2) Can we interchangeably use different BadHood blacklists to protect a target from Spam?

The motivation for this study is the following: if most of the BadHoods attacking a target are listed on third-party BadHood blacklists, then the administrator could *effectively* protect the target by using these blacklists in BadHood-based defense mechanisms. This would allow the administrator to choose between the available third-party lists, selecting the one which best suits their requirements, for example, in terms of size and update-policy. Furthermore, spam filters could switch between blacklists in case of unavailability of the main blacklist source[1] without suffering from performance degradation.

To answer our research questions, we consider five raw blacklists containing the IP addresses of spamming hosts. Two lists are public Spam blacklists, generated by distributed Spam traps. The other lists have been generated from the

[1]PSBL users experienced a 4-day outage in November 2011 due to bad weather conditions [9].

Name	# Distinct Entries	Description
Composite Blocking List (CBL) [11]	14,344,508	Spam blacklist containing IP addresses that have reached CBL spamtraps, distributed over different networks and countries. It was used in the following works: [7], [8], [12]–[14]
Passive Spam Blocking List (PSBL) [9]	3,528,855	Spam blacklists of IP addresses that have spammed PSBL distributed traps. Employed in [7], [8]
Provider A	1,668,204	Spam blacklist from mail filter log files of a major hosting provider in the Netherlands
University of Twente/EWI (UT/EWI) [15]	429,197	Spam blacklist generated based on mail filter logs of the Electrical Engineering, Mathematics, and Computer Science Department of the University of Twente
Security Incident Response Team/RNP – (CAIS/RNP) [16]	36,938	Spam blacklist from the mail filter of CAIS of Brazilian Research Network (RNP) mail server, employed in [7]

TABLE I. RAW BLACKLISTS EVALUATED

mail server logs of a hosting provider and two organizations. From these raw lists, we create BadHood blacklists based on /24 netblock aggregation. Using a simple BadHood-based Spam detection procedure, we investigate the capability of the BadHood blacklists to protect target systems from Spam.

The rest of this paper is structured as follows. In Section II, we present the raw blacklists and we describe how we generate BadHood blacklists from them. In Section III, we compare the different BadHood blacklists and study how much they overlap. The effectiveness of the different lists in detecting Spam messages is analyzed in Section V. Finally, conclusions and future work are presented in Section VI.

II. DATA SETS

In this section we describe the data sets used in this paper. We first present five raw blacklists in Section II-A. We then aggregate these raw lists into /24 Bad Neighborhood blacklists in Section II-B.

A. The Raw Blacklists

A large number of public Spam blacklists is available on the Internet. In this work, we have selected a subset of these satisfying the following two criteria: (i) the blacklist has been previously investigated by academic and Internet security communities; and (ii) the organization hosting the list provides bulk-access to the blacklist data, to ensure we have a complete view of the malicious IP addresses. These criteria led us to choose two widely known Spam blacklists relying on distributed spam traps, namely the Composite Block List (CBL [11]) and the Passive Spam Block List (PSBL [9]). The lists contain the IP addresses of hosts that have sent one or more Spam messages to their respective traps.

In addition, we have used three blacklists obtained via private agreement with a Dutch hosting provider (Provider A) and two organizations, the Computer Science department of the University of Twente (UT/EWI) [15], and the Security Incident Response Team of the Brazilian Research Network (CAIS/RNP) [16]. Differently from public blacklists described above, these three lists have been generated from the spam filters logs of single mail servers and, hence, reflect a more "local" view on Spam activity in the Internet (but not employing CBL or PSBL in the process). Furthermore, the mail filter logs allow us to calculate the number of Spam messages per malicious IP. For the UT/EWI data set, we also obtained statistics on the number of legitimate messages (Ham).

The considered raw blacklists cover a monitoring period of one week, from April 19th to April 26th, 2010. Table I shows more details. As can be seen, the lists differ considerably in size. CBL has the largest number of entries, followed by PSBL. In contrast, UT/EWI, and especially CAIS/RNP are relatively small blacklists. As we can observe, the size of a blacklist is related to the number of monitoring points deployed (distributed spam traps vs. individual mail servers). The list of Provider A takes an intermediate position between the two public lists and the two institutional lists. We believe that this is due to the nature of the provider A: as a hosting provider with thousands of customers, it is much more visible to spammers than UT/EWI and CAIS/RNP.

B. The Bad Neighborhood Blacklists

The five raw blacklists have been used to generate Bad Neighborhood blacklists in the following way. We obtain a BadHood blacklist by aggregating the spammers' IP addresses listed in a raw blacklist to /24 BadHoods. As we have discussed in [8], we have chosen /24 because this is the prefix that incurs less aggregation error, and it is also the smallest prefix that can be "routed" on the Internet [17]. For a blacklist source S, we define its BadHood blacklist L_S as

$$L_S = \{\langle B_i, nHosts_S(B_i)\rangle, i = 1 \ldots n_S\} \quad (1)$$

where n_S is the number of different /24 IP prefixes we observe in S, B_i is a /24 IP prefix observed in S, and $nHosts_S(B_i)$ is the number of (spamming) IP addresses in S with prefix B_i. Note that, by definition, $nHosts_S(B_i) > 0$ for all B_i of L_S. Please also note that a BadHood is not necessarily a Class C network, and, hence, the number of hosts can be as large as 256.

