Evaluating the Potential of Collaborative Anomaly Detection

Haakon Ringberg[*], Augustin Soule[†], and Matthew Caesar[‡]

[*]Princeton University, [†]Thomson, [‡]UIUC

Abstract. Unwanted traffic is a serious problem for users and operators of networks. *Collaboration* amongst victim machines or networks, for example by exchanging lists of suspected attackers, has been proposed to mitigate this problem. However, the performance of such techniques on real Internet traffic is not well understood. Here, we improve upon this understanding by correlating several large spam corpora with flow traces from an ISP network, detecting malicious behavior of end hosts, and evaluating the ability of end-hosts to coordinate to block these attacks.

We have made several key findings. First, malicious hosts often attack many victims within relatively short time periods. Hence, collaboration techniques that maintain a small window of history work well, and are often sufficient to block the majority of attacks. We found that small groups of 100 collaborative end-hosts is enough to mitigate up to 90% of anomalous events such as DoS or port scans. Second, malicious hosts commonly alternate between a range of malicious behavior (including DDoS, scanning, and spamming). Based on this finding, we propose *cross-class* anomaly detection, where hosts monitor and share information across different kinds of attacks, further increasing the benefits of collaboration.

1 Introduction

Unwanted traffic from malicious hosts is a tremendous problem in the Internet today. DDoS attacks, exploit scanning, email and instant-message spam, click fraud, and other forms of malicious behavior are a common occurrence [1–3]. Vulnerabilities in network software have led to the rapid proliferation of automated attack methods (worms, botnets, viruses), and it is estimated that 25% of all personal computers may be infected by malware [4]. Organizations are estimated to lose billions of dollars per year to malware [5], and single botnet was recently discovered that contained over a million hosts and had caused over $20 million USD in economic losses [6].

Defending against attack traffic can be extremely challenging. The stealthy nature of many attacks, where malicious hosts emulate the characteristics of well-behaved traffic, limits the ability of any one host or network to detect or filter malicious activity in isolation. In order to counter this emerging threat, previous work has proposed that victim sites *collaborate* to build a shared defense against attacks [7–10].

While the notion of victim collaboration has been previously proposed in the literature, the extent to which it improves the ability to detect and isolate malicious traffic has not been rigorously evaluated. In order to design the most effective mitigation techniques, and to determine how existing collaborative architectures would perform in practice, we need to build an understanding of the sorts of workloads botnets generate across different host sites. Building this understanding requires studying what botnet activity looks like when viewed across several vantage points, and precisely how these vantage points may monitor traffic and exchange information to best isolate attacks. To the best of our knowledge, our work is the first to directly measure the benefits of victim cooperation on ISP-level traffic traces. In particular, we apply standard network anomaly detectors to identify unwanted traffic, and analyze the ability of a representative set of collaboration schemes to assist the victims in isolating and mitigating these attacks.

Our measurement study is based on IP flow traces from GEANT, a European ISP operated by a consortium of research and educational institutions. We used traces from the twenty routers that make up GEANT's Eurpean backbone network. Our data set spans five months and contains roughly 24.4 billion individual flows. We used standard anomaly detectors [11, 12] to identify in these traces unwanted traffic that collaborating hosts and networks may wish to detect and remove, including DDoS, DoS, port scanning, and IP scanning events. Our final results calculate the number and percentage of attacks that could have been mitigated by a set of collaborating victim end-hosts.

To reduce probability of false positives, we correlate our detected anomalous events with email spam logs from three domains. That is, our hypothesis is that a detected anomaly, such as a port scan, is much less likely to be a false-positive if the port scanning host also sent spam within some short period of time of the port scan. In our experiements we set this interval to one hour in order to eliminative the vast majority of DHCP changes [13].

Is this story better with Project-Honeypot/Spamhaus traces? What is the change in our results?

Our experiments show that malicious hosts often have a high degree of fan-out, with 1% of attackers collectively attacking 99% of victims. These high-profile attackers tend to be visible to a wide number of victims, and our results indicate that if only 100 victims participate, they can collectively block 85% of their attacks via a blacklist-based filtering mechanism. Second, malicious hosts are *repeat offenders*, often attacking the same set of hosts multiple times within a short period of time. This result indicates that the blacklisting architecture should keep track of history of attacks, and we find that keeping a small window of history is sufficient. Third, malicious hosts often alternate between a wide range of malicious behavior, executing DDoS, scanning, and spamming. This observation underscores the need for victims to perform *cross-class* collaborative filtering, in which victims maintain and share information about multiple different attack types.

