Traffic Aggregation for Malware Detection

Michael K. Reiter Ting-Fang Yen

December 16, 2007
CMU-CyLab-07-017

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Traffic Aggregation for Malware Detection

Michael K. Reiter[*†]
reiter@cs.unc.edu

Ting-Fang Yen[†]
tyen@andrew.cmu.edu

Abstract

Stealthy malware, such as botnets and spyware, are hard to detect because their activities are subtle and do not disrupt the network, in contrast to DoS attacks and aggressive worms. Stealthy malware, however, does communicate to exfiltrate data to the attacker, to receive the attacker's commands, or to carry out those commands (e.g., send spam). Moreover, since malware rarely infiltrates only a single host in a large enterprise, these communications should emerge from multiple hosts within coarse temporal proximity to one another. In this paper, we describe a system called T$\bar{\text{A}}$MD (pronounced "tamed") with which an enterprise can identify infected computers within its network by finding new communication "aggregates" involving multiple internal hosts, i.e., communication flows that share common characteristics. We describe characteristics for defining aggregates—including flows that communicate with the same external network, that share similar payload, and/or that involve internal hosts with similar software platforms—and justify their use in finding infected hosts. We also detail efficient algorithms employed by T$\bar{\text{A}}$MD for identifying such aggregates, and demonstrate a particular configuration of T$\bar{\text{A}}$MD that identifies new infections for multiple bot and spyware examples with very few false detections, within traces of traffic recorded at the edge of a university network. This is achieved even when the number of infected hosts comprise only about 0.0065% of all internal hosts in the network.

1 Introduction

It is clearly in the interest of network administrators to detect computers within their networks that are infiltrated by spyware or bots. Such stealthy malware can exfiltrate sensitive data to adversaries, or lie in wait for commands from a bot-master to forward spam or launch denial-of-service attacks, for example. Unfortunately it is difficult to detect such malware, since by default it does little to arouse suspicion: e.g., generally its communications neither consume significant bandwidth nor involve a large number of targets. While this changes if the bots are enlisted in aggressive scanning for other vulnerable hosts or in denial-of-service attacks—in which case they can easily be detected using known techniques (e.g., [33, 24])—it would obviously be better to detect the bots prior to such a disruptive event, in the hopes of averting it. Moreover, such easily detectable behaviors are uncharacteristic of significant classes of malware, notably spyware.

We hypothesize, however, that even stealthy, previously unseen malware is likely to exhibit communication that is detectable, if viewed in the right light. First, since emerging malware rarely infects only a single victim, we expect its characteristic communications, however subtle, to appear roughly coincidentally (e.g., within an hour of one another) at multiple hosts in a large network. Second, we expect these communications to share certain features that differentiate them from other communications that are typical of that

[*]University of North Carolina, Chapel Hill, NC, USA
[†]Carnegie Mellon University, Pittsburgh, PA, USA

network. Of course, these two observations may pertain equally well to a variety of communications that are not induced by malware, and consequently the challenge is to refine these observations so as to be useful for detecting malware in an operational system.

In this paper we describe such a system, called TAMD, an abbreviation for "Traffic Aggregation for Malware Detection". As its name suggests, TAMD distills *traffic aggregates* from the traffic passing the edge of a network, where each aggregate is defined by certain characteristics that the traffic grouped within it shares in common. By refining these aggregates to include only traffic that shares multiple relevant characteristics, and by using past traffic as precedent to justify discarding certain aggregates as normal, TAMD constructs a small set of new aggregates (i.e., without previous precedent) that it recommends for examination by a human analyst. The key to making TAMD recommend to the human analyst only those aggregates that represent emerging malware communications is the characteristics on which it aggregates traffic, which include:

- **Common destinations**: TAMD analyzes the networks with which internal hosts communicate, in order to identify aggregates of communication to busier-than-normal external destinations. Spyware reporting to the attacker's site or bot communication to a bot-master (e.g., with IRC, HTTP, or another protocol) might thus form an aggregate under this classification.
- **Similar payload**: TAMD identifies traffic with similar payloads or, more specifically, payloads for which a type of edit distance (*string edit distance matching with moves* [7]) is small. Intuitively, command-and-control traffic between a bot-master and his bots should share significant structure and hence, we expect, would have a low edit distance between them. Similarly, replicas of spam messages sent from multiple bots would also have low edit distance, if their bodies are largely the same.
- **Common internal-host platforms**: TAMD uses traffic to passively fingerprint platforms of internal hosts, and forms aggregates of traffic involving internal hosts that share a common platform. Traffic caused by malware infections that are platform-dependent should form an aggregate by use of this characteristic.

Alone, each of these methods of forming traffic aggregates would be far too coarse to accurately identify malware-infected hosts, as legitimate traffic can form aggregates under these characterizations, as well. In combination, however, they can be quite powerful at extracting aggregates of malware communications. To demonstrate this, we detail a particular configuration of TAMD that employs these aggregation techniques to identify internal hosts infected by malware that reports to a controller site external to the network. Since botnets have been observed to switch controllers or download updates frequently, as often as every two or three days [16, 10], each such event gives TAMD an opportunity to identify these communications. We show that with traffic generated from real spyware and bot instances, TAMD was able to reliably extract this traffic from all traffic passing the edge of a university network, with very few false detections.

In addition to identifying aggregates and ways of combining them to find malware-infected hosts, the contributions of TAMD include algorithms for computing these aggregates efficiently. Our algorithms draw from diverse areas including signal processing, data mining and metric embeddings. We will detail each of these algorithms here.

2 Related Work

Botnet detection Previous approaches to botnet detection rely on heuristics that assume certain models of botnet architecture or behavior, such as IRC-based command-and-control [6, 4, 23, 10], the presence of scanning activities, long idle time and short response time for bots compared to humans [29], etc. Karasaridis

et al. [16] proposed an approach for identifying botnet controllers by combining heuristics that assume the use of IRC communication, scanning behavior, and known models of botnet communication. BotHunter [12] models all bots as sharing common infection steps—namely target scanning, infection exploit, binary download and execution, command-and-control channel establishment, and outbound scanning—and then employs Snort with various malware extensions to raise an alarm when a sufficient subset of these are detected. Consequently, malware not conforming to this profile (e.g., spyware or bots engineered differently) would seemingly go undetected by their approach. Ramachandran et al. [32] observed that botmasters lookup DNS blacklists to determine whether their bots are blacklisted. They thus passively monitor lookups to a DNS-based blackhole list to identify bots.

We believe our approach to be fundamentally different from the above approaches in the following respect. While these approaches work from models of malware behavior (not unlike signature-based intrusion detection), our approach simply seeks to identify new aggregates of communication that are not explained by past behavior on the network being monitored. Like all anomaly-detection approaches, our challenge is to demonstrate that the number of identified anomalous aggregates is manageable, but it has the potential to identify a wider range of as-yet-unseen malware. In particular, the assumptions underlying previous systems present opportunities for attackers to evade these systems by changing the behavior of botnets, and these systems will fail to detect other types of malware (e.g., spyware) that do not meet these assumptions.

Various prior works on botnet detection use honeypots (e.g., [2, 30]). As honeypots can only approximately mimic (at best) real user behavior, they may not attract spyware or bots that rely on human action to infect users' machines. Our approach, in not requiring a honeypot, places no assumptions about the infection vector by which attacks occur and whether these vectors present themselves in a honeypot. In doing so, we hope to make our approach as general as possible.