In Table II we summarize the high-level characteristics of the BadHood blacklists obtained by the above procedure. The first row in the table shows the number of entries, i.e., the number of /24 BadHoods observed by the lists. We can observe that there is a direct relationship between the size of the raw blacklists and the size of the BadHood lists: Again, the CBL list is the largest, followed by PSBL and the other lists.

	CBL	PSBL	Prov. A	UT/EWI	CAIS/RNP
#BadHoods	1,140,005	732,731	548,866	248,947	34,096
Min(nHosts)	1	1	1	1	1
Max(nHosts)	256	248	227	101	28
Mean(nHosts)	12.58	4.81	3.03	1.72	1.08
Std(nHosts)	29.32	9.44	4.81	1.77	0.44

TABLE II. OVERVIEW ON THE BAD NEIGHBORHOOD BLACKLISTS BASED ON /24 NETBLOCKS

However, Table II also shows that the number of spamming IP addresses (nHosts) per Bad Neighborhood significantly varies among the different lists. Clearly, the CBL list contains not only the most "malicious" BadHood (with 256 spamming IP addresses in a /24 netblock; row "Max(nHosts)" in the table) but it also observes, in average, 2.6 times more spamming hosts per BadHood than PSBL, 4.1 times more hosts than Provider A, and this ratio grows to 7.3 for UT/EWI and 11.6 for CAIS (row "Mean(nHosts)" in the table). The numbers illustrate that there is a high correlation between the average number of spamming hosts per BadHood and the total number of spamming hosts observed. The reason for this is the concentration of spammers in certain parts of the Internet, as studied in [7]. The more spammers are observed, the higher the probability that two spammers are located in the same /24 BadHood.

III. Comparing BadHood Blacklists

The goal of this section is to address the first research question by analyzing the information overlap between different BadHood blacklists. We describe the used methodology and the considered scenario in Section III-A and discuss the results of the overlap analysis in Section III-B.

The second research question will be answered in Section IV, where we study the effectiveness of the different blacklists for Spam filtering.

A. Methodology and Considered Scenario

In the remainder of the paper, we will consider the scenario shown in Figure 1. We regard the mail servers of Provider A, UT/EWI, and CAIS/RNP as targets to be protected from Spam by using a BadHood blacklist (`BL1`, `BL2`, and `BLn`). As the administrator of such a target, we have the choice between the five BadHood blacklists presented in Section II-B. The blacklist of the target itself will serve as reference because it contains exactly those BadHoods that have sent Spam to the target.

We first evaluate to what extent a blacklist L_S and the blacklist L_T of the target contain the same BadHoods. To this purpose, we calculate the *intersection blacklist* $I_{S\cap T}$ which we define as

$$\begin{aligned} I_{S\cap T} \quad = \quad & \{\langle B, nHosts_S(B), nHosts_T(B)\rangle \mid \\ & \langle B, nHosts_S(B)\rangle \in L_S, \\ & \langle B, nHosts_T(B)\rangle \in L_T\}. \end{aligned} \tag{2}$$

In order to quantify the overlap between the two lists, we calculate the overlap ratio

$$v_{S,T} = \frac{|I_{S\cap T}|}{|L_T|}. \tag{3}$$

Note that we will only compare a target's BadHood blacklist with another BadHood blacklist if the latter is larger, i.e., we will not compare the UT/EWI BadHood blacklist to the Provider A target's BadHood list. Because of the size difference, the UT/EWI list will, independently of its content, never achieve a high overlap ratio, and, therefore, would be uninteresting for the network manager of Provider A.

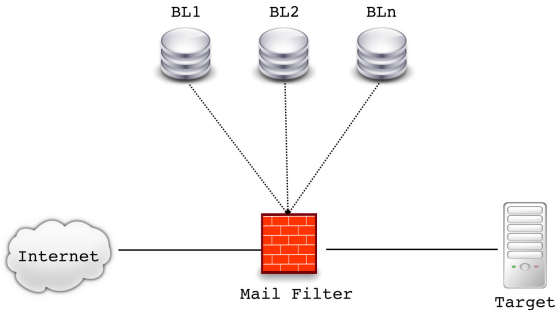


Fig. 1. Blacklist Sources for Target Protection

In addition, we obtain the BadHoods that have exclusively attacked S but not T from $Irr_{S,T} = L_S - (L_S \cap L_T)$ and refer to them as the entries "irrelevant" to T.

Second, we investigate whether the activity of a BadHood in the intersection $I_{S\cap T}$, in terms of number of malicious hosts per BadHood, is perceived differently by S and T and how the badhoods in the non-intersecting part behave (again, in terms of number of malicious hosts). This will become important in Section IV, where we will study a simple spam detection procedure that flags spams based on the activity of the sender BadHood.

B. Intersecting BadHood Blacklists

Table III shows the overlap ratios in percents of the five BadHood blacklists with respect to the BadHood blacklists of the three targets. One row in the table corresponds to one target. The numbers in parentheses give the numbers of BadHoods in the respective intersection blacklists. An overlap ratio close to 100% indicates the target's BadHood blacklist is approximately a subset of the compared BadHood blacklist.

The reported ratios show that the public sources, CBL and PSBL, capture most of the BadHoods that attack individual targets (from 88.03% to 98.74%). From the point of view of the network administrator, this is a very satisfactory result, since it confirms that such public blacklist sources can be relied upon to protect the individual targets. Also, the Provider A blacklist achieves, thanks to its size, high overlap ratios (91.89% and 94.8%) when compared to the UT/EWI and CAIS/RNP. As expected, the much smaller UT/EWI only yields a low ratio of 68.3%.

