Characterization of Blacklists and Tainted Network Traffic

Jing Zhang[1], Ari Chivukula[1], Michael Bailey[1], Manish Karir[2], and Mingyan Liu[1]

[1] University of Michigan
Ann Arbor, Michigan, USA
[2] Cyber Security Division, Department of Homeland Security,
Washington DC, USA

Abstract. Threats to the security and availability of the network have contributed to the use of Real-time Blackhole Lists (RBLs) as an attractive method for implementing dynamic filtering and blocking. While RBLs have received considerable study, little is known about the impact of these lists in practice. In this paper, we use nine different RBLs from three different categories to perform the evaluation of RBL tainted traffic at a large regional Internet Service Provider.

1 Introduction

A variety of threats, ranging from misconfiguration and mismanagement to botnets, worms, SPAM, and denial of service attacks, threaten the security and availability of today's Internet. In response, network operators have sought to adopt security policies that minimize their impact. Real-time Blackhole Lists (RBLs) are a form of coarse-grained, reputation-based, dynamic policy enforcement in which real-time feeds of malicious hosts are sent to networks so that connections to these hosts may be rejected.

Existing work has studied how these lists can be created [14], evaluated their effectiveness [17, 23], and explored the properties of the networks that make them effective [24, 26, 22]. In this paper, rather than focusing solely on the lists themselves, we analyze the *impact* of nine popular blacklists on Merit Network [8], a large Internet Service Provider (ISP). By examining what network traffic is tainted by these blacklists, we gain better insight into the utility of these mechanisms and the nature of malicious traffic on our networks. Our findings include:

- While stable in size, the RBL populations are highly dynamic, growing between 150% to 500% over a one week period.
- Classes of RBLs show significant internal entry overlap, but little similarity is seen between classes.
- RBL classes share affinity for specific geographic distributions (e.g., RIPE and APNIC dominate SPAM; ARIN and RIPE dominate phishing and malware).

- A surprisingly high proportion, about 17%, of the collected network traffic is tainted by at least one of the nine RBLs.
- Our network only saw traffic to a small portion, between 3% and 51%, of IP addresses within the blacklists.
- Heavy hitters account for a significant number of the tainted bytes to the network.

2 Data Collection Methodology

Netflow We collected records of the traffic at Merit to understand the impact of RBLs. Merit is a large regional ISP, which provides high-performance computer networking and related services to educational, government, healthcare, and non-profitable organizations located primarily in Michigan. This network experiences a load which varies daily from a low of four Gbps to a high of eight Gbps. Though Merit has over 100 customers, the top five make up more than half of the total traffic, and HTTP accounts for more than half of the traffic volume.

Our traffic data was collected via NetFlow [7] with a sampling ratio of 1:1. The traffic was monitored at all peering edges of the network for a period of one week, starting on June 20, 2012. During this period, we experienced several collection failures, each lasting from one to seven hours, for a total of 17 hours lost. The collected NetFlow represents 118.4TB of traffic with 5.7 billion flows and 175 billion packets.

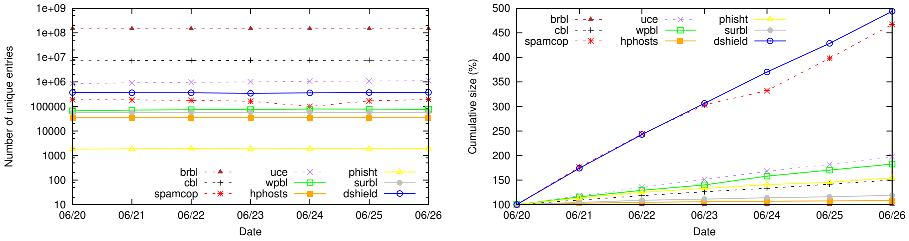
RBL Type	**RBL Name**
SPAM	CBL[3], BRBL[2], SpamCop[16], WPBL[13], UCEPROTECT[12]
Phishing/Malware	SURBL[11], PhishTank[9], hpHosts[5]
Active attack/probing behavior	Dshield[4]

Table 1: Reputation data sources and types.