Roadmap: We start by presenting our data sources and methodology in Section 2. In Section 3 we study the benefit of performing *cross-class* anomaly detection at a single host. In Section 4 we investigate victim collaboration, both with and without cross-class anomaly detection. We then briefly discuss related work in Section 5 and conclude in Section 6.

2 Methodology

To evaluate the potential of victim collaboration for anomaly detection, we require a large, representative set of unwanted traffic. We accomplish this by extracting attacker IP addresses from several spam corpora, which we correlate with IP flow traces to determine what other victims these attackers affected. These traces and our procedure for correlating them are described in Section 2.1. In addition, we wish to study the benefit of victim collaboration across multiple types of attacks, which requires classifying attack patterns from our IP flow traces. We leveraged anomaly detection techniques presented in previous work [11, 12], which we describe in Section 2.2. Finally, some hosts in the Internet are dynamically assigned IP addresses. Since we use IP addresses to label attackers, some of our long-term results could be affected. In Section 2.3 we describe how we designed our experiments to mitigate this effect.

2.1 Data Sources

Spam traces: We collected spam feeds from three large domains ranging over a five-month period. Given that non-malicious hosts are unlikely to generate spam email (and given that the vast majority of spam arises directly from malicious hosts [14, 3]), we use this data set to extract IP addresses associated with malicious hosts. From these feeds we extracted 1.6 million malicious hosts. Our first spam feed is from a medium-sized software company from which we received a total of 6.6 million spam emails at an average rate of $7\,500$ spam emails per hour. Another of our large feeds includes a log from anti-spam software from a large university in the United States. This data set was collected over a shorter period of time in early 2008 and contained 302 thousand spam emails. Our third feed was collected from a privately-owned "sinkhole" domain specifically established to collect spam. This feed comprised 1.2 million spam emails collected over a 2-month period. As compared to [15, 3, 16], our combined sources yielded 8 million total spam emails (for a total of 50 gigabytes) from which we extracted 1.6 million unique IP addresses.

NetFlow traces: Next, in order to understand the behavior of malicious hosts across the wide-area, we correlate the set of malicious host IP addresses from the spam feeds with flow traces collected from the GEANT ISP network backbone [17]. The GEANT network interconnects 30 National Research and Education Networks representing 34 countries across Europe. GEANT maintains multiple redundant connections to the Internet and provides transit service to its customers. The network operation center routinely collects flow and routing information and make them available to the research community. Each of the 20 Geant routers samples 1-in-1000 packets and exports the flow headers to a central collector using a NetFlow-like [18] format. In this paper we used flow traces collected between October and December 2007. This data set represents 264 gigabytes of flow header information, and we saw 205 million distinct IP addresses during this period.

Roughly 85% of the malicious IP addresses extracted from our spam traces (which we refer to henceforth as *spammer IPs*) were also found in our GEANT NetFlow traffic traces. This was despite the fact that we only observe a small fraction of all Internet

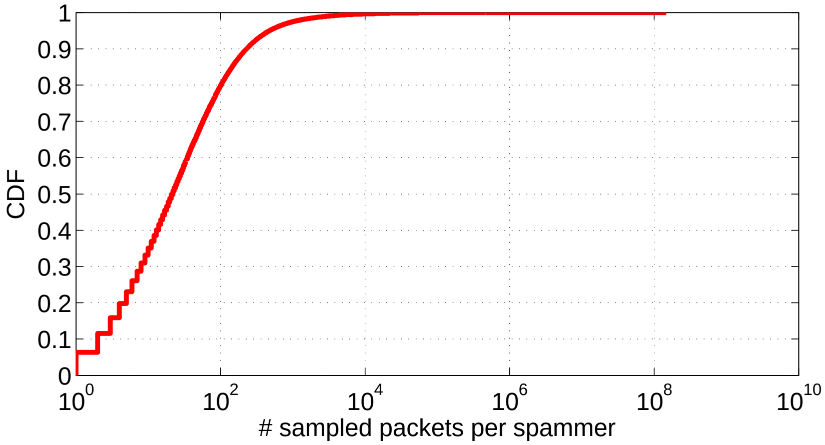


Fig. 1: The amount of *sampled* traffic that we observe for the spammers visible in our flow traces from Geant

traffic in the GEANT trace (since our trace only covers a single ISP, which heavily samples its traffic). Moreover, we observed a large amount of traffic directed from and to these spammer IPs. Figure 1 shows the distribution of sampled packets observed for a given spammer over the entire three months of flow traces. Our traces contain at least 100 sampled packets for over 80% of hosts in our data set (*i.e.*, an expected 100 thousand packets). We also note that an active minority of hosts sent millions of sampled packets.