However, this result comes with a price, namely in the form of irrelevant entries. Table IV shows the number of irrelevant entries for the five BadHood blacklists and the three targets as percentages of the number of entries in the targets' blacklists (the numbers in parentheses give the number of irrelevant entries). We observe that, even though CBL has captured 98.74% of Provider A's BadHoods, CBL has still observed 598,038 BadHoods that did not spam Provider A's mail server — which is more than the size of Provider A's BadHood blacklist itself. This means that when applying a public blacklist source, the network administrator should keep in mind that, while providing a significant match with the target, such a blacklist is likely to observe a larger number of entries that are not observed by the target itself. This can make the list unsuitable for the use in embedded devices such as firewalls integrated in routers.

	CBL	PSBL	Provider A	UT/EWI
Provider A	98.74% (541,967)	88.03% (483,179)	–	–
UT/EWI	97.95% (243,855)	93.13% (231,856)	91.89% (228,777)	–
CAIS/RNP	99.12% (33,799)	96.88% (33,034)	94.8% (32,347)	68.3% (23,316)

TABLE III. OVERLAPS FOR THE THREE TARGETS (AS RATIOS IN %; NUMBER OF ENTRIES IN PARENTHESES)

	CBL	PSBL	Provider A	UT/EWI
Provider A	108.95% (598,038)	45.46% (249,552)	–	–
UT/EWI	359.97% (896,150)	201.19% (500,875)	319,889 (128,49%)	–
CAIS/RNP	3,244.38% (1,106,206)	2052.13% (699,697)	516,519 (1,514.89%)	225,631 (661,75%)

TABLE IV. IRRELEVANT ENTRIES FOR THE THREE TARGETS (AS RATIOS IN %; NUMBER OF ENTRIES IN PARENTHESES)

Finally, we analyze how the activity of the same BadHood, in terms of number of malicious hosts, is perceived by a target and by a third-party BadHood blacklist source. Figure 2 shows for all BadHoods in the intersection $I_{CBL \cap ProviderA}$ the number of spamming hosts monitored by CBL (x-axis) and monitored by Provider A (y-axis). As expected, the much larger CBL blacklist sees more hosts for the same BadHood than Provider A. This phenomenon has been already discussed in Section II-B.

In Figure 3(a), we show the histogram of the number of hosts per BadHood for the BadHoods in the intersection $I_{CBL \cap ProviderA}$, as monitored by CBL and by Provider A. The histogram for the "irrelevant" BadHoods in $CBL - (CBL \cap ProviderA)$ are shown in Figure 3(b). We observe that the blacklist of Provider A, also much smaller, already contains the most malicious BadHoods of the CBL list. The BadHoods not seen by Provider A, i.e., those in the set $CBL - (CBL \cap ProviderA)$, do not show much activity (in terms of number of malicious hosts). Hence, if an administrator had to choose between the CBL list and the Provider A list, one could argue that the latter contains more valuable information *relative to its size*. A similar observation can be made when using the PSBL list instead of the CBL list (Figures 3(c) and 3(d)). For the UT/EWI target (Figure 4), the effect is weaker but still present: the intersection sets contain more highly active BadHoods than the irrelevant sets. Noteworthy, this is not true anymore for CAIS/RNP (Figure 5). This is due to the fact that the blacklist of CAIS/RNP is so small that is is outperformed by the larger lists.

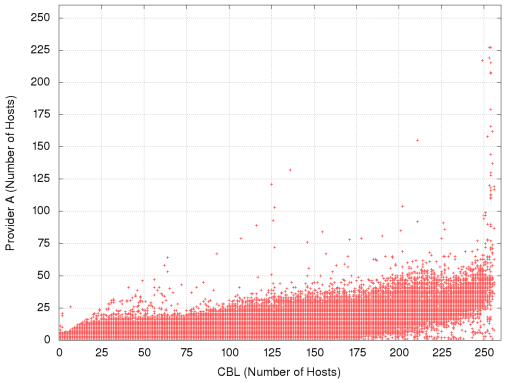
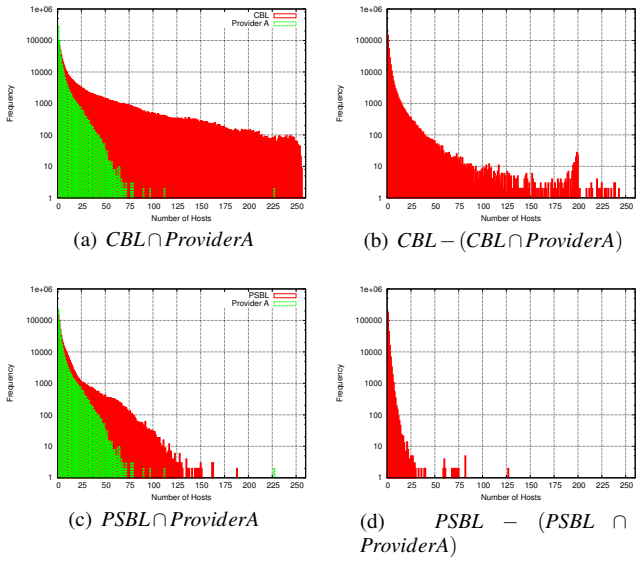