Techniques The techniques we employ for aggregation, specifically on the basis of external subnets to which communication occurs, include some drawn from the signal processing domain (e.g., principal component analysis (PCA)). While others have drawn from this domain in the detection of network traffic anomalies, our approach has different goals and hence applies these techniques differently. Coarsely speaking, past approaches extract packet header information, such as the number of bytes or packets transferred for each flow, counts of TCP flags, etc., in search of volume anomalies like denial-of-service attacks, flash crowds, or network outages [34, 3, 19]. Lakhina et al. [22] studied the structure of network flows by decomposing OD flows (flows originating and exiting from the same ingress and egress points in the network) using PCA. They expressed each OD flow as a linear combination of smaller "eigenflows", which may belong to deterministic periodic trends, short-lived bursts, or noise, in the traffic. Terrell et al. [35] focused on multi-variate data analysis by grouping network traces into time-series data and selecting features of the traffic from each time bin, including the number of bytes, packets, flows, and the entropy of the packet size and port numbers. They applied Singular Value Decomposition (SVD) to the time-series data, and by examining the low-order components, they were able to detect denial-of-service attacks. In general, transient and light-weight events would go unnoticed by these approaches, such as spammers that send only a few emails over the course of a few minutes, as found in a study by Ramachandran et al. [31]. Our work, on the other hand, is targeted at such lighter-weight events and so employs signal processing techniques differently, not to mention techniques from other domains (e.g., metric embeddings, passive fingerprinting).

Our use of signal processing techniques was most directly inspired by the approach of Xie et al. [39] on the Seurat system. This system detects anomalous changes to file-change patterns in hosts within an enterprise, as a host-based way of identifying malware infections. Seurat correlates file changes from different hosts using wavelet analysis and PCA. By filtering out file changes by single hosts and other periodic file

update patterns, Seurat detects aggregate anomalous file system events that happen during worm outbreaks, which might otherwise go unnoticed if only a single host was examined. Our approach to aggregating traffic by external destination can be viewed as an adaptation of their approach to the networking domain.

Another technique we employ is payload inspection, specifically to aggregate flows based on similar content. Payload inspection has been applied within methods for detecting worm outbreaks and generating signatures. Many previous approaches assume that malicious traffic is significantly more frequent or wide-spread than other traffic, and so the same content will be repeated in many different packets or flows (e.g., [33, 18, 26, 15, 27]); we do not make this assumption here. Previous approaches to comparing payloads includes matching substrings [37, 26, 15, 25], hashing blocks of the payload [33, 19], or searching for the longest common substring [21]. Compared to these methods, our edit distance metric is more sensitive and accurate in cases where parts of the message are simply shifted or replaced. Goebel et al. [10] inspected packet payload to find IRC bots with formatted nicknames. They observed that often IRC bots have nicknames with common patterns, such as long random numbers, bot names, or country codes. However, this approach can detect only bots for which the nickname format is known.

Other tools for intrusion analysis include IDABench (`http://idabench.ists.dartmouth.edu`) and the commercial product StealthWatch from Lancope (`http://www.lancope.com`). IDABench is an intrusion analysis system that allows the network administrator to process observed traffic by deploying a combination of plug-ins. However, it is not intended to be a intrusion detection system, and only provides a framework for facilitating the deployment of existing tools, such as tcpdump, ngrep, Snort, or p0f. StealthWatch monitors all traffic at the network border, checking for policy violations or signs of anomalous behavior by looking for higher-than-usual traffic volumes. Although this is similar to our approach of using past traffic as a baseline for identifying busier-than-normal external destinations, StealthWatch does not refine this information using, e.g., payload or platform aggregation as we do here. Consequently, StealthWatch is primarily useful for detecting only large-volume anomalies like port scanning and denial-of-service attacks, in contrast to the subtle behaviors that TAMD identifies.

3 Defining Aggregates

Given a collection of bi-directional flow records observed at the edge of an enterprise network, our system aims to identify infected internal hosts by finding communication "aggregates", which consist of internal hosts that share common network characteristics. Our observation is that, even though stealthy malware do not manifest themselves in volume or frequency, they rarely infect only a single victim in the network, and, if viewed in the right light, the infected hosts exhibit common characteristics that differentiate them from others typical of that network.

More specifically, TAMD deploys three aggregation functions that identify hosts sharing the following characteristics. First, TAMD identifies aggregates of communications from internal hosts that contribute to busier-than-usual destinations, as may be the case for botnets switching controllers or downloading updates from designated sites, or for spyware-infected machines attempting to "phone home". Second, TAMD identifies traffic with payloads for which a type of edit distance is small. This should include command-and-control traffic between a bot-master and his bots, which is likely to have structured syntax, or replicas of spam email messages. Third, TAMD identifies aggregates of traffic involving internal hosts that share a common platform, in order to capture activities caused by malware infections that are platform-dependent.

The aggregation functions take as input collections of flow records, Λ, and output either groups (aggregates) of internal hosts that share particular properties or a value indicating the amount of similarity between the input flow record collections. We presume that each flow record $\lambda \in \Lambda$ includes the IP address of the

internal host $\lambda.\mathsf{internal}$ involved in the communication and the external subnet $\lambda.\mathsf{external}$ with which it communicates. (In our evaluation in Section 5, $\lambda.\mathsf{external}$ is a /24 prefix, for example.) λ also includes some portion of the payload $\lambda.\mathsf{payload}$ of that communication, packet header fields (port numbers, protocol, TTL, etc.), and the start and end time of the communication.

3.1 Destination Aggregates

Previous studies show that the destination addresses with which a group of hosts communicates exhibit pattern and stability over time, both in the amount of traffic sent and in the set-membership of the destinations [1, 20]. Spyware reporting to the attacker's site, bots communicating with a bot-master or executing the bot-master's commands, or other malware activities are thus likely to exhibit communication patterns outside the norm, i.e., contacting destinations that the internal hosts would not have contacted otherwise.

The destination aggregation function $\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$ takes as input two sets $\Lambda, \Lambda_{\mathsf{past}}$ of communication records. The variables α, τ, and ρ are parameters to the analyses of the functions, as described later in this section. By analyzing the external addresses with which internal hosts communicate in Λ and Λ_{past}, the function outputs a set $\mathsf{SuspiciousSubnets}$ of destination subnets for which there is a larger number of interactions with the internal network, using Λ_{past} as a baseline. The function also outputs an integer $\mathsf{numAggs}$ and a set of pairs $(\mathsf{Agg}_i, \mathsf{Sim}_i)$ $(1 \leq i \leq \mathsf{numAggs})$, where Agg_i and Sim_i are sets of internal hosts (IP addresses) that originated traffic in Λ and Λ_{past}, respectively. Intuitively, the hosts in Agg_i contributed to larger-than-usual number of interactions with an external destination subnet in $\mathsf{SuspiciousSubnets}$, and Sim_i consists of hosts that previously demonstrated communication patterns similar to nodes in Agg_i, as described later in this section.

At a high level, the set of selected "suspicious" external destinations, $\mathsf{SuspiciousSubnets}$, is determined after filtering out periodic and regular activities in the communications of the network as represented in the past traffic Λ_{past}. External destinations observed in Λ that do not follow the norm, i.e., that according to Λ_{past} are busier than usual or have not been contacted before, are thus output in $\mathsf{SuspiciousSubnets}$.

In the following analyses, each internal host is represented as a binary vector $v = (v[1], v[2], \cdots, v[k])$ for which the dimensionality k is equal to the number of destinations in $\mathsf{SuspiciousSubnets}$. A dimension $v[i]$ is set to 1 if the internal host communicated with destination i in $\mathsf{SuspiciousSubnets}$ (according to Λ), and 0 otherwise. This vector representation makes it possible to find communication patterns among the internal hosts in terms of contacted external addresses, through efficient signal processing and machine learning algorithms.

Below we describe the three processing steps in $\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$: (i) Trend filtering, which selects the set of suspicious external destinations; (ii) Dimension reduction, which reduces the dimensionality of the above vector v, while preserving most of the information; and (iii) Clustering, which forms clusters of the vectors (i.e., internal hosts) by the destinations they contact.