2.2 Anomaly Detectors

Collaborative network anomaly detection is not a new idea. Communal blacklists are widely used to mitigate spam [19–22], and researchers have proposed leveraging victim collaboration in response to other threats, such as worms [8] and other self-propagating code [7]. Our study differs in that we wish to compare the benefit of victim collaboration across multiple different types of anomalies. In order to do this, we must first classify attack behavior of malicious hosts into classes. This paper studies four representative kinds of attacks: IP scans, port scans, Denial of Service (DoS) attacks, and Distributed Denial of Service (DDoS) attacks.

To identify these attacks, and to categorize attack patterns of malicious hosts, we leverage mechanisms developed as part of previous work. To identify port scans, host scans, and DoS attacks, we used what we coin “definitional anomaly detectors”. These detectors precisely specify what constitutes an anomaly in terms of observable features, which means that every detected event has exactly the specified characteristics. For example, to detect port scans, we leverage a method proposed in [12], which triggers a detection whenever a given source IP address contacts more than α different destination port numbers on a single destination IP address within a Δ-second time window. Definitional anomaly detectors have been used by previous work [23, 11]. We selected these approaches to ensure that anomalies are precisely defined, to simplify understanding of our results and to allow other researchers to exactly reproduce our results given our data sets. To identify DDoS attacks, we leveraged the LADS algorithm designed by Sekar *et. al.* [11]. LADS aims to be an efficient DDoS detection algorithm with low operational complexity, and works by detecting volume anomalies in traffic traces. In particular, LADS monitors for each host the number of flows in the trace that have that host as the destination. LADS then searches for periods of time when the number of flows to a host exceeds a threshold number α of standard deviations larger than the average.

The description of our anomaly detectors would be incomplete if we did not specify the values of α (the threshold for detection) and Δ (the length of the time window). In the case of the port scan detector, for example, this means determining the number of ports that must be contacted on a single host to qualify as a port scan. We chose to tune these parameters by observing the distribution of the underlying parameter over our Geant traffic trace. For example, the distribution of "number of times a source host contacts α destination ports on a single destination IP address within a Δ-second time window" is used to determine a threshold for our port-scanning detector. A part of this specific distribution can be seen in

Type	Definition	n	% Flows
DoS	More than n connections initiated to a single destination host and port pair	200	0.038
DDoS	More than n inbound connections initiated to a single destination host	100	$3.5e^{-5}$
Port Scan	More than n destination ports contacted on a single host	25	0.012
IP Scan	More than n hosts contacted	100	0.024

Table 1: Anomaly Detectors

Based on these distributions, we configure our anomaly detectors by selecting thresholds that capture only the largest attacks. We do this so as to focus our study on the most significant attacks, and to limit the number of false positives. Moreover, our anomaly detectors are only used to classify traffic behavior of known spammers, further reducing false positive rate. The precise specification of all our definitional anomaly detectors can be seen in Table 1. Note that all definitions are given with respect to sampled data. We chose to use a detection window length Δ of 1 minute. The fraction of IP flows in our Geant trace that trigger detection The distribution of different types of attacks, computed by applying our detectors to the flow data set, is shown in Table 2.

The very low fraction of flows that trigger detection (as shown in Table 1) indicates that only genuine attacks are detected. However, there is still a possibility that a few

action	DDoS	DoS	IPscan	Portscan
# anomalous flows	$2.6e^6$	$7e^6$	$232e^6$	$1.7e^6$

Table 2: Number of anomalous flows for various kinds of attacks

of our detected attacks are false positives. Quantifying the likelihood of false positives by manually labeling each of the roughly 24.4 billion in our traces in order to discover the ground truth does not seem tractable. Instead, we correlated the source IP addresses of the detected attacks with the SpamHaus blacklist [21]. Only $12\,\%$ of our anomalous flows were from source IPs that were not in the SpamHaus blacklist *at the time of the attack*. Given that the SpamHaus blacklist is commonly trusted as being accurate enough for widespread use (it is used to protect over 500 million email accounts [21]) false positives should provide little impact on our results.