Fig. 2. Activity of BadHoods, as perceived by Provider A and CBL



(a) $CBL \cap ProviderA$ (b) $CBL - (CBL \cap ProviderA)$

(c) $PSBL \cap ProviderA$ (d) $PSBL - (PSBL \cap ProviderA)$

Fig. 3. Histogram of the number of hosts per BadHood for Provider A

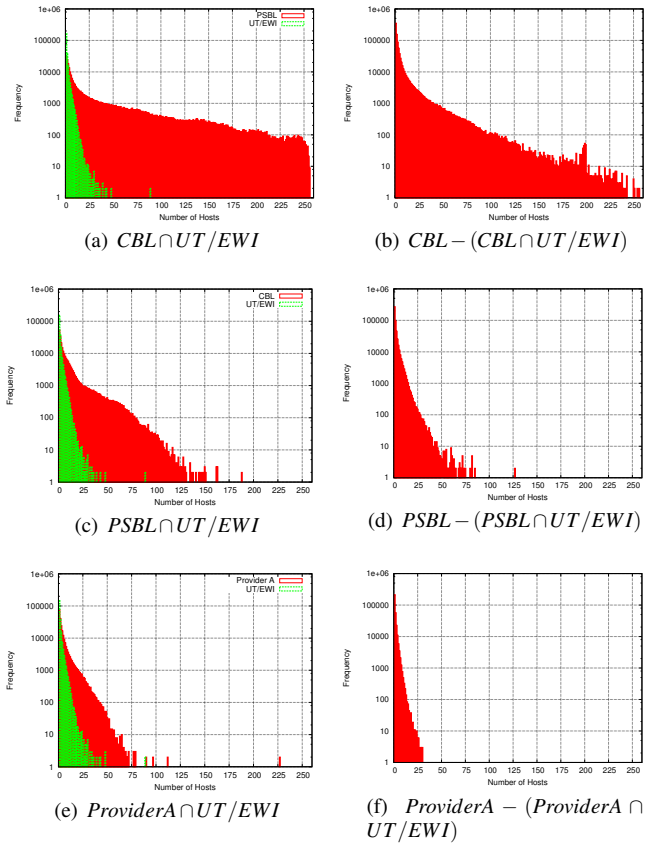
IV. EFFECTIVENESS ON DETECTING SPAM

We have learned in the previous section that, for the majority of the cases (9 out of 11), there is a significant overlap between BadHood blacklists ($> 90\%$ w.r.t the smaller of the two). These results suggest that we can employ BadHood blacklist from various sources to detect Spam. Therefore, in this section, we evaluate the effectiveness of the third-party BadHood blacklists on Spam detection by employing them to protect three distinct targets (mail servers).In [4], the authors have also filtered e-mail employing third-party blacklists. However, they combine several blacklists and only evaluate a small data set ($<$ 5k messages), for a single day and a single source. We have evaluated the third-party blacklists individually, and tested it against several sources, which allowed us to analyze more than 1M messages for several days.

We first describe our methodology and the considered scenario in Section IV-A, followed by a discussion of the achieved results in Section IV-B.

A. Methodology and Considered Scenario

In [4], the authors have presented an approach for Spam filtering based on analyzing the origin of e-mail messages and the URL's within the messages to malicious websites. One of the criteria used in their approach is whether the number of malicious hosts in the origin BadHood of the e-mail is