Trend Filtering Trend filtering aims to remove regular and periodic communications from Λ, so that external destinations showing behavior outside the norm are identified. In particular, the "norm" is defined, for each external destination subnet, by the average number of internal hosts that communicate with that subnet in various periodic intervals, as recorded in Λ_{past}. For example, periodic patterns, such as Windows machines connecting to the Windows update server on a weekly basis or banking websites experiencing traffic spikes on pay day each month, can be inferred from Λ_{past}. The change in activity of a destination in Λ can then be measured by how much more traffic it received in Λ compared to its average values for previous time intervals in Λ_{past}.

Trend filtering is parameterized by the increase percentage threshold, α, that determines if a particular external destination is abnormally busy. A destination is selected to be in $\mathsf{SuspiciousSubnets}$ if the number of internal hosts communicating with it, in Λ, exceeds all periodic average values in Λ_{past} by more than $\alpha\%$.

Dimension Reduction As described above, given $\mathsf{SuspiciousSubnets}$, each internal host can be represented as a k-dimensional binary vector. However, these vectors may be unnecessarily large, and the dimensions may also be redundant or dependent on one another; e.g., retrieving a web page can cause other web servers to be contacted. To identify such relationships between the destinations and to further dimension reduction, we apply Principal Component Analysis (PCA).

PCA [14] is a method for analyzing multivariate data. It enables data reduction by transforming the original vectors onto a new set of orthogonal axes, i.e., principal components, while still preserving most of the original information. This is done by having each principal component capture as much of the variability in the data as possible. After PCA, each new axis represents a weighted sum of the original dimensions.

While a vector originally has length equal to the number of suspicious destinations, the transformed vector after PCA has a dimensionality that is the number of selected principal components, with each dimension now representing a linear combination of the external destinations. The number of selected principal components depends on the amount of variance we want to capture in the data, denoted as the parameter τ. The more variance to be captured, the more accurate the transformation represents the original data, but, at the same time, more principal components are needed, increasing the dimensionality.

Clustering PCA reduces the vector dimensionality significantly, after which hosts connecting to the same combinations of destinations can be identified efficiently through clustering. $\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$ forms clusters of the vectors (i.e., internal hosts) whose traffic is present in Λ using a K-means clustering algorithm [17], which does not require the number of clusters to be known in advance.

1. Randomly select a vector as the first cluster hub. Assign all vectors to this cluster.
2. Select the vector furthest away from its hub as the new cluster hub. Re-assign all vectors to the cluster whose hub it is closest to.
3. Repeat step 2 until no vector is further from its hub than half of the average hub-hub distance.

The distance metric used for comparing vector distances is Cosine distance, i.e., $\mathsf{CosineDist}(v_1, v_2) = \cos^{-1}((v_1 \bullet v_2)/(|v_1||v_2|))$, for two vectors v_1 and v_2, where the symbol $\bullet$ is the dot product between the two vectors, and $|v_1|$ is the length of vector v_1. Cosine distance is essentially a normalized dot product of the vectors, where a particular dimension would contribute to the final sum if and only if both vectors have a nonzero value in that dimension. In our case, each vector represents a particular internal source host, and each dimension represents a linear combination of destination subnets. Cosine distance thus captures well the relationship between source hosts based on the common destinations they contacted.

Let $\mathsf{numAggs}$ denote the number of clusters from the above K-means algorithm, and let Agg_i ($i = 1 \ldots \mathsf{numAggs}$) denote the hosts whose vectors comprise the i-th cluster. As such, Agg_i is an aggregate of internal hosts interacting with the same busier-than-usual external subnets in Λ. We then construct a k-dimensional binary vector for each internal host in Λ_{past} and transform these vectors using the same axes, i.e., principal components, used to transform the vectors for hosts in Λ. These "old" vectors are then evaluated in terms of their similarity to the aggregates. More specifically, this phase takes a parameter ρ. A host from Λ_{past} is added to Sim_i if the vector representing its communications to $\mathsf{SuspiciousSubnets}$ is within ρ times the cluster radius of the hub for Agg_i. In this sense, this host is "similar" to those hosts in Agg_i. Again,

all of SuspiciousSubnets, numAggs and $\{(\mathsf{Agg}_i, \mathsf{Sim}_i)\}_{1 \le i \le \mathsf{numAggs}}$ are output from $\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$.

3.2 Payload Aggregates

Payload inspection algorithms for malware detection have previously focused on either modeling byte-frequency distributions (e.g., [33, 18, 26, 15]), which assumes that malicious traffic should exhibit an observably different byte-frequency distribution from that of normal traffic, or substring matching (e.g., [37, 25]). In contrast to these approaches, our measure of payload similarity is *edit distance with substring moves*, which we choose because it is capable of capturing syntactic similarities between strings, even if parts of one string are simply shifted or replaced. To our knowledge, ours is the first work that detects malicious traffic by comparing (a type of) string edit distance.

For two character strings s_1 and s_2, the edit distance with substring moves $\mathsf{EditDist}(s_1, s_2)$ is defined as the number of character insertions, deletions, or substitutions, or substring moves, required to turn s_1 into s_2. Given a string $s = s[1] \cdots s[\mathsf{len}(s)]$, a substring move with parameters i, j, and k transforms s into $s[1] \cdots s[i-1], s[j] \cdots s[k-1], s[i] \cdots s[j-1], s[k] \cdots s[\mathsf{len}(s)]$ for some $1 \le i \le j \le k \le \mathsf{len}(s)$. For example, swapping labeled parameters in a parameter list would be a substring move in a command string.

The payload comparison function $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda, \Lambda', \mathsf{distinct})$ that we introduce for use in Section 4 takes as input two sets Λ, Λ' of communication records and a boolean distinct, and outputs a value in the range $[0, 1]$. It is parameterized by an edit distance threshold δ_{Ed} that determines if two communication records λ, λ' are "close enough", i.e., if $\mathsf{EditDist}(\lambda.\mathsf{payload}, \lambda'.\mathsf{payload}) \le \delta_{\mathsf{Ed}}$. Its output indicates from among all record pairs $(\lambda, \lambda') \in \Lambda \times \Lambda'$ such that $\lambda.\mathsf{external} = \lambda'.\mathsf{external}$ (i.e., that involve the same external subnet), the (approximate, see below) fraction for which $\mathsf{EditDist}(\lambda.\mathsf{payload}, \lambda'.\mathsf{payload}) \le \delta_{\mathsf{Ed}}$. If distinct is true, only pairs (λ, λ') such that $\lambda.\mathsf{internal} \neq \lambda'.\mathsf{internal}$ are considered; otherwise, all such pairs are considered.

Since Λ and Λ' can be large, computing $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda, \Lambda', \mathsf{distinct})$ by computing $\mathsf{EditDist}(\lambda.\mathsf{payload}, \lambda'.\mathsf{payload})$ for each relevant (λ, λ') pair individually can be prohibitively expensive, i.e., requiring time proportional to $|\Lambda| \cdot |\Lambda'|$, where $|\Lambda|$ denotes the cardinality of Λ. A contribution of our work is an algorithm for approximating the fraction of relevant record pairs (λ, λ') that satisfy $\mathsf{EditDist}(\lambda.\mathsf{payload}, \lambda'.\mathsf{payload}) \le \delta_{\mathsf{Ed}}$ in time roughly proportional to $|\Lambda| + |\Lambda'|$ if δ_{Ed} is small.