2.3 Dynamic Addresses

Some of our results involve estimating the gain of sharing *blacklists* as a form of victim collaboration. These blacklists contain IP addresses of hosts engaging in malicious behavior, so that remote networks, NIDS, and end hosts may block connections from these hosts or give them lower priority. For example, an attacker IP χ that attacks a victim ν at some time t_1 is blacklisted by our system at time t_1, which means that an attack by χ at some later time t_2 will be deemed ineffective. However, if the IP addresses of some of these attackers are dynamically assigned, then estimating this gain can be challenging. For example, if the IP χ is dynamically assigned, the host associated with χ at time t_1 may not be the same as the host associated with χ at time t_2. This presents a problem, since well-behaved hosts may be erroneously denied access due to the blacklisting scheme.

We minimize the chance that dynamic addressing will influence our results by leveraging a technique used by some DNSBLs to deal with this problem [21]. Namely, we associate an expiry timeout with entries included in the blacklist. That is, instead of permanently blocking (or reducing priority) of the attacker's IP address, our approach would only affect the attacker during the interval $[t_1 + \epsilon, t_1 + \mathit{\Delta}]$ where $\mathit{\Delta}$ is the "blacklist duration" parameter, which determines how long a an entry remains on the blacklist.

The use of the parameter $\mathit{\Delta}$ has two key benefits for our study. First, by tuning $\mathit{\Delta}$ to a small value, this allows us to minimize the potential influence of dynamic addressing on our experimental results. While tuning $\mathit{\Delta}$ to a small value also reduces the benefits of collaboration, we later show that even relatively small values of $\mathit{\Delta}$ (i.e. smaller than the vast majority of DHCP lease times [13]) still provides most of the benefits of collaboration. Secondly, incorporating $\mathit{\Delta}$ into our study allows us to explore its effects as part of a real system. For example, $\mathit{\Delta}$ may provide a useful knob to enterprises leveraging a blacklisting scheme by allowing them to trade off the security of the site with inconvenience of customers that have inadvertently triggered a false-alarm detection. Unless otherwise mentioned, our experiments measure the fraction of attacks that could have been mitigated using blacklisting as a function of the duration parameter $\mathit{\Delta}$. To set a reasonable default value for $\mathit{\Delta}$, we leverage the results of Xie *et. al.* [13], which shows that over 95% of hosts retain their dynamic IP address for longer than one hour. Our mitigation evaluation experiments that do not include Δ as an explicit parameter therefore have it specifically set to one hour, which means that dynamic address issues should not significantly influence our results.

3 Cross-Class Anomaly Detection

In this section we evaluate the benefit that a *single* victim ν—an IP address that was affected by the set of malicious hosts—can receive from cross-class IP blacklisting. We start by studying attack patterns from the standpoint of a single victim host, and then evaluate the benefit of performing cross-class anomaly detection at that host.