Reputation Black Lists RBLs are lists managed by various organizations that contain IP addresses believed to have originated some malicious behavior. RBLs generally focus on some specific suspicious behavior. Merit collects nine commonly used RBLs on a daily basis, which are typically fetched directly from the publisher via rsync or wget. These lists can be categorized into three types: SPAM, Phishing/Malware, or Active (and prolific) malicious activity (as shown in Table 1).

3 Properties of Reputation Blacklists

Timing We examined the stability of each RBL with respect to *the daily number of unique IP addresses.* As shown in Figure 1a, the size varied across RBLs with BRBL being much larger than the others, but the size of RBLs was consistent over the week measured. In order to understand the churn of unique IP addresses, we calculated the relative size of cumulative entries in Figure 1b. Spamcop and Dshield updated their entries aggressively, with nearly 500% turnover in one week, while BRBL, hpHosts, and SURBL were relatively static during the week, with less than 110% turnover.



(a) Number of unique entries.

(b) Relative cumulative size (%).

Fig. 1: Daily size and cumulative size of RBLs.

	Spam					Phishing/Malware			Active
	BRBL	CBL	Spamcop	UCE	WPBL	hpHosts	Phisht	SURBL	Dshield
AFRINIC	3.02	7.70	5.89	6.37	4.19	0.20	0.58	0.04	2.19
APNIC	25.20	47.14	51.94	48.45	51.27	8.45	11.56	5.58	36.19
ARIN	6.23	1.05	2.53	1.84	6.17	53.32	43.93	54.70	13.54
LACNIC	17.11	16.19	12.15	15.89	10.59	1.66	5.32	1.44	8.54
RIPENCC	48.44	27.93	27.50	27.44	27.77	36.37	38.6	38.24	39.53

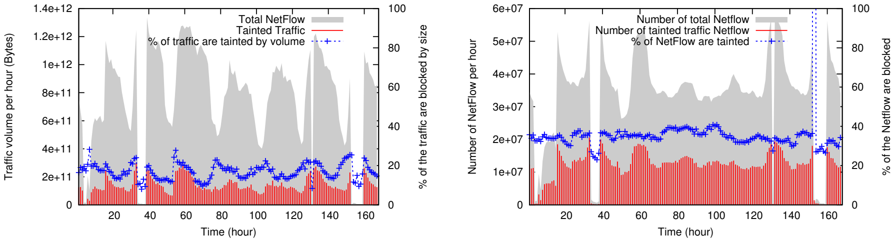
Table 2: Geographic distribution of IPs for each RBL (%).

Regional Characteristics We mapped the blacklisted IP addresses to their registries by using the IP to ASN mapping services provided by Team Cymru [21]. Table 2 demonstrates that a given class of RBLs has consistent geographical properties. SPAM- and Active-attack-related lists have more entries in the APNIC (Asia/Pacific) and RIPENCC (Europe) regions, while ARIN (North America) and RIPENCC are the most common regions in Phishing/Malware RBLs. Even though monitoring position and listing methodologies are different for each RBL, they share consistent views of the regional distribution of malicious activity.

	Spam					**Phishing/Malware**			**Active**
	BRBL	**CBL**	**Spamcop**	**UCE**	**WPBL**	**hpHosts**	**Phisht**	**SURBL**	**Dshield**
BRBL	100.0	75.2	94.6	89.8	93.8	5.3	10.0	30.7	33.2
CBL	3.9	100.0	98.1	91.7	70.2	0.5	0.7	6.2	9.3
Spamcop	0.1	2.3	100.0	12.6	21.5	0.1	0.1	0.8	1.2
UCE	0.6	12.1	69.4	100.0	50.6	0.3	1.5	1.2	4.8
WPBL	0.0	0.7	8.8	3.7	100.0	0.0	0.2	0.9	0.4
hpHosts	0.0	0.0	0.0	0.0	0.0	100.0	33.7	7.3	0.0
Phisht	0.0	0.0	0.0	0.0	0.0	1.8	100.0	1.7	0.0
SURBL	0.0	0.0	0.3	0.1	0.7	11.8	52.8	100.0	0.1
Dshield	0.1	0.4	2.4	1.8	2.2	0.4	0.7	0.3	100.0

Table 3: The average % (of column) overlap between RBLs (row, column).