To perform this approximation, we first *embed* the EditDist metric within L1 distance L1Dist, where for two vectors $v_1 = v_1[1 \ldots m]$, $v_2 = v_2[1 \ldots m]$, $\mathsf{L1Dist}(v_1, v_2) = \sum_{i=1}^{m} |v_1[i] - v_2[i]|$. That is, we transform each $\lambda.\mathsf{payload}$ into a vector v_λ so that if $\mathsf{EditDist}(\lambda.\mathsf{payload}, \lambda'.\mathsf{payload}) \le \delta_{\mathsf{Ed}}$ then $\mathsf{L1Dist}(v_\lambda, v_{\lambda'}) \le \delta_{\mathsf{L1}}$ for a known value δ_{L1}. We do so using an algorithm due to Cormode et al. [7] called *Edit Sensitive Parsing* (ESP). For this algorithm, the ratio of δ_{L1} over δ_{Ed} is bounded by $O(\log n \log^* n)$, where n is the length of $\lambda.\mathsf{payload}$.[1] In our evaluation in Section 5, $n = 64$ and we set $\delta_{\mathsf{L1}} = \delta_{\mathsf{Ed}} \cdot \log_{10} 64$.

The embedding of EditDist into L1Dist is essential to our efficiency gains, since it enables us to utilize an approximate nearest-neighbor algorithm called *Locality Sensitive Hashing* (LSH) [9] to find vectors (and hence payload strings) near one another in terms of L1Dist (and hence in terms of EditDist), in time roughly proportional to $|\Lambda| + |\Lambda'|$. Briefly, LSH hashes each vector using several randomly selected hash functions; each hash function maps the vector to a *bucket*. LSH ensures that if $\mathsf{L1Dist}(v_1, v_2) \le \delta_{\mathsf{L1}}$, then the buckets to which v_1 and v_2 are hashed will overlap with high probability (and will overlap with much lower probability if not), where probabilities are taken with respect to the random selection of the hash functions. Conse-

[1] $\log^* n$ denotes the *iterated logarithm* of n, i.e., the number of times the logarithm must be iteratively applied before the result is less than or equal to one.

quently, we hash v_λ for each $\lambda \in \Lambda \cup \Lambda'$, and explicitly confirm that $\mathsf{EditDist}(\lambda.\mathsf{payload}, \lambda'.\mathsf{payload}) \leq \delta_{\mathsf{Ed}}$ only for pairs (λ, λ') for which v_λ and $v_{\lambda'}$ hash to at least one overlapping bucket.

While edit distance may not be meaningful for encrypted messages, we can generalize the payload comparison function to define encrypted payload (e.g., detected by its entropy) as "similar". Exploring payload aggregation using other metrics is part of ongoing work; see Section 6.

3.3 Platform Aggregates

Forming traffic aggregates based on platform can be useful in identifying malware infections that are platform dependent. That is, suspicious traffic common to a collection of hosts becomes even more suspicious if the hosts share a common software platform.

Much host platform information can be inferred from traffic observed passively. Passive tools, unlike active fingerprinting tools like Nmap (`http://insecure.org`), do not probe hosts, but rather listen silently. The most comprehensive passive operating system fingerprinting tool of which we are aware is p0f (`http://lcamtuf.coredump.cx/p0f.shtml`), which extracts various IP and TCP header fields from SYN packets and uses a rule-based comparison algorithm. However, p0f cannot be applied to traffic traces in the flow-record format available to us (see Section 5), since most individual packet information (including for SYN packets) is not retained.

At the time of this writing, TĀMD employs two simple heuristics for fingerprinting internal host operating systems passively. The first employs time-to-live (TTL) fields witnessed at the network border in packets from internal hosts. It is well-known that in many cases, different operating system types select different initial TTL values (e.g., see `http://secfr.nerim.net/docs/fingerprint/en/ttl_default.html`). With a detailed map of the internal network, the observed TTL values can be used to infer the exact initial TTL value and so narrow the possibilities for operating system the host is running. However, a detailed map is typically unnecessary, as routes in most enterprise networks are sufficiently short that witnessing TTLs of packets from internal hosts as those packets leave the network enables the initial TTL values to be inferred well enough.

The second heuristic employed in TĀMD watches for host communications characteristic of a particular operating system platform. For example, Windows machines connect to the Microsoft time server by default during system boot for time sychronization, and the FreeBSD packages FTP server is more likely to be accessed by FreeBSD machines to install software updates. Once characteristic communications for different platforms are identified, TĀMD can monitor for these to learn the platform of an internal host.

There are at least two limitations of such passive fingerprinting approaches for our purposes. First, DHCP-assigned IP addresses can be assigned to hosts with different operating systems over time, leading to inconsistent indications of the host operating system associated with an IP address. This suggests that TĀMD should weigh recent indications more heavily than older (and hence potentially stale) indications. Second, a machine with a compromised kernel could, in theory, alter its behavior to masquerade as a different operating system. In the absence of a possible IP address reassignment (e.g., for address ranges not assigned via DHCP), such a shift in behavior should itself be detectable evidence that a compromise may have occurred. In general, however, this limitation is intrinsic to *any* fingerprinting technique, passive or active, except those based on attestations from trusted hardware (e.g., TCG's Trusted Platform Module, `https://www.trustedcomputinggroup.org/groups/tpm/`). While we are unaware of malware that employs such a masquerading strategy, should platform-based aggregation for malware detection become commonplace, such systems would presumably need to migrate to attestation-based platform identification as it matures, in order to detect kernel-level compromises. User-level compromise should not affect platform-based aggregation using conventional fingerprinting techniques, however.

```
FindSuspiciousAggregates(Λ, Λ_past)
100: SuspiciousAggregates ← ∅
101: (SuspiciousSubnets, numAggs, {(Agg_i, Sim_i)}_{1≤i≤numAggs}) ← ByDestination^{α,τ,ρ}(Λ, Λ_past)
                                                        /* Form aggregates by external subnet */
102: for i ∈ 1 ... numAggs do
103:     Λ_i ← {λ ∈ Λ : λ.internal ∈ Agg_i}                          /* Traffic from hosts in Agg_i */
104:     Λ_i^susp ← {λ ∈ Λ_i : λ.external ∈ SuspiciousSubnets}      /* Traffic from hosts in Agg_i to suspicious subnets */
105:     Λ_past^sim ← {λ ∈ Λ_past : λ.internal ∈ Sim_i ∧ λ.external ∈ SuspiciousSubnets}
                                                        /* Past traffic from hosts in Sim_i to suspicious subnets */
106:     if ByPayload^{δ_Ed}(Λ_past^sim, Λ_i^susp, false) = 0 then   /* Keep if no similar past traffic to same external subnet */
107:         if ByPayload^{δ_Ed}(Λ_i^susp, Λ_i^susp, true) > 0.2 then   /* Keep if traffic to same external subnet is self-similar */
108:             if ByPlatform(Λ_i) > 0.5 then                      /* Keep if most of aggregate consists of one platform */
109:                 SuspiciousAggregates ← SuspiciousAggregates ∪ {Agg_i}
110: return SuspiciousAggregates
```

Figure 1: The function used to find suspicious aggregates in the example construction given in Section 4. $\mathsf{ByDestination}^{\alpha,\tau,\rho}$ (line 101), $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}$ (lines 106, 107), and ByPlatform (line 108) are defined in Sections 3.1, 3.2 and 3.3, respectively.

Presently TAMD uses the aforementioned heuristics based on TTL values and communication with characteristic sites to identify platforms. For use in Section 4, we embody this in a function $\mathsf{ByPlatform}(\Lambda)$ that returns the largest fraction of internal hosts in Λ (i.e., among the hosts $\{\lambda.\mathsf{internal} : \lambda \in \Lambda\}$) that can be identified as having the same operating system, based on these heuristics applied to the traffic records Λ.