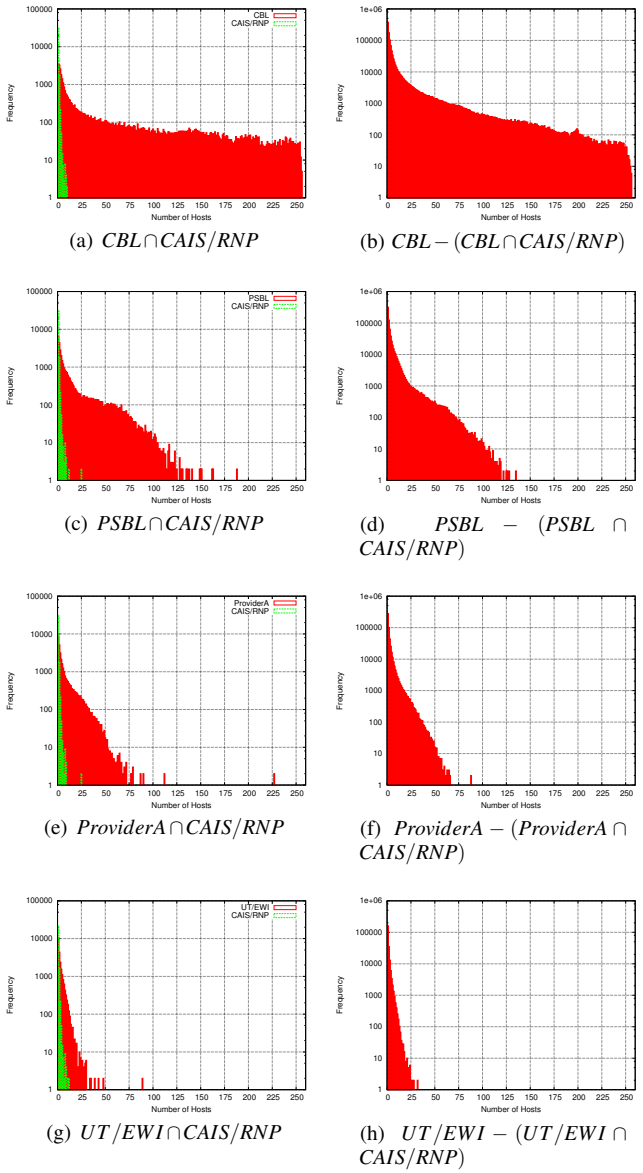


(a) $CBL \cap UT/EWI$ (b) $CBL - (CBL \cap UT/EWI)$

(c) $PSBL \cap UT/EWI$ (d) $PSBL - (PSBL \cap UT/EWI)$

(e) $ProviderA \cap UT/EWI$ (f) $ProviderA - (ProviderA \cap UT/EWI)$

Fig. 4. Histogram of the number of hosts per BadHood for UT/EWI



(a) $CBL \cap CAIS/RNP$ (b) $CBL - (CBL \cap CAIS/RNP)$

(c) $PSBL \cap CAIS/RNP$ (d) $PSBL - (PSBL \cap CAIS/RNP)$

(e) $ProviderA \cap CAIS/RNP$ (f) $ProviderA - (ProviderA \cap CAIS/RNP)$

(g) $UT/EWI \cap CAIS/RNP$ (h) $UT/EWI - (UT/EWI \cap CAIS/RNP)$

Fig. 5. Histogram of the number of hosts per BadHood for CAIS-RNP

above a certain threshold. Similar to our work, the authors used publicly available blacklists to build the list of BadHoods.

To evaluate how the individual BadHood blacklists perform in detecting Spam, we propose for the following experiments a simple Spam detection system that implements the threshold-based criterion described above. It should be emphasized that a real-world BadHood-based mail filter, like the one in [4], should combine different techniques, including whitelisting, in order to optimize the overall detection performance.

Consider L_S as the BadHood blacklist to be used for Spam detection. Whenever a new message M arrives, the mail filter extracts the source /24 prefix address of the sender ($M_{/24}$) and checks it against the list L_S. If $M_{/24}$ is found in L_S, then the mail filter will classify the message as Spam if $nHosts(M_{/24}) > \theta$, where θ ($0 \leq \theta <= 256$[2]) can be seen as a threshold on how malicious a BadHood is. This procedure is summarized in Algorithm 1.

To evaluate the effectiveness of the different BadHood blacklists, we split each data set into a training data set of seven full days (April 19th to April 25th) and a test set of 1 day (April 26th). The resulting five training sets will then be used to build training blacklists as described in Section II-B. Table V shows the number of malicious hosts (distinct /32 hosts), and the number of BadHoods in each training data set.

We then employ the training blacklists to protect the mail servers of Provider A, UT/EWI, and CAIS/RNP from Spam (as in Figure 1). We apply Algorithm 1 to each target T for different values of θ and for the different training blacklists

Algorithm 1 Spam detection procedure used in the experiments

Input: $L_S = \{\langle B_i, nHosts_S(B_i)\rangle, i = 1 \ldots n_S\}$
Input: θ
Input: $M_{/24}$
Output: true, if spam detected; false, otherwise
1: **if** $M_{/24} \in L_S$ **and** $nHosts_S(M_{/24}) > \theta$ **then**
2: **return** true
3: **else**
4: **return** false
5: **end if**

[2]A /24 prefix can have up to 256 malicious IP addresses depending on how addresses are allocated. *E.g.*, if an ISP allocates addresses as /22, as in 130.89.10.0/22 (which covers the /32 addresses 130.89.8.0 — 130.89.11.255), the addresses 130.89.10.255 and 130.89.10.0 are valid "routable" IP addresses.

Dataset	# /32 IPs	# /24 BadHoods
CBL	13,668,909	1,123,492
PSBL	3,301,159	714,466
Provider A	1,498,991	522,522
UT/EWI	377,571	228,445
CAIS/RNP	31,012	28,883

TABLE V. TRAINING DATA SETS, APRIL 19TH–25TH, 2010

Dataset	# /32 IPs	# /24 BadHoods	# Spam
Provider A	296,596	206,980	879,856
UT/EWI	68,748	59,739	221,179
CAIS/RNP	6,447	6,314	13,187

Dataset	# /32 IPs	# /24 Hoods	# Ham
HAM: UT/EWI	1,540	978	7,950

TABLE VI. TEST DATA SETS, APRIL 26TH, 2010

L_S. For each mail in the test set of the target, the algorithm has to decide whether the mail should be flagged as Spam or not. The achieved Spam detection rate $r_{S,T}(\theta)$ is defined as

$$r_{S,T}(\theta) = \frac{\text{number of Spam mails detected}}{\text{total number of Spam mails received by } T}. \quad (4)$$

The total number of Spam mails received by the different targets on the test day are shown in the first three rows of Table VI. The table also gives the number of spammers (/32 IP addresses) and the number of observed BadHoods on that day. For UT/EWI, we also have the number of Ham messages received, as shown in the fourth row of the table.

B. Experimental Results and Discussion

Figures 6(a), 6(b), and 6(c) show the results (in percent) for detecting the Spam directed to Provider A, UT/EWI, and CAIS/RNP, respectively, as function of the threshold θ, using the different training blacklists. Again (see Section III-A), we have only used a training blacklist if it is larger than the target's training blacklist (e.g., we have not applied the UT/EWI list to the Provider A target).