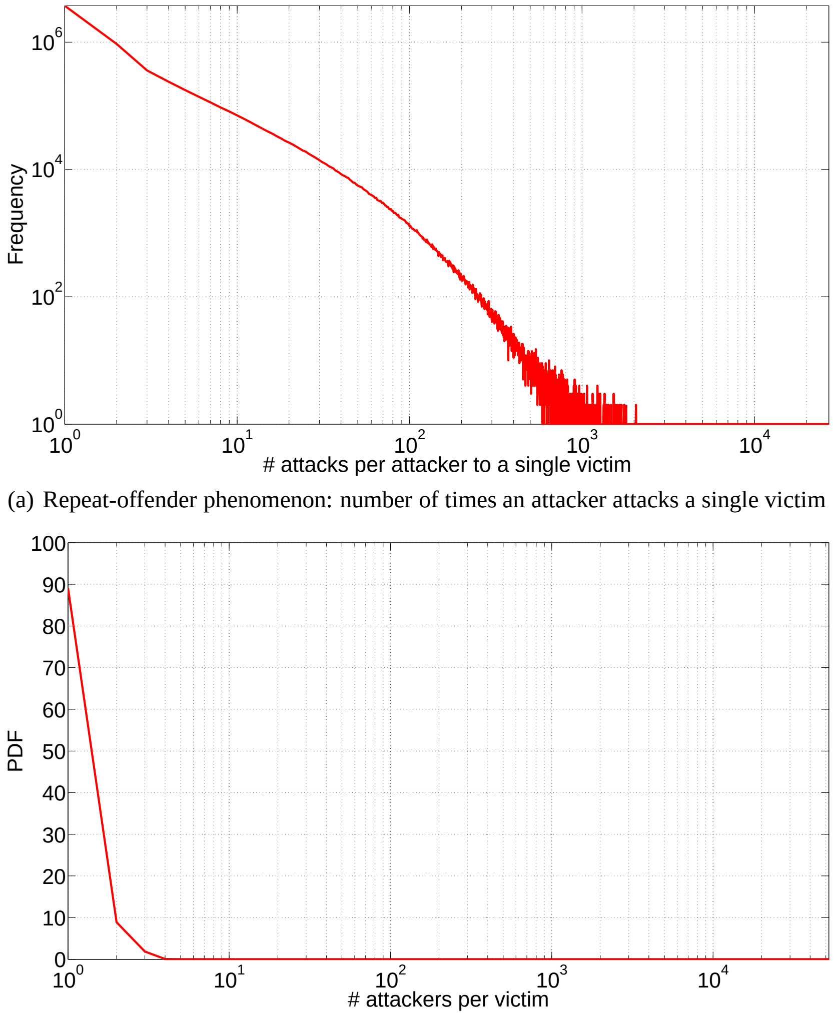
same victim.

Attack patterns at a single victim: Here, we study the entire set of attacks, across all classes of attacks, observed at individual hosts. Our results indicate that a large fraction of malicious hosts attack the same victim host many times. This *repeat-offender* phenomenon is shown in Figure 2(a). There are tens of thousands of instances where a single victim is attacked repeatedly by a single attacker. There are also extreme cases where a victim is attacked thousands of times by a single malicious attacker. Moreover, there are victims that are attacked by millions of attackers. This can be seen from Figure 2(b), where we plot the number of unique attackers per victim. These facts bode well for blacklist-based techniques—a single victim can mitigate a large fraction of attacks by performing cross-class anomaly detection.

Benefit of cross-class anomaly detection: To characterize the benefit of performing cross-class anomaly detection at an individual host, we define a simple model of a blacklisting scheme, and evaluate performance of that scheme on our traces. In our model a malicious host χ that attacks a victim ν at time t_1 is blacklisted for the next Δ hours, and hence any repeated attack by χ on ν within Δ hours is counted as ineffective (or *avoided*). The number of ineffective attacks under this model is shown in Figure 3. Each curve in the plot represents a different anomaly (DoS, DDoS, IP scan, and port scan). The curve marked *all* plots the fraction of ineffective attacks if hosts perform cross-class anomaly detection. Interestingly, we find that blacklisting is substantially more effective for certain anomalies than others. For example, DDoS attackers are much less likely to return to the same victim within a short period of time, greatly reducing the effectiveness of blacklisting. DDoS attacks tended to comprise large bursts of flows sent to hosts within a small time window, and hence the number of times a particular malicious host was observed within the window was smaller than for other kinds of attacks. Overall, however, 70% of attacks could have been rendered ineffective if victims blocked attackers for only an hour.

4 Victim Collaboration

Next, we study the benefit individual victims can gain by *collaborating*, i.e., sharing information about their blacklists. We first study the visibility of attacks across victims, and then characterize the ability of hosts to cooperate to block these attacks.



(a) Repeat-offender phenomenon: number of times an attacker attacks a single victim

(b) Attackers per victim

Fig. 2: Visibility of attack patterns at a single victim host

Attack patterns across victims: The effectiveness of victim collaboration is a function of three parameters. First, the distribution of the number of attacks per attacker to a given victim, which we previously showed in Figure 2(a). Second, the number of different victims that a single malicious host attacks is shown in Figure 4. If a malicious host attacks k victims once each then the upper-bound on the number of attacks that could have been prevented by victim collaboration is $k - 1$. Interestingly, there is a wide variation in this number across hosts, meaning that some attackers can be more easily rendered ineffective by collaboration than others. Finally, the fraction of victims