4 Example Configuration

In this section, we detail a configuration of TAMD that identifies internal hosts infected by malware by employing the functions described in Section 3. This configuration identifies platform-dependent malware infections that report to common sites, e.g., IRC channels for receiving commands, public servers for downloading binaries, denial-of-service victims to attack, or database servers for uploading stolen information, as is typical of most bots and spyware. This configuration is based on several observations about such malware:

- O1. For even moderately aggressive malware, it is rarely the case that only a single victim exists in a large enterprise network, and so we hypothesize that stealthy malware is likely to generate traffic that appears within the same, coarse window of time (e.g., within the same hour) from multiple infected hosts. Moreover, we would expect that the controller site is located in a subnet that would not be a common one with which benign hosts interact (though we do not presume that benign hosts *never* interact with it). As such, malware interacting with the controller site should generate a noticeable increase in the number of interactions with the controller's subnet in that window of time.
- O2. We expect that the contents of the malware communications would be syntactically unlike the contents of previous communication with that subnet (assuming these infections are new, as would be the case in the emergence of a new malware strain). We also expect that the multiple instances of the malware communication to the controller site would themselves be similar.
- O3. In the case of platform-dependent malware, the malware communications to the controller site will involve internal hosts all having the same host platform.

Using these observations, we have assembled the aggregation functions described in Section 3 into an algorithm FindSuspiciousAggregates to identify such malware infections, shown in Figure 1. The input to this

function is a set Λ of traffic records observed in a fixed time interval (e.g., one hour) at the border of the network, and a set Λ_{past} of records previously observed at the border of the network. FindSuspiciousAggregates assembles and returns (in line 110) a set SuspiciousAggregates comprised of suspicious aggregates, where each aggregate is a set of internal hosts (IP addresses) that is suspected of being infected by malware.

FindSuspiciousAggregates first exploits observation O1, using $\mathsf{ByDestination}^{\alpha,\tau,\rho}$ from Section 3.1 to find suspicious external subnets SuspiciousSubnets responsible for noticeably greater communication with the monitored network than in the past, and to find aggregates $\{\mathsf{Agg}_i\}_{1\leq i\leq \mathsf{numAggs}}$, each of which includes internal hosts that interacted with one or more of these subnets. In line with observation O2, FindSuspiciousAggregates then compares the payloads of communications with those suspicious subnets to past communications with those subnets by hosts with similar communication patterns, i.e., hosts in Sim_i. Each aggregate Agg_i for which this communication does not match this previous communication to those subnets (line 106) remains in consideration. Each such aggregate is then tested in line 107 to determine if distinct hosts ($\mathsf{distinct} = \mathrm{true}$) in the aggregate communicate with suspicious subnets using similar payload. Finally, as motivated by observation O3, for each aggregate that has survived these tests, the platforms of the hosts in the aggregate are inferred to the extent possible using ByPlatform and, if the aggregate is adequately homogenous (line 108), then it is added to SuspiciousAggregates (line 109).

There are numerous constants in Figure 1 that we have chosen on the basis of our evaluation that we will present in Section 5. These constants include $\alpha = 150\%$, $\tau = 90\%$, and $\rho = 20$ for $\mathsf{ByDestination}^{\alpha,\tau,\rho}$, 0.2 in line 107 and 0.5 in line 108, and the choice to eliminate aggregates as aggressively as possible in line 106, i.e., whenever $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda^{\mathsf{sim}}_{\mathsf{past}}, \Lambda^{\mathsf{susp}}_i, \mathrm{false}) > 0$. In addition, as we will describe in Section 5, the data on which we perform our evaluation includes 64 bytes of payload per record λ, for which we found $\delta_{\mathsf{Ed}} = 5$ to be an effective value. However, we emphasize that all of these constants can be adjusted in order to make this configuration of TAMD more conservative or liberal in its selection of suspicious aggregates, and we plan to continue evaluation of the alternatives in ongoing work. That said, in Section 5, we show that with traffic generated from real spyware and bot instances, this configuration of TAMD was able to reliably extract malware traffic from all traffic passing the edge of our university network with very few false detections. This reliability is achieved even in tests where the number of simulated infected hosts comprise only about 0.0065% of the total number of internal hosts in the network, calculated as the maximum number of internal IP addresses observed communicating in any one hour period during our data collection (see Section 5).

5 Evaluation

We present an evaluation of the particular configuration of TAMD described in Section 4, using traffic from real spyware and bot instances, which are overlayed onto flow records recorded at the edge of the Carnegie Mellon campus network. The performance of TAMD as observed in this evaluation is described in Appendix C.

5.1 Data collection

Our network traffic traces were obtained from the edge routers of the Carnegie Mellon campus network, which consists of two /16 subnets. The packets are organized into bi-directional flow records by Argus (Audit Record Generation and Utilization System, `http://www.qosient.com/argus`), which is a real time flow monitor based on the RTFM flow model [5, 13]. Argus inspects each packet and groups together those with the same attribute values into one bi-directional record. In particular, TCP and UDP flows are identified by the 5-tuple (source IP address, destination IP address, source port, destination port,

protocol)[2], and packets in both directions are recorded as a summary of the communication, namely, an Argus flow record.

The fields extracted from Argus records are listed in Table 1. The rate of the traffic from the edge of the CMU campus network is about 5000 flow records per second. The traces were collected for six hours daily, from 9am until 3pm.

IP Header	TCP Header	Flow Attribute
Source IP	Source Port	Byte Count
Destination IP	Destination Port	Packet Count
Protocol	Sequence Number	Payload (64 bytes)
TTL	Window Size	

Table 1: Extracted Flow Fields

In our experiments, we assumed that the recorded campus traffic was benign, and so for testing we additionally generated malicious traffic using four instances of malware collected from the internet: Bagle, IRCbot, Mybot, and Sdbot. The characteristics of these malware instances are described in Appendix A, as is our method for collecting flows from them. For testing, we then overlayed flows recorded from these malware instances onto one hour of our recorded campus network traffic, and assigned the malware traffic to originate from randomly selected, supposedly benign, internal hosts, observed to be active during that hour. This makes our testing scenario much more realistic, since the internal hosts to be identified still exhibit their normal connection patterns, in addition to subtle malware activities, as is common for stealthy malware.

We overlayed onto each hour of the recorded campus network traffic eight instances of Bagle, two instances of IRCbot, five instances of Mybot, or five instances of Sdbot. These numbers of instances were chosen to represent a very small fraction of the total campus hosts, specifically at most 0.0065% based upon the number of campus hosts (IP addresses) observed sending traffic in the busiest hour.

In our initial tests, we found that these malware-infected hosts were obscured by certain apparently-benign hosts with highly unusual behavior, which turned out to be PlanetLab (`http://www.planet-lab.org`) and Tor (`http://tor.eff.org`) nodes. The experience of identifying these hosts and their exclusion from our dataset for the experiments reported in Section 5.2 is described in Appendix B.

5.2 Detecting Malware

As described in Section 5.1, the particular configuration of TĀMD was given all traffic collected at the edge of the Carnegie Mellon campus network in hourly batches. For each malware, we overlayed the malware traffic onto one hour in the campus data, assigning the traffic to originate from randomly selected internal hosts observed to be active during that period. The same analysis steps are repeated for each hour over an entire week in March 2007. The number of infected hosts for each experiment varied from 2 to 8, while the maximum number of internal hosts observed communicating in any one hour period was more than 30,000.

The granularity of external destinations was set to be /24 subnets. While the communication records from the current hour was given to FindSuspiciousAggregates as Λ, the set Λ_{past} was selected from communication records in the past that represented the general trend and the periodicity in the traffic. In particular, for the purpose of our experiments, Λ_{past} consisted of traffic from, in reference to the time frame for Λ, (i) the same hour from the same days of the week, (ii) the same hour from the same days of the month, and (iii) all hours from the previous two days.