Overlap We examined to what extent RBLs overlap with other; we expected that overlap within the same category of RBLs would be significantly larger than the overlap among different classes. Our results in Table 3 match our expectation: BRBL and CBL, the two largest SPAM blacklists, cover about 90% of other SPAM-related lists, and the intersection within hpHosts, PhishTank, and SURBL is also large. Meanwhile, the overlaps between different classes are trivial.



(a) By traffic volume (bytes).

(b) By number of flows.

Fig. 2: Total traffic v.s. tainted traffic.

4 Impact of Reputation

One of the key questions we considered in our study was, *what fraction of traffic carries a negative reputation?* In our study, if one or both of the collected NetFlow's source and destination IPs are listed by any RBL, the NetFlow is considered tainted. As some IP addresses are shared via mechanisms like Network Address Translation (NAT), some traffic was tainted due to "guilt by association", but we did not explore this effect. While we expected that perhaps as much as 10% of network traffic might be potentially malicious [6], we found that tainted traffic accounted for an average of 16.9% of the total traffic volume over the week. When measured by flow count, the proportion is even larger, with 39.9% of the flows being tainted (Figure 2b).

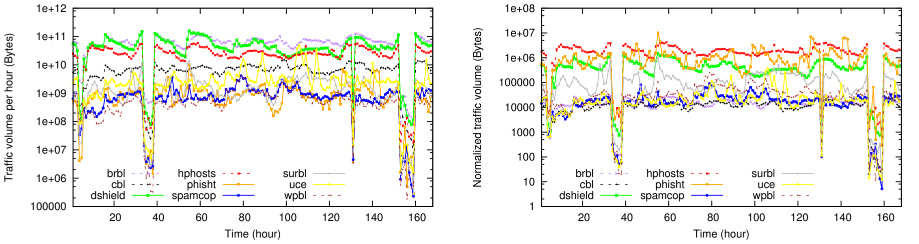
	Spam					**Phishing/Malware**			**Active**
	BRBL	CBL	Spamcop	UCE	WPBL	hpHosts	Phisht	SURBL	Dshield
Touched entries	4,142,394	577,583	44,383	134,024	16,288	13,989	983	14,043	105,918
% of the list	2.8%	7.7%	29.3%	39.5%	51.2%	25.2%	24.4%	13.9%	22.1%

Table 4: RBL entries touched by our network traffic.

Next, we investigated *the potential impact of global reputation blacklists when applied locally.* Prior work in this area has suggested that there might be some entries in global blacklists that are never used by an organization [26], and our results validated this argument. In Table 4, we show the average number of daily entries touched for each RBL. Only a small fraction of entries were touched by our network traffic. For our ISP, only small portions of RBLs are relevant, even though these portions may change over time.

Finally, we examined *whether lists, or a class of lists, have the greatest impact on our traffic.* The traffic volume tainted by each RBL is shown in Figure 3a. There is a clear variance among tainted traffic volumes, ranging from more than ten GB per hour by Dshield, BRBL, and hpHosts to about tens of MB per hour by Spamcop, PhishTank, and SURBL.

Since the number of entries in each RBL differs, we then normalized the volume of tainted traffic (i.e. $\frac{Volume\ of\ tainted\ traffic\ by\ the\ RBL}{Number\ of\ touched\ entries\ in\ the\ RBL}$) in Figure 3b. Interestingly, we show that each entry in hpHosts, PhishTank, and Dshield taints about one MB of traffic on average; but, the contribution of entries in the SPAM-related RBLs is about two orders of magnitude lower.



(a) Total tainted traffic.

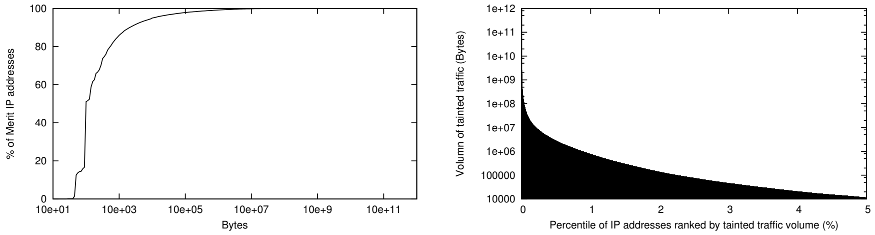
(b) Normalized tainted traffic volume.