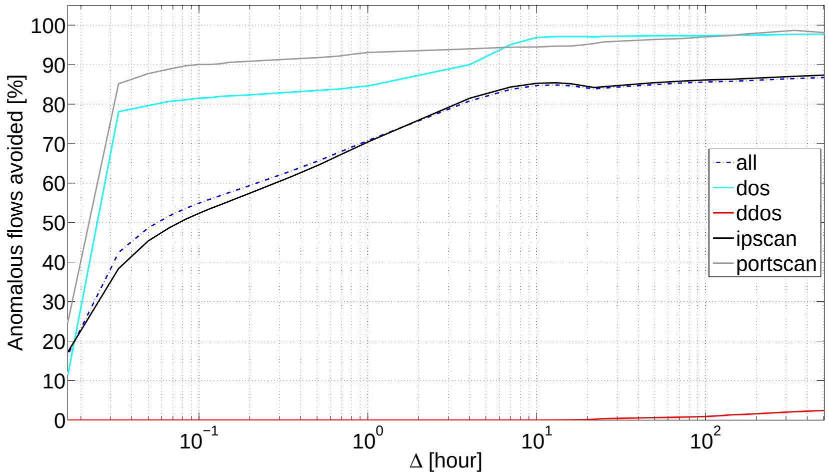


Fig. 3: Effectiveness of cross-class anomaly detection by victims

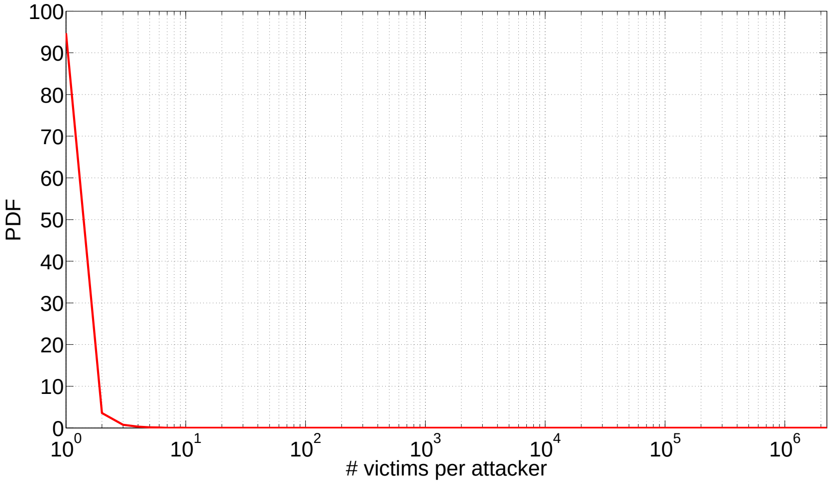
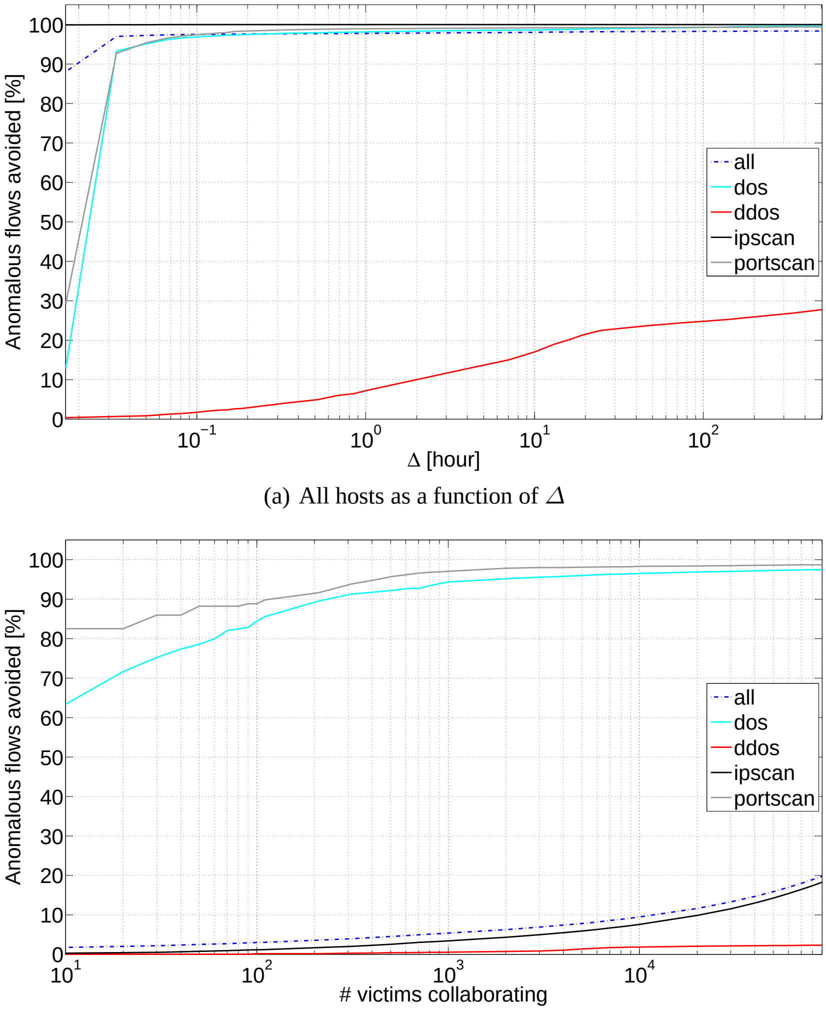