In all experiments, TĀMD was able to identify the infected hosts with very few false positives. Figure 2 shows the number of aggregates that our particular configuration of TĀMD yielded after applying each

[2]Since Argus records are bi-directional, the source and destination IP addresses are swappable in the logic that matches packets to flows. However, the source IP address in the record is set to the IP address of the host that initiated the connection.

aggregation function to the network traffic, for each malware instance, averaged over all test hours. We assume that the campus traffic represented normal, benign communications, and so in each experiment involving each malware instance, the number of true positives (i.e., the actual malware aggregates) should then always be one, corresponding to the group of randomly selected internal hosts to which we assigned the malware traffic. Figure 2 shows that the number of aggregates is reduced after each aggregation function. The single aggregate consisting solely of infected hosts is always identified, in every malware experiment.

			Bagle	IRCbot	Mybot	Sdbot
1.	$\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$	μ	69.23	101.90	69.38	74.08
	(line 101)	σ	42.06	34.02	40.76	43.87
2.	$\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda^{\mathsf{sim}}_{\mathsf{past}}, \Lambda_i^{\mathsf{susp}}, \mathrm{false}) = 0$	μ	59.15	89.20	61.13	63.46
	(line 106)	σ	32.48	27.77	31.60	33.52
3.	$\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda_i^{\mathsf{susp}}, \Lambda_i^{\mathsf{susp}}, \mathrm{true}) > 0.2$	μ	2.77	3.30	3.00	2.92
	(line 107)	σ	2.17	2.01	2.09	2.22
4.	$\mathsf{ByPlatform}(\Lambda_i) > 0.5$	μ	1.46	2.2	1.81	1.62
	(line 108)	σ	0.66	1.62	1.28	0.96

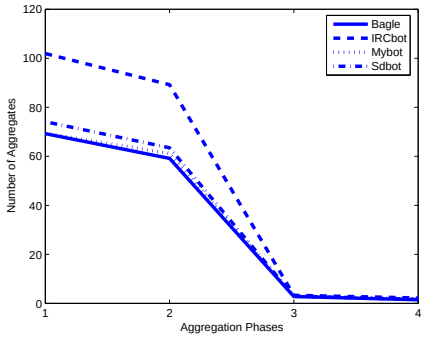

Figure 2: Aggregates remaining after each function in Figure 1 (μ = mean, σ = std. dev.)

5.3 Benign Aggregates

As indicated in Figure 2, our methodology detected a small number of apparently "benign" aggregates (about 0.7 per hour, on average) in addition to the one aggregate of infected hosts that we overlayed on the trace. We found that some of these same benign aggregates regularly appeared for that hour of input data, across different malware experiments. Further investigation of these aggregates, based on the 64 bytes of flow payload available to us, port numbers, and protocol field (for privacy reasons, the IP addresses were anonymized), showed that these aggregates included ICMP probes, SMTP connection timeout messages, and advertising-related HTTP requests; several of these suggest that additional investigation may be warranted. Others included peer-to-peer (P2P) file transfers and connections to online game servers. All of these aggregates consisted of internal hosts contacting rare sites, and often consisted of less than five hosts sharing one or two common destination subnets.

In theory, a group of internal hosts visiting a new popular website (i.e., the "slashdot" effect) could also form an aggregate. However, it is unlikely that all of the hosts would come from the same platform, and in our experiments, we believe we saw very few such false detections. While we consider these results to be promising, our approach is a significant first step that we (and, we hope, others) will build upon to reduce false detections further.

6 Discussion and Ongoing Work

Approaches by which malware writers might attempt to avoid detection by our techniques include encrypting their malware traffic, so that our payload comparisons will be ineffective. To accommodate encryption, our techniques can be generalized to define encrypted content (which itself is generally easy to detect) as "similar"; we are exploring the impact of this adaptation in ongoing work. Malware writers could go further and have their malware communicate steganographically, though at the cost of greater sophistication and

lower bandwidth. Detecting steganographic communication is itself an active area of research (e.g., [28]) from which TAMD could benefit.

A second way that malware writers could try to avoid detection by TAMD is with alternative botnet architectures. Although the vast majority of spyware and botnets found today use a centralized IRC command-and-control server, other botnet architectures have been reported, such as P2P botnets (Phatbot[3], Trojan.Peacomm bot [11], Sinit P2P trojan[4]) or HTTP-based botnets (Clickbot.A [8]). Still others have been proposed, such as hybrid P2P and centralized botnets [36, 38].

Even among these alternative architectures, a large number exhibit characteristics that we believe should be detectable via FindSuspiciousAggregates in Section 4. For example, Trojan.Peacomm bots, while using a P2P network to transfer addresses of compromised web servers among them, still connect to these web servers to download malicious executables for sending spam or performing DoS attacks. This activity of collectively contacting web servers matches the behavior that our techniques successfully detected in our evaluations. The same detection method can also be applied to HTTP-based bots, such as Clickbot.A [8], which commit click frauds by having bots connect to a compromised web server for a list of websites and search keywords, or for a URL to download updated bot versions. Vogt et al. [36] suggested a "superbotnet", where the botnet is composed of individual smaller centralized botnets, and the controllers from each smaller botnet peer together in a P2P network. Since the individual smaller botnets still use a centralized architecture, this should be still be detectable via our techniques. Wang et al. [38] proposed a hybrid P2P botnet where each bot maintains its own peer list and polls other bots periodically for new commands. However, in order to monitor the IP address and resources of each individual bot, the botnet supports a command by which the botmaster can solicit all bots to report to a specific compromised server. Again, this behavior should be detectable by FindSuspiciousAggregates.

That said, some P2P bots avoid contacting a common server for the transfer executables or other tasks, such as Phatbot and the Sinit trojan. While Phatbots find peers by registering themselves as Gnutella clients, the Sinit trojan sends out random probes for peer discovery. In both cases, forming aggregates based on payload similarity should remain effective, provided that similarity is generalized as described above to accommodate encrypted traffic (which Phatbot utilizes). Similarly, platform-based aggregation should also be effective, as both are platform-dependent. We are evaluating these directions in ongoing work, as well as alternative aggregation methods to help identify these types of malware.

7 Conclusion

In this paper, we presented TAMD, a system that detects stealthy malware within a network by identifying internal hosts that share common network characteristics. TAMD employs three aggregation functions to group hosts based on the following characteristics. First, the destination aggregation function, $\mathsf{ByDestination}^{\alpha,\tau,\rho}$, forms aggregates of internal hosts that contact the same combination of busier-than-usual external destinations. A binary vector is formed for each internal host, with each dimension representing one of the selected external destinations. The vectors are processed by PCA for dimension reduction, and clustered by K-means clustering. New clusters are selected as those that do not conform to preceding communication patterns. Second, the payload aggregation function, $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}$, identifies communications with similar payloads in terms of a type of edit distance. This is done by first embedding the payload strings into vectors in L1 space, and then finding close vectors by an approximate nearest-neighbor algorithm. Third, the platform aggregation function, ByPlatform, forms aggregates that involve hosts running on common platforms,

[3] See `http://www.secureworks.com/research/threats/phatbot`.

[4] See `http://www.secureworks.com/research/threats/sinit`.

as inferred using TTL values or platform-specific sites to which they connect.

We detailed a configuration of TĀMD that employs these functions in combination to detect platform-dependent malware infections that report to common sites. A common site might be an IRC channel for receiving commands, a public webserver for downloading binaries, a denial-of-service victim they are instructed to attack, or a database server for uploading stolen information, as is typical of most bots and spyware. Our experiments showed that, with traffic generated from real spyware and bot instances, this configuration of TĀMD reliably extracted malware traffic from all traffic passing the edge of our university network, with very few false detections. This is achieved even in tests where the number of simulated infected hosts comprised only about 0.0065% of all internal hosts in the network.