Fig. 3: Tainted traffic per RBL.

5 Impact of Heavy Hitting IPs

In this section, we investigate whether any specific IPs are responsible for skewing the traffic distribution. Toward this end, we divided the traffic into two categories: those IP addresses belonging to Merit (internal IP addresses) and those not belonging to Merit (external IP addresses).

5.1 External IP Addresses



(a) CDF of traffic volume per IP.

(b) Tainted traffic volume of top 5% of IPs.

Fig. 4: Tainted traffic to/from external IP addresses.

Of the 11,016,520 unique external IP addresses in the tainted traffic, 99.5% of them had less than 10 MB of tainted traffic each (as shown in Figure 4a). However, the top contributors had more than 100 GB of tainted traffic associated with each of them (Figure 4b). In fact, the top 50 external IP addresses contributed about 40% of total tainted traffic. In the following analysis, we try to define *what these hitters are* and *what comprises their traffic.*

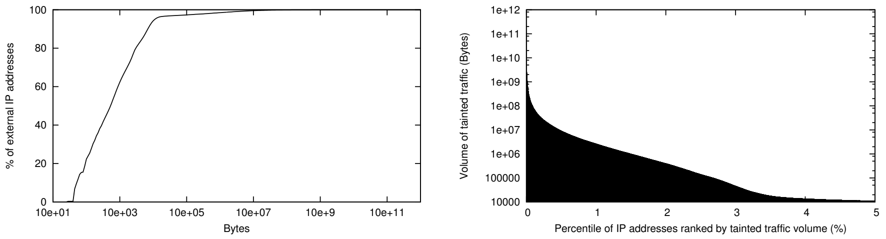
External Heavy Hitters Among the top 50 external IP addresses, 39 are listed in at least one RBL. It is surprising to see that 27 of those are hosting service providers or caching servers, including Amazon Web Services hosts (10 IPs listed on hpHosts, Phisht, SURBL, or Dshield), Facebook content distribution network (CDN) servers (six IPs listed on Dshield), Pandora media servers (six IPs listed on Dshield), EDGECAST Network hosts (three IPs listed on hpHosts, Phisht, or Dshield), and BOXNET servers (two IPs listed on BRBL). These hosts are owned by popular service providers and their traffic is dominated by HTTP, as shown in Table 5.

Ports	*80*	*443*	*1935*	*1256*	*1509*	*1046*	*1077*	*1224*	*1121*	*1065*
% of volume	60.65	35.31	3.48	1.12	1.06	1.03	0.71	0.66	0.64	0.58

Table 5: Distribution over TCP/UDP ports for top blacklisted external IPs.

Top External Hitters Not Blacklisted The remaining 11 external IP addresses in the top 50 are IP addresses communicating with tainted Merit hosts, who send large volumes of traffic. Of these external destinations, 10 are owned by Netflix and one belongs to Yahoo!. 99% of the tainted traffic within these 11 IP addresses was over HTTP.

5.2 Internal IP Addresses



(a) CDF of traffic volume per IP. (b) Tainted traffic volume of top 5% IPs.

Fig. 5: Tainted traffic to/from internal IP addresses.

Analysis of the 2,515,080 Internal IP addresses observed in the tainted traffic also showed the existence of heavy internal hitters (as shown in Figure 5). In this case, the top 50 internal IP addresses contributed 38% of the total tainted traffic.