Fig. 4: Victims per attacker

that participate in the scheme also affects performance. We study sensitivity of this final parameter below.

Benefit of victim collaboration:

To estimate benefit of victim collaboration we define the following model. A malicious host χ that attacks a victim $\nu_i \in V$ at time t_1 is blacklisted for the next Δ hours, which means that any repeat attack by χ on any victim $\nu_j \in V$ within Δ hours is counted as ineffective. V represents the set of hosts that are participating in the blacklist collaboration scheme. While this model represents a simplified generalization of a real system, we define it thus so as to characterize expected behavior across a wide range of collaboration techniques.



(a) All hosts as a function of Δ

(b) $\Delta = 1$ hour

Fig. 5: Effectiveness of cross-class victim collaboration

To *upper-bound* the potential gain of victim collaboration in our model, we first consider the idealized scenario where *all* hosts on the Internet collaborate. Results from this experiment are shown in Figure 5(a), which indicates it is sufficient to blacklist attackers for a mere 10 minutes to mitigate more than 90% of DoS, IP scans, and port scans. Moreover, whereas Figure 3 indicates that blacklisting is not effective at mitigating

DDoS attacks when used by a single victim, it is much more effective in a collaborative setting. Nearly 20% of DDoS events would have been mitigated through blacklisting attackers for 10 hours. While assuming all hosts collaborate is unrealistic, performing in-network detection by a small number of cooperating *networks* is known to have very high visibility [24], and hence achieving a fraction of this performance might not be inconceivable. Next, we also consider the more realistic scenario where only a subset of victims participate. In Figure 5(b) we plot the fraction of attacks rendered ineffective assuming the n most victimized hosts collaborated. Here, we see that the vast majority of one-to-one attacks (port scans, DoS) affect a small set of hosts. These victims can collaborate to substantially mitigate these attacks. In fact, 200 participants render 90% of attacks ineffective. One-to-many attacks (IP scans) and many-to-one attacks (DDoS) need more participants, requiring tens of thousands to approach 25% of IP scans. Hence, to block these attacks, some form of network-level collaboration, where network routers or NIDS work together to perform collaboration on behalf of hosts (and thus gain benefit of large groups of hosts collaborating), seems necessary. Finally, we note from the figure that there is an incremental benefit for hosts participating in the scheme. Any non-participant is very likely to improve its own detection accuracy by participating, providing a natural incentive to join, which may accelerate deployment.

5 Related Work

A great deal of research work has been done to allow enterprises and networks to correlate observations from various vantage points in order to improve anomaly detection. The majority of this work has analyzed traffic traces and leveraged general statistical techniques, *e.g.*, [25, 26]. While these techniques have shown promise for intranetwork anomaly detection, they have not been extended to cross-organizational settings where there will be many more vantage points and thus the computational expense of the correlation is much greater. The most well-known technique for such victim collaboration is the sharing of spam blacklists, *e.g.*, [21, 22], but researchers have also proposed leveraging victim collaboration in response to other threats such as worms [8] and self-propagating code [7]. As far as we are aware, however, our work is the first to evaluate the potential benefit of cross-organizational collaboration over a range of relevant attacks. Our results motivate architectures such as the one proposed in [10].