Acknowledgements

We are grateful to Moheeb Rajab and other members of the Johns Hopkins Honeynet Project (`http://hinrg.cs.jhu.edu/jhuhoneynet/`) for providing the malware binaries that we used in our evaluations. We are also grateful to Chas Difatta, Mark Poepping and other members of the EDDY Initiative (`http://www.cmu.edu/eddy/`) for facilitating access to the network traffic records from Carnegie Mellon University used in this research. This research was supported in part by NSF awards 0326472 and 0433540.

References

[1] W. Aiello, C. Kalmanek, P. McDaniel, S. Sen, O. Spatscheck, and J. Van der Merwe. Analysis of communities of interest in data networks. In *Proceedings of Passive and Active Measurement Workshop*, 2005.

[2] P. Bächer, T. Holz, M. Kötter, and G. Wicherski. Know your enemy: Tracking botnets. Technical report, The Honeynet Project and Research Alliance, 2005.

[3] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop (IMW)*, 2002.

[4] J. R. Binkley and S. Singh. An algorithm for anomaly-based botnet detection. In *Proceedings of the Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, 2006.

[5] N. Brownlee, C. Mills, and G. Ruth. Traffic flow measurement: Architecture. RFC 2722, 1999.

[6] E. Cooke, F. Jahanian, and D. McPherson. The zombie roundup: Understanding, detecting, and disrupting botnets. In *Proceedings of the Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, 2005.

[7] G. Cormode and S. M. Muthukrishnan. The string edit distance matching problem with moves. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2002.

[8] N. Daswani, M. Stoppelman, the Google Click Quality, and Security Teams. The anatomy of clickbot.A. In *Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots)*, 2007.

[9] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Symposium on Computational Geometry*, 2004.

[10] J. Goebel and T. Holz. Rishi: Identify bot contaminated hosts by IRC nickname evaluation. In *Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots)*, 2007.

[11] J. B. Grizzard, V. Sharma, C. Nunnery, B. B. Kang, and D. Dagon. Peer-to-peer botnets: Overview and case study. In *Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots)*, 2007.

[12] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee. BotHunter: Detecting Malware Infection Through IDS-Driven Dialog Correlation. In *Proceedings of the USENIX Security Symposium*, 2007.

[13] S. Handelman, S. Stibler, N. Brownlee, and G. Ruth. New attributes for traffic flow measurement. RFC 2724, 1999.

[14] I. T. Jolliffe. *Principal Component Analysis*. Spring-Verlag, 1986.

[15] V. Karamcheti, D. Geiger, Z. Kedem, and S. M. Muthukrishnan. Detecting malicious network traffic using inverse distributions of packet contents. In *Proceedings of the ACM SIGCOMM Workshop on Mining Network Data (MineNet)*, 2005.

[16] A. Karasaridis, B. Rexroad, and D. Hoeflin. Wide-scale botnet detection and characterization. In *Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots)*, 2007.

[17] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley, 1990.

[18] H. Kim and B. Karp. Autograph: Toward automated, distributed worm signature detection. In *Proceedings of the USENIX Security Symposium*, 2004.

[19] S. S. Kim, A. L. N. Reddy, and M. Vannucci. Detecting traffic anomalies using discrete wavelet transform. In *Proceedings of the International Conference on Information Networking (ICOIN)*, 2004.

[20] E. Kohler, J. Li, V. Paxson, and S. Shenker. Observed structure of addresses in IP traffic. *IEEE/ACM Transactions on Networking*, 14(6), 2006.

[21] C. Kreibich and J. Crowcroft. Honeycomb - creating intrusion detection signatures using honeypots. In *Proceedings of the ACM SIGCOMM Workshop on Hop Topics in Networks (HotNets)*, 2003.

[22] A. Lakhina, K. Papagiannaki, and M. Crovella. Structural analysis of network traffic flows. In *Proceedings of ACM SIGMETRICS/Performance*, 2004.

[23] C. Livadas, B. Walsh, D. Lapsley, and T. Strayer. Using machine learning techniques to identify botnet traffic. In *Proceedings of the IEEE LCN Workshop on Network Security (WoNS)*, 2006.

[24] D. Moore, G. M. Voelker, and S. Savage. Inferring internet denial-of-service activity. In *Proceedings of the USENIX Security Symposium*, 2001.

[25] J. Newsome, B. Karp, and D. Song. Polygraph: Automatic signature generation for polymorphic worms. In *IEEE Security and Privacy Symposium*, 2005.

[26] J. J. Parekh, K. Wang, and S. J. Stolfo. Privacy-preserving payload-based correlation for accurate malicious traffic detection. In *Proceedings of the ACM SIGCOMM Workshop on Large Scale Attack Defense*, 2006.

[27] R. Perdisci, G. Gu, and W. Lee. Using an ensemble of one-class SVM classifiers to harden payload-based anomaly detection systems. In *Proceedings of the International Conference on Data Mining (ICDM)*, 2006.

[28] N. Provos and P. Honeyman. Detecting steganographic content on the Internet. In *Proceedings of the 2002 ISOC Network and Distributed System Security Symposium*, February 2002.

[29] S. Racine. Analysis of Internet Relay Chat Usage by DDoS Zombies. Master's thesis, Swiss Federal Institute of Technology Zurich, 2004.

[30] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. A multifaceted approach to understanding the botnet phenomenon. In *ACM SIGCOMM/USENIX Internet Measurement Conference (IMC)*, 2006.

[31] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *Proceedings of ACM SIGCOMM*, 2006.

[32] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using DNSBL counter-intelligence. In *Proceedings of the Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, 2006.

[33] S. Singh, C. Estan, G. Varghese, and S. Savage. Automated worm fingerprinting. In *Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI)*, 2004.

[34] C. Taylor and J. Alves-Foss. NATE - network analysis of anomalous traffic events, a low-cost approach. In *Proceedings of the New Security Paradigms Workshop (NSPW)*, 2001.

[35] J. Terrell, L. Zhang, Z. Zhu, K. Jeffay, H. Shen, A. Nobel, and F. Donelson Smith. Multivariate SVD analyses for network anomaly detection. In *Poster Proceedings of ACM SIGCOMM*, 2005.

[36] R. Vogt, J. Aycock, and Jr. M. J. Jacobson. Army of botnets. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2007.

[37] K. Wang, J. J. Parekh, and S. J. Stolfo. Anagram: A content anomaly detector resistant to mimicry attack. In *Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection(RAID)*, 2006.

[38] P. Wang, S. Sparks, and C. C. Zou. An advanced hybrid peer-to-peer botnet. In *Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots)*, 2007.

[39] Y. Xie, H. Kim, D. R. O'Hallaron, M. K. Reiter, and H. Zhang. Seurat: A pointillist approach to anomaly detection. In *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID)*, 2004.

A Malware Instances

For our testing described in Section 5, traffic from four malware instances was collected using virtual machine hosts infected with each malware. The virtual hosts were all running the Windows XP Professional operating system with the same VMWare image file. Each run of traffic collection is one hour long.

Bagle[5] is spyware that, on execution, runs as a background process and attempts to download other malicious executables from various sites, while generating pop-up windows and hijacking the web browser to advertising websites. As with other types of spyware and adware, Bagle initiates connections to numerous destinations that are set up to exclusively host advertisements or other malicious content. We collected Bagle traffic by simultaneously running eight instances of Windows XP virtual machine hosts infected with Bagle.

IRCbot[6] is a backdoor trojan that connects to an IRC server and waits for commands from the attacker. In addition, after successfully connecting to the command-and-control center, the bot downloads an update executable from a designated webserver, and goes on to scan the local /16 subnet attacking other machines with the LSASS vulnerability on port 445[7] and the NetBIOS vulnerability on port 139[8]. We collected traffic from two instances of IRCbot running on two Windows XP virtual machine hosts.