The figures indicate that it is possible to effectively detect Spam messages based on the different BadHood blacklists. This is especially true for large blacklists, like CBL, which always provides the best detection rate. However, and especially for the smaller lists, the figures also show that the rate decreases fast with increasing values of θ, a fact that most likely is due to the presence of high-volume spammers in the data sets.

A second insight provided by these results is that the value of θ should be adjusted to the considered BadHood blacklist. For the same θ, the detection rate changes considerably among BadHood blacklists. At first sight, this seems to suggest that the best choice for an administrator is the largest BadHood blacklist, just due to the fact that it has observed a higher number of spamming hosts. However, large BadHood blacklists might suffer of drawbacks like a high number of irrelevant entries, as indicated in Section III-B.

Considering this fact, we then investigate if smaller BadHoods blacklists can still potentially provide similar detection rates if the threshold θ is chosen appropriately. Let θ_{CBL} be the threshold that we have used to calculate the detection rate for the CBL training blacklist (the biggest in our data set).

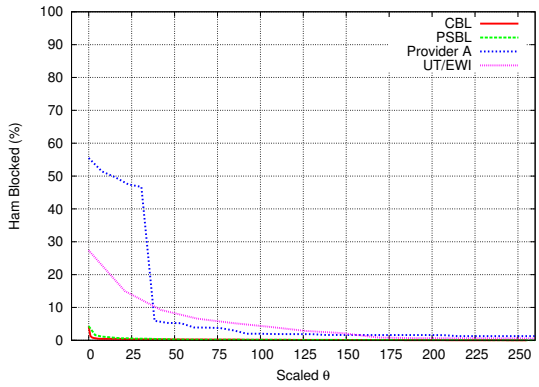


Fig. 7. Percentage of Ham erroneously blocked at UT/EWI – scaled θ

We choose the threshold θ_S for the training list L_S as

$$\theta_S = \theta_{CBL} \cdot \frac{\sum_{B_i \in I_{CBL \cap S}} nHosts_{CBL}(B_i)}{\sum_{B_i \in I_{CBL \cap S}} nHosts_S(B_i)} \quad (5)$$

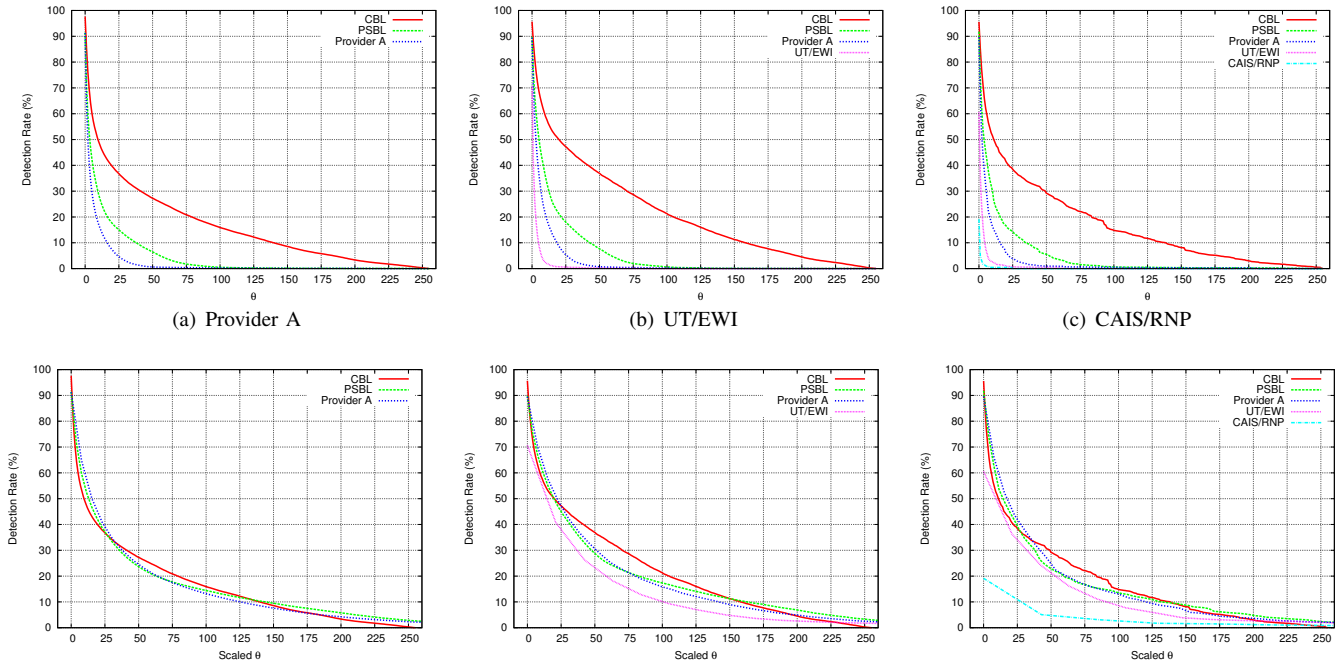
For example, instead of comparing the detection rate between the CBL and PSBL lists for $\theta = 100$ to a particular target T, we then can use $\theta_{CBL} = 100$ for the CBL list and $\theta_{PSBL} = \frac{100}{4.14} = 22.6$ for the PSBL list because the latter contains 4.14 times less IP addresses than the CBL list. In this way, we compensate the fact that PSBL has, in average, observed less malicious hosts than CBL.