6 Conclusions

Given the extreme challenges in identifying and filtering unwanted traffic, some form of victim collaboration seems necessary. This paper characterizes the ability of victims to establish a shared defense against attacks, over a variety of attack types and workloads. Towards this goal, we mine and correlate observations across several large data sets. We also propose the use of cross-class collaboration to further increase the benefit of victim collaboration.

Moving forward, we plan to pursue several key directions. First, some hosts and networks may be unwilling to share certain kinds of traces and anomalies with each other. We therefore plan to develop and evaluate a scalable *privacy-preserving* architecture [10], to detect correlations without forcing participants to reveal sensitive information. Our other ongoing efforts involve collecting and correlating across a wider range of traces, including botnet IRC logs and instant message spam.

References

1. Franklin, J., Paxson, V., Perrig, A., Savage, S.: An inquiry into the nature and cause of the wealth of internet miscreants. In: Proc. ACM CCS. (October 2007)
2. Ianelli, N., Hackworth, A.: Botnets as a vehicle for online crime. In: CERT Coordination Center. (December 2005)
3. Ramachandran, A., Feamster, N.: Understanding the network-level behavior of spammers. In: ACM SIGCOMM. (2006)
4. Weber, T.: Criminals 'may overwhelm the web'. In: BBC News. (January 2007) `http://news.bbc.co.uk/1/hi/business/6298641.stm`.
5. Peterson, P.: The billion dollar problem (interview). In: IT Security. (January 2007) `http://www.itsecurity.com/interviews/billion-dollar-problem-ironport-\malware-012607/`.
6. Wilson, T.: FBI nabs eight in second 'bot roast'. In: Dark Reading. (November 2007) `http://www.darkreading.com/document.asp?doc_id=93625`.
7. Kannan, J., Subramanian, L., Stoica, I., Katz, R.: Analyzing cooperative containment of fast scanning worms. In: SRUTI. (July 2005)
8. Moore, D., Shannon, C., Voelker, G., Savage, S.: Internet quarantine: Requirements for containing self-propagating code. In: IEEE Infocom, San Francisco, CA, USA (2003)
9. Soule, A., Ringberg, H., Silveira, F., Rexford, J., Diot, C.: Detectability of traffic anomalies in two adjacent networks. Passive And Active Measurement Conference (2007)
10. Allman, M., Blanton, E., Paxson, V., Shenker, S.: Fighting coordinated attackers with cross-organizational information sharing. In: HotNets-V. (November 2006)
11. Sekar, V., Duffield, N., Spatscheck, O., van der Merwe, J., Zhang, H.: Lads: Large-scale automated DDoS detection system. In: USENIX Annual Technical Conference. (2006)
12. Allman, M., Paxson, V., Terrell, J.: A brief history of scanning. In: ACM IMC. (October 2007)
13. Xie, Y., Yu, F., Achan, K., Gillum, E., Goldszmidt, M., Wobber, T.: How dynamic are ip addresses? ACM SIGCOMM (2007)
14. Berinato, S.: Attack of the bots. In: Wired. (November 2006) `http://www.wired.com/wired/archive/14.11/botnet.html`.
15. Kreibich, C., Kanich, C., Levchenko, K., Enright, B., Voelker, G.M., Paxson, V., Savage, S.: On the spam campaign trail. Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET) (2008)
16. Anderson, D.S., Fleizach, C., Savage, S., Voelker, G.M.: Spamscatter: Characterizing internet scam hosting infrastructure. USENIX Security Symposium (2007)
17. Geant Network: `http://www.geant.net/`.
18. Cisco NetFlow: `http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html`.
19. Project Honey Pot: `http://www.projecthoneypot.org/?rf=43429`.
20. DShield: `http://www.dshield.org`.
21. SpamHaus: `http://www.spamhaus.org`.
22. Spamcop: `http://www.spamcop.net`.

23. Kompella, R.R., Singh, S., Varghese, G.: On scalable attack detection in the network. In: ACM IMC. (2004)
24. Feamster, N., Dingledine, R.: Location diversity in anonymity networks. In: WPES. (2004)
25. Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. In: ACM SIGCOMM. (2004)
26. Soule, A., Salamatian, K., Taft, N.: Combining filtering and statistical methods for anomaly detection. In: ACM IMC. (2005)