Organization	**CDN**	**EDU**				**LIB**	**MED**
	Akamai	**University**	**College**	**Intermediate**	**Regional**		
Num of IPs	9	6	4	1	1	4	4
Total	9	12				4	4

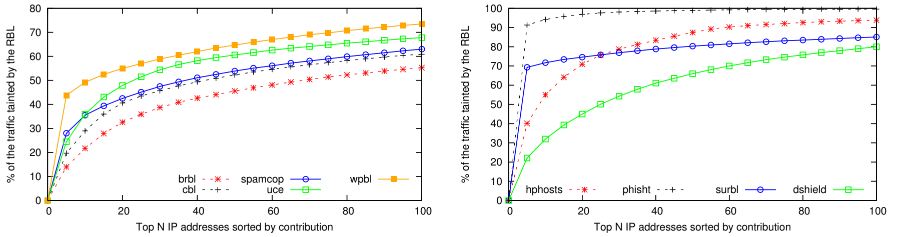
Table 6: Organization of blacklisted internal IP addresses.

Internal Heavy Top Hitters Our results showed that there are only 35 IP addresses in the top 50 listed by the RBLs, and of the 35 IP addresses, only 29 were resolvable to host names. When categorized by owner (as shown in Table 6), we see that nine of these blacklisted IP addresses are owned by Akamai [1], a provider of content delivery network (CDN) and shared hosting services; others are hosts registered by educational institutions, library network providers, and medical centers. Interestingly, there are two Virtual Private Network servers, a mail server, and one web site server from educational institutions.

Top Internal IP Addresses not on a RBL We found the top three internal heavy hitters, which accounted for 12% of total tainted traffic, are not themselves on an RBL, and 81.6% of their traffic is HTTPS traffic. Furthermore, by inspecting the blacklisted hosts they communicated with, we noticed that about 80% of their tainted traffic is to/from Amazon Web Services (AWS) IP addresses that are blacklisted.

5.3 Top Hitters Distribution

Heavy hitters constitute a significant portion of tainted traffic. *Are these top hitters distributed across RBLs?*



(a) SPAM RBLs. (b) Phishing/Malware and Active RBLs.

Fig. 6: Cumulative contributions of the top N entries per RBL.

To understand the heavy hitters in each RBL, we defined the contribution of $entry_i$ in RBL_j as $\frac{V_{entry_i}}{V_{RBL_j}}$, where V_{entry_i} is the volume of traffic tainted by $entry_i$ and V_{RBL_j} is the total volume of traffic tainted by RBL_j. We then sorted the entries by their contribution in decreasing order for each RBL, and then derived the cumulative contribution of the top N entries (Figure 6). The top entries contribute greatly to the RBLs — the traffic tainted by the top 50 entries accounted for more than half of the total tainted traffic of each. In the case of Phishing/Malware RBLs, the top 50 entries contributed even more (80%) of the tainted traffic (as shown in Figure 6b). Once again, we find a small amount of entries dominating the tainted traffic.

BRBL	**CBL**	**Spamcop**	**UCE**	**WPBL**
80 (59.62)	*80* (34.01)	*80* (26.394)	*3389* (27.03)	*25* (26.71)
443 (22.30)	*443* (21.26)	*44794* (16.51)	*53* (14.16)	*80* (23.30)
1935 (2.22)	*4444* (11.78)	*4025* (16.16)	*25345* (12.80)	*44794* (19.30)
3578 (1.26)	*25* (6.67)	*25* (11.14)	*80* (12.54)	*4025* (18.89)
17391 (1.21)	*3389* (4.96)	*37101* (7.60)	*25* (8.18)	*1080* (9.73)

(a) SPAM.

hpHosts	**Phisht**	**SURBL**
80 (84.99)	*80* (65.05)	*443* (52.30)
443 (15.00)	*443* (32.32)	*80* (44.84)
1256 (1.95)	*49729* (2.96)	*25* (1.85)
1121 (1.10)	*42652* (1.80)	*1288* (1.51)
1605 (1.01)	*52951* (1.48)	*1032* (1.12)

(b) Phishing/Malware.

Dshield
80 (60.75)
443 (32.26)
1935 (3.55)
993 (1.68)
1509 (1.16)

(c) Active.

Table 7: Top TCP/UDP ports for traffic tainted by top 50 contributors per RBL.

Next, we characterized the tainted traffic by the top 50 contributors for each RBL (Table 7). Though not dominating, SMTP (port 25) traffic occupied a large proportion of the tainted traffic for each of the SPAM related blacklists (except

BRBL). This matches our expectation that SPAM related IP addresses send email more aggressively than other hosts. In the other RBLs, we see a higher proportion of Web related traffic. This could be associated with either Phishing and Malware distribution activities or other, potentially benign, traffic from these hosts.