In Figures 6(d), 6(e), and 6(f), we present the detection rates obtained for the different targets using the rescaled θ defined in Eq. (5). For the CBL list, the threshold as indicated on the x-axis is used. For the other lists, we compute the rescaled theta according to Eq. (5).

The figures show that, once the size factor is removed by using Eq. (5), all the considered BadHood blacklists detect Spam with comparable performance for a wide range of θ, although the CBL list still provides the best results when the whole lists ($\theta = 0$) are used. The only exception is the CAIS/RNP list which only achieves detection rates below 20%, due to its small size (see Figure 6(f)). These results therefore indicate that Eq. (5) offers an operational way for choosing values of θ for different blacklists such that the blacklists are similarly effective in identifying Spam. In fact, one may be tempted to conclude from these results that all blacklists perform similarly independently of their size.

However, a different picture is obtained when calculating the number of legitimate mail traffic erroneously flagged as Spam – that is, the number of false positives. Figure 7 shows the percentage of legitimate mail messages received by the mail server of UT/EWI that are labeled as Spam for varying values of the scaled threshold θ. While for CBL and PSBL the percentages of blocked Ham is less than 5% and rapidly falls to zero, for UT/EWI and Provider A we observe that up to approximately 60% of legitimate mail would be labeled as Spam if a very low value of θ is chosen. On the other hand, also in the case of Provider A and UT/EWI, the percentage of blocked Ham is decreasing rapidly for increasing values of θ.

Our results highlight therefore a trade-off between (i) the size of the blacklist, (ii) the Spam detection rate and (iii) the



(a) Provider A (b) UT/EWI (c) CAIS/RNP

(d) Provider A – scaled (e) UT/EWI – scaled (f) CAIS/RNP – scaled

Fig. 6. Spam detection rate for varying values of the threshold θ in Fig. (a)-(c) and the scaled threshold in Fig.(d)-(f)

percentage of blocked Ham. Very large lists, such as CBL and PSBL, achieve a high Spam detection rate with a low percentage of blocked Ham but contain a large number of irrelevant entries. In contrast, small and mid-sized lists, that is, Provider A and UT/EWI, contain much less irrelevant entries and can achieve detection rates comparable to those of the larger lists. However, for $\theta < 100$, a relatively high number of false positives can be expected.

V. Related Work

We have seen in the literature several research works focusing on the concentration of malicious hosts on the Internet. Ramachandran and Feamster have investigated the network level behavior of spammers in [1], by showing that most of spam comes from a few concentrated part of IP address space (IPv4). Collins *et. al* [2], on the other hand, have defined the concept of uncleanliness, which "works as an indicator for how likely the network is to contain compromised hosts". DNS blacklists [10], such as [9], [11], [18], also suggest the same concentration.

Taking these observation in consideration, the Bad Neighborhood concept was introduced in [4]. In that study, the authors have developed a mail filter that employed Spam BadHoods to tell whether a message is Spam or not. They have proposed a combination of several rules to classify a message based on characteristics of the message itself, among which the most closely related to our work are (i) the number of different BadHood blacklist containing the IP address and (ii) the number of Hosts for the /24 subnet. In [4], the authors have used as threshold for the number of hosts in a /24 block values equal to 2, 12, and 24. In our work, we have investigated in Section IV how the detection rate varies accordingly to the number of hosts (θ), for the whole interval ($0 \leq \theta < 256$). In addition, we have proposed and evaluated a scaled value for θ, which considers the size of the BadHood blacklists.

In [19], Soldo *et. al* have employed a recommendation system to predict /24 prefixes that were likely to attack "neighboring" targets (or victim networks). The authors have evaluated the D-Shield data set [20], a community-shared firewall log system, and employed a neighborhood model (popular approach in recommendation systems) to predict attack sources by "trusting similar peers" [19]. Our BadHood definition is not a recommendation system technique and it is defined in terms of neighboring sources of attacks – and not on neighboring targets. Moreover, the authors do not differentiated BadHoods according to the application used in the attack, while we consider application-tailored BadHoods.

Taking the previous studies into account, we have investigated in [7] the specifics of Spamming BadHoods. We have proposed four definitions for Spamming BadHoods, each of them addressing a particular part of the "Spam picture". We have found that botnets (and individual bots) are responsible for most of Spam; however we cannot neglect the impact of massive Spamming BadHoods – that is, BadHoods that have been observed having very few spamming hosts which have sent a large number of spam messages.

In [8], we have proposed and evaluated two IP-based techniques to aggregate malicious hosts into network prefixes other than /24 (from /24 to /8). We have found that BadHood can be viably aggregated into smaller prefixes; however, the smaller the prefix, the larger the aggregation error. This result allowed us to provide meaningful feedback to network administrators wishing to make use of BadHood-based filtering.

VI. Conclusions

In this paper we have evaluated the effectiveness of third-party BadHood blacklists to protect a target system from Spam. In particular, we have investigated if and to what extend a BadHood blacklist is specific to its source and whether a network administrator can interchangeably use different third-party BadHood blacklists to detect Spam messages. We have conducted our analysis using five BadHood blacklists generated from third-party sources.

