J Comput Virol (2010) 6:91–103
DOI 10.1007/s11416-009-0119-3

ORIGINAL PAPER

Auto-Sign: an automatic signature generator for high-speed malware filtering devices

Gil Tahan · Chanan Glezer · Yuval Elovici · Lior Rokach

Received: 15 June 2008 / Accepted: 12 January 2009 / Published online: 7 February 2009

Abstract This research proposes a novel automatic method (termed Auto-Sign) for extracting unique signatures of malware executables to be used by high-speed malware filtering devices based on deep-packet inspection and operating in real-time. Contrary to extant string and token-based signature generation methods, we implemented *Auto-Sign* an automatic signature generation method that can be used on large-size malware by disregarding signature candidates which appear in benign executables. Results from experimental evaluation of the proposed method suggest that picking a collection of executables which closely represents commonly used code, plays a key role in achieving highly specific signatures which yield low false positives.

1 Introduction

The time interval from the release of new malicious software to the wild till the time that the security software/hardware companies detects the malware, analyze it, generate a signature and release an update to its clients is highly critical. During this time interval the malware is undetectable by most of the signature-based commercial solutions and thus it can easily propagate [1]. For that reason, it is highly important to detect a new malware as soon as possible and rapidly generate a good signature so that abundant signature-based solutions can be updated.

One way to protect organizations from malware is to deploy high-speed malware filtering appliances on the communication lines that connect the organization to the outside world. Such appliances are based on deep-packet inspection in real-time and thus support very simple signatures for detecting malware. Security appliances are an appealing solution because they require no local overhead to manage them, yet they work continuously to protect the enterprise [2].

This research focuses on automating the process of generating signatures to be installed on such appliances for known malware that needs to be filtered by the appliances. Various techniques have been proposed to derive malware signatures automatically, including among others: vulnerability-based signatures [1]; payload-based signatures [3,4]; content sifting [5]; semantic-aware signatures [6]; The Amd algorithm [7]; Honeypot-based signatures [8–10], and polymorphic content-based signatures [2,11]. These studies examine code by matching and analyzing the distribution of string patterns in communication packets; classifying unsuccessful connections; and modeling invariant code structures. Such signatures were tested and reported to be effective for small-sized malware (usually worms) [1,3,4]. Nevertheless, the employed approaches for signature generation ignore the fact that many types of malware appear as full-fledged executable and therefore contain a significant portion of repetitive code emanating from code generators, development tools and platforms.

Considering the fact that signature-based systems do not account for such large common code segments, the quality of existing signature generation mechanisms is degraded. Such quality is measured using sensitivity (low false negative for

This work has been supported by Deutsche Telekom AG.

G. Tahan · C. Glezer (✉) · Y. Elovici · L. Rokach
Deutsche Telekom Laboratory at Ben Gurion University,
84105 Beersheba, Israel
e-mail: Chanan@bgu.ac.il

G. Tahan
e-mail: Gilta@bgu.ac.il

Y. Elovici
e-mail: Elovici@bgu.ac.il

L. Rokach
e-mail: liorrk@bgu.ac.il

malware) and specificity (low false positive for innocuous traffic) measures.

In order to address the problems stated above, this research proposes and evaluates a signature generation technique, termed *Auto-Sign*, for generating signatures which can be used by Network Intrusion Detection and Protection Systems (NIDS/NIPS) operating as malware filtering devices [12]. For such devices, Auto-Sign needs to generate a very simple signature that a network appliance can use for filtering malware in real-time. A very simple signature is actually a string of bytes or a simple regular expression of bytes at the most. To improve its precision, Auto-Sign employs an exhaustive and structured technique which first sanitizes malware from segments of common benign code and only then generates unique signatures which can be later used for detection of malware traffic.

The scope of this research is on tackling malicious code in the form of adware, spyware, Trojans, and viruses. Auto-Sign was tested also on large, full-fledged malicious applications and not necessarily on short stream-based malware where common code is not relevant.

Auto-Sign raises many questions with regards to various aspects of the proposed technique. In this research we were interested in finding the optimal length and selection criteria of a signature among several candidates in order to minimize false positives as well as the size and type of the training set in order to minimize false positives. This research describes the Auto-Sign technique and a set of experiments which were performed on a collection of malicious and benign executables. Auto