	Spam					**Phishing/Malware**			**Active**
	BRBL	CBL	Spamcop	UCE	WPBL	hpHosts	Phisht	SURBL	Dshield
CDN	2	0	0	0	0	35	3	1	26
HOST	0	0	1	0	2	3	19	17	12
TOR	1	11	0	0	0	1	0	0	0
MAIL	0	0	0	3	5	0	1	0	1
VPN	3	0	0	1	0	0	0	0	0
Total	**10**	**13**	**1**	**4**	**7**	**39**	**23**	**18**	**39**

Table 8: Service hosts in top 50 contributors for each RBL.

Finally, we looked at the network and domain information of the top contributers (shown in Table 8). We found that 60 of these IP addresses are used by content delivery networks and 51 of them are owned by hosting companies. Four VPN servers are listed in BRBL and UCEProtector, while 11 Tor nodes are shown in CBL. Nine different mail servers (some of them belonging to LinkedIn) are also in the top 50 entries of some RBLs. These entries form a sizable fraction of network traffic. This holds especially true for the Phishing/Malware and Active RBLs, whose tainted traffic included from 29% to 68% of these heavy hitters.

6 Related Work

While there is a great deal of prior work on generating reputation blacklists [15, 20, 24, 26], there are fewer studies which characterize the RBLs themselves or their impact. Prior work has focused on understanding the makeup of RBLs from geographical and topological perspectives [18], as well as the correlation between seven popular RBLs [17]. Other related work has discussed the effectiveness and limitation of blacklists. For example, researchers have shown that blacklists often contain numerous false positives [23] and outdated entries [22]. The study in [19] finds that very few sections of IP space account for the majority of SPAM (meaning that a small, stable RBL would be highly effective at blocking SPAM), and that a small, but increasing, amount of SPAM comes from random and short-lived hijacked prefixes (whose entries in RBLs would quickly become outdated). In [26], the author argues that entries in common blacklists which are never used within an organization should be removed to reduce costs. Our work is complementary to these efforts, as our focus in this study is to gain a better understanding of the key properties of RBLs themselves and their impact on traffic from the perspective of an ISP.

7 Limitations

In our study, we adopted a liberal approach to tainted traffic analysis: tainting all the traffic of a host by all the entries in all the blacklists. As a result, our estimate of 17% of the tainted traffic provides an upper bound estimation. We conjecture that there are two main sources of overestimation: some RBLs are intended to taint only one kind of application traffic instead of an entire host, and the RBLs may contain false positives. While a detailed exploration is beyond the scope of this work, we have briefly examined these two problems in an effort to provide a lower bound estimation of tainted traffic. We looked at the reduction in tainted traffic when lists are applied solely to the type of traffic they pertain to (e.g., SPAM blacklists are only applied to SMTP traffic). The results show that 10.5% of total traffic was tainted by our more conservative approach. In addition, we recognized that there may be false positives, for example, Amazon Web Services and Facebook CDNs, in the RBLs. Although some previous work has shown that the cloud services have been used for malicious activities [25], and are not clearly false positives, to be conservative we whitelisted these service providers. As a result, the volume of tainted traffic was reduced to 7.5% of total traffic. Therefore, we believe a realistic value for tainted traffic is likely to lie within the range of 7.5% to 17% of the total traffic by bytes.

8 Conclusion

In this study, we characterized nine RBLs and their impacts on traffic from a live operational network. The RBLs are highly dynamic, growing between 150% to 500% over a period of one week. While there is a significant overlap among RBLs within the same class, little similarity is seen between classes. We demonstrated that almost 17% of the traffic could be considered tainted, as it flowed to or from addresses on various RBLs. We also show the relative contribution of different entries on a RBL towards this tainted traffic, and we show that heavy hitters dominate both tainted traffic as well as RBLs.

Reputation information is a useful resource for organizations to evaluate and design their security policies. Our work indicates that an organizational view of network threats can differ from the global perspective. Therefore, it is important to consider local information in conjunction with global RBLs in order to build more accurate reputation information.