We observe a significant overlap between the third-party BadHood blacklists. This is particularly true for large, public blacklists such as CBL and PSBL, which cover up to 99% of the BadHoods spamming the targets. Our research also shows that this intersection captures the most aggressive BadHoods (in terms of number of malicious hosts). However, we also found that large blacklists also contain a large number of "irrelevant" entries, which are BadHoods not observed by the target itself. Such extra entries might impose a burden if used in resource-restricted security mechanisms, such as firewalls.

Based on the high overlapping observed, we have therefore applied BadHood blacklists to Spam detection. For this, we have evaluated, for each individual blacklist, the performance of a simple BadHood-based mail filter that takes into account the number of malicious hosts in the BadHood from which a message originates. We have found that the largest lists provide the best detection rates. However, the results from smaller lists can be significantly improved if the detection threshold is properly chosen. We therefore provide an operational way to adjust the detection threshold according to the BadHood blacklist size. However, we also show that the choice of the detection threshold might cause a high percentage of misclassified legitimate mail messages (false positives), especially for smaller blacklists. Therefore, when deploying a BadHood-enhanced Spam filter, a network administrator should be aware of the trade-off among the BadHood blacklist size, the Spam detection rate, and the percentage of blocked legitimate messages in the case of medium to small BadHood blacklists.

As future work, we will investigate if the attacks to distinct applications are carried by a same set of Internet BadHoods. In addition, we will analyze how Internet BadHoods change over time to be able to predict attacks from unobserved BadHoods based on its neighbors behavior.

Acknowledgments: The authors would like to thank CAIS/RNP, Frederico Costa, Jürgen Rochol, Liliana Solha, Lisandro Granville, Marc Berenschot, Provider A, and UT/EWI for their support for this research. Special thanks to the maintainers of CBL and PSBL blacklists.

References

[1] A. Ramachandran and N. Feamster, "Understanding the Network-level Behavior of Spammers," in *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '06, New York, NY, USA, 2006.

[2] M. P. Collins, T. J. Shimeall, S. Faber, J. Janies, R. Weaver, M. De Shon, and J. Kadane, "Using Uncleanliness to Predict Future Botnet Addresses," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 93–104.

[3] M. A. Rajab, F. Monrose, and A. Terzis, "On the effectiveness of distributed worm monitoring," in *Proceedings of the 14th conference on USENIX Security Symposium - Volume 14*. Berkeley, CA, USA: USENIX Association, 2005, pp. 15–15.

[4] W. van Wanrooij and A. Pras, "Filtering Spam from Bad Neighborhoods," *International Journal of Network Management*, vol. 20, no. 6, pp. 433–444, November 2010.

[5] V. Fuller and T. Li, "RFC 4632: Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan," August 2006.

[6] G. C. M. Moura, "Internet Bad Neighborhoods (to appear)," Ph.D. dissertation, University of Twente, 2013.

[7] G. C. M. Moura, R. Sadre, and A. Pras, "Internet Bad Neighborhoods: The Spam Case," in *7th International Conference on Network and Services Management (CNSM 2011)*, Paris, France), 2011.

[8] G. C. M. Moura, R. Sadre, A. Sperotto, and A. Pras, "Internet Bad Neighborhoods Aggregation," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*, April 2012, pp. 343 –350.

[9] Passive Spam Block List, 2011. [Online]. Available: http://psbl.surriel.com/

[10] J. Levine, "DNS Blacklists and Whitelists," RFC 5782 (Informational), Internet Engineering Task Force, Feb. 2010.

[11] CBL, "Composite Blocking List," 2012. [Online]. Available: http://cbl.abuseat.org/

[12] J. P. John, A. Moshchuk, S. D. Gribble, and A. Krishnamurthy, "Studying spamming botnets using botlab," in *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, ser. NSDI'09. Berkeley, CA, USA: USENIX Association, 2009.

[13] V. Chandra and N. Shrivastava, "Ways to evade spam filters and machine learning as a potential solution," in *Communications and Information Technologies, 2006. ISCIT '06. International Symposium on*, 18 2006-sept. 20 2006, pp. 268 –273.

[14] J. Franklin, V. Paxson, A. Perrig, and S. Savage, "An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants," in *Proceedings of the 14th ACM conference on Computer and communications security*, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 375–388.

[15] UT/EWI, "Faculty of Electrical Engineering, Mathematics and Computer Science," 2012. [Online]. Available: http://www.utwente.nl/en/education/eemcs/

[16] CAIS, "Security Incident Response Team (In Portuguese: *Centro de Atendimento a Incidentes de Segurança*," 2012. [Online]. Available: http://www.rnp.br/en/cais/

[17] RIPE NCC, "PI Assignment Size ," July 2006. [Online]. Available: http://www.ripe.net/ripe/policies/proposals/2006-05

[18] The Spamhaus Block List, 2011. [Online]. Available: http://www.spamhaus.org/sbl/

[19] F. Soldo, A. Le, and A. Markopoulou, "Predictive blacklisting as an implicit recommendation system," in *Proceedings of the 29th conference on Information communications*, ser. INFOCOM'10. Piscataway, NJ, USA: IEEE Press, 2010, pp. 1640–1648.

[20] DSHIELD.org, "dshield Home — DShield; Cooperative Network Security Community - Internet Security," May 2012. [Online]. Available: http:/www.dshield.org