Mybot[9] is spyware, a worm, and a bot that connects to an IRC server to wait for commands, and also records keystrokes and steals other personal information on the victim host. This malware is especially subtle in its communications. When it is only waiting for commands on the IRC server, the bot initiates one connection every 90 seconds, in the form of IRC PING/PONG messages. In the hour of our traffic collection, Mybot simply waited for commands on the IRC channel, and its only outbound connections were these PING/PONG messages. We collected traffic for five Mybot instances.

Sdbot[10] is a trojan and a bot that opens a back door to connect to an IRC server. Similar to Mybot, when it is waiting for commands from the attacker, Sdbot only makes outbound connections once every 90 seconds, in the form of IRC PING/PONG messages. We collected Sdbot traffic from simultaneously running five instances of Windows XP virtual machine hosts infected with this malware.

B Outlier Hosts

In the early stages of our analysis described in Section 5, we found that often TAMD failed to detect the malware-laden hosts, but rather identified other internal hosts as more symptomatic of malware, instead. Upon further inspection, we identified the internal hosts that resulted in these false alarms: PlanetLab nodes (`http://www.planet-lab.org`) and a Tor node (`http://tor.eff.org`).

In the case of PlanetLab nodes, we noticed that during the destination aggregation function, the vectors after PCA analysis often had very low dimensionality, e.g., two, where two principal components were able to cover over 90% of the data variance. Clustering these vectors resulted in a few outliers forming their own individual clusters, unlike any of the other vectors in Λ (i.e., the "new vectors"), or even those from Λ_{past} (the "old vectors"). This is shown in Figure 3. The two axes correspond to the top two principal components on which the original data is projected. The outliers were found to be PlanetLab nodes, which, being a development and testing platform, exhibit behavior deviating from other hosts. Their existence was also the reason why PCA analysis was able to reduce the vector dimensionality down to only two, since PlanetLab

[5] See `http://www.trendmicro.com/vinfo/virusencyclo`.
[6] See `http://www.symantec.com/enterprise/security_response/threatexplorer/threats.jsp`.
[7] See `http://www.microsoft.com/technet/security/Bulletin/MS04-044.mspx`.
[8] See `http://msdn2.microsoft.com/en-us/library/ms913275.aspx`.
[9] See `http://www.sophos.com/security/analyses/w32rbotxf.html`.
[10] See `http://www.symantec.com/enterprise/security_response/threatexplorer/threats.jsp`.

nodes' behavior is so different from other hosts that only two principal components were needed to capture most of the data variance.

In another example from experiments involving the Bagel trojan spyware, we noticed that even though TAMD was able to form a final aggregate containing all spyware traffic and spyware traffic only, at times it also combined another supposedly benign host into the spyware-hosts aggregate, both in the destination aggregation function and the payload aggregation function. Similar investigations revealed that this additional node is a Tor router inside the campus network. Tor offers online anonymity by routing packets over random routes between Tor servers so that the source and destination of the packet is obfuscated. Because the traffic comes from different anonymous hosts, it is possible that, even though the Tor router itself is not infected, another host routing traffic through the Tor node may be a spyware victim.

For the rest of this work, we removed the PlanetLab and Tor nodes from our analysis.

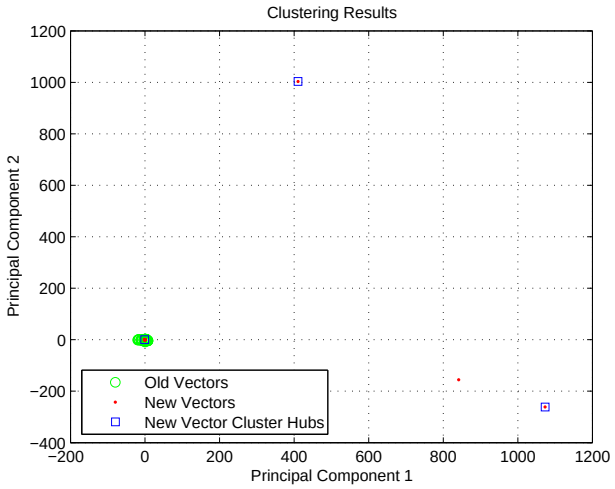

Figure 3: Clustering results after dimension reduction by PCA. The three outliers were found to be PlanetLab nodes.

C Performance

The top half of Table 2 shows the run times in seconds for each aggregation function and for each malware instance, averaged over the week's worth of traffic (in one-hour intervals) we used to performed our experiments. In our present implementation of TAMD, $\mathsf{ByDestination}^{\alpha,\tau,\rho}$ is implemented in Matlab, and $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}$ and ByPlatform are implemented in C. For the numbers reported in Table 2, $\mathsf{ByDestination}^{\alpha,\tau,\rho}$ was run on a PC with a Pentium IV 3.2 GHz processor and 3 GB of RAM, and $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}$ and ByPlatform were run on a Dell PowerEdge server with dual core 3 GHz processors and 4 GB of RAM.

The running times of the aggregation functions depend on several factors, including the number of external destinations identified as suspicious (i.e., SuspiciousSubnets as computed by $\mathsf{ByDestination}^{\alpha,\tau,\rho}$) and the number of "old" vectors from Λ_{past} that represent communication to one of those suspicious destinations; averages for these numbers are also listed in Table 2. The amount of traffic in Λ_{past} is especially critical to the performance of the destination and payload aggregation functions, $\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$ and $\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda^{\mathsf{sim}}_{\mathsf{past}}, \Lambda^{\mathsf{susp}}_{i})$, since both of them access significant amounts of historical data (i.e., Λ_{past} and $\Lambda^{\mathsf{sim}}_{\mathsf{past}}$) to define the "normal" behavior for this network. While the implementation of TAMD is not yet optimized, retrieving historical data from database contributed to the majority of the slowdown. This problem can be alleviated in the future by performing these calculations in advance and storing them statically, only updating incrementally as more data is collected.

		Bagle	IRCbot	Mybot	Sdbot
$\mathsf{ByDestination}^{\alpha,\tau,\rho}(\Lambda, \Lambda_{\mathsf{past}})$	μ	197.89s	598.52s	186.28s	186.87s
(line 101)	σ	412.46s	345.39s	393.59s	391.53s
$\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda^{\mathsf{sim}}_{\mathsf{past}}, \Lambda_i^{\mathsf{susp}}, \mathrm{false}) = 0$	μ	222.58s	90.87s	268.25s	96.34s
(line 106)	σ	572.68s	236.20s	635.96s	287.19s
$\mathsf{ByPayload}^{\delta_{\mathsf{Ed}}}(\Lambda_i^{\mathsf{susp}}, \Lambda_i^{\mathsf{susp}}, \mathrm{true}) > 0.2$, and	μ	1.25s	6.91s	1.53s	1.38s
$\mathsf{ByPlatform}(\Lambda_i) > 0.5$	σ	1.62s	2.95s	1.80s	1.79s
(line 107 and line 108)					
Total run time	μ	421.73s	306.78s	268.25s	284.60s
	σ	975.75s	678.47s	635.96s	672.74s
Size of $\mathsf{SuspiciousSubnets}$	μ	515.17	3652.00	474.52	475.47
	σ	309.32	456.39	309.41	309.43
Number of internal hosts contacting $\mathsf{SuspiciousSubnets}$					
in Λ	μ	982.94	1428.00	974.64	974.52
in Λ_{past}	μ	2273.58	4900.00	2165.12	2165.23

Table 2: Run time of each phase in seconds (s) of algorithm in Figure 1 and statistics of factors impacting performance (μ = mean, σ = std. dev.)