Dataset Availability Our RBLs are provided under a licensing agreement that requires they not be publicized or redistributed. NetFlow data represents potentially private information about users of our network. Therefore, we are unable to provide the raw data of RBLs and traffic NetFlow used in our work. An anonymized version of NetFlow traffic annotated with the RBL matches is available through the PREDICT project [10].

Acknowledgments This work was supported in part by the Department of Homeland Security (DHS) under contract number D08PC75388; the National Science Foundation (NSF) under contract numbers CNS 1111699, CNS 091639, CNS 08311174, and CNS 0751116; and the Department of the Navy under contract number N000.14-09-1-1042.

References

1. Akamai. `www.akamai.com/`.
2. Barracuda reputation blocklist. `http://www.barracudacentral.org/`.
3. Cbl: Composite blocking list. `http://cbl.abuseat.org/`.
4. Dshield. `http://www.dshield.org/`.
5. HpHosts for your pretection. `http://hosts-file.net/`.
6. Internet has a garbage problem, researcher says. `http://www.pcworld.com/article/144006/article.html`.
7. Introduction to Cisco IOS NetFlow. `http://www.cisco.com/en/US/products/ps6601/prod_white_papers_list.html`.
8. Merit Network INC. `http://www.merit.edu/`.
9. Phishtank. `http://www.phishtank.com/`.
10. PREDICT: Protected Repository for the Defense of Infrastructure Against Cyber Threats. `https://www.predict.org/`.
11. SURBL: URL Reputation Data. `http://www.surbl.org/`.
12. Uceprotector network. `http://www.uceprotect.net/`.
13. Wpbl: Weighted private block list. `http://www.wpbl.info/`.
14. Manos Antonakakis, Roberto Perdisci, David Dagon, Wenke Lee, and Nick Feamster. Building a Dynamic Reputation System for DNS. In *USENIX Security Symposium*, pages 273–290, 2010.
15. Holly Esquivel, Aditya Akella, and Tatsuya Mori. On the effectiveness of IP reputation for spam filtering. In *Proceedings of COMSNETS '10*, pages 1–10, 2010.
16. Cisco Systems Inc. SpamCop Blocking List (SCBL). `http://www.spamcop.net/`.
17. Jaeyeon Jung and Emil Sit. An empirical study of spam traffic and the use of DNS black lists. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 370–375, New York, NY, USA, 2004. ACM.
18. Kyle Creyts Manish Karir and Nathan Mentley. Towards network reputation - analyzing the makeup of rbls, June 2011.
19. Anirudh Ramachandran and Nick Feamster. Understanding the network-level behavior of spammers. In *Proceedings of SIGCOMM '06*, pages 291–302, 2006.
20. Anirudh Ramachandran, Nick Feamster, and Santosh Vempala. Filtering spam with behavioral blacklisting. In *Proceedings of the 14th ACM conference on Computer and communications security*, 2007.
21. Team Cymru Community Services. IP to ASN Mapping. `http://www.team-cymru.org/Services/ip-to-asn.html`.
22. Craig A. Shue, Andrew J. Kalafut, and Minaxi Gupta. Abnormally malicious autonomous systems and their internet connectivity. *IEEE/ACM Trans. Netw.*, 20(1):220–230, February 2012.
23. Sushant Sinha, Michael Bailey, and Farnam Jahanian. Shades of Grey: On the Effectiveness of Reputation-based "blacklists". In *Proceedings of MALWARE '08*, pages 57–64, October 2008.
24. Shobha Venkataraman, Subhabrata Sen, Oliver Spatscheck, Patrick Haffner, and Dawn Song. Exploiting network structure for proactive spam mitigation. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*. USENIX Association, 2007.
25. Yinglian Xie, Fang Yu, Kannan Achan, Eliot Gillum, Moises Goldszmidt, and Ted Wobber. How dynamic are ip addresses? In *Proceedings of SIGCOMM '07*, pages 301–312, 2007.
26. Jian Zhang, Phillip Porras, and Johannes Ullrich. Highly Predictive Blacklisting. In *Usenix Security 2008*, August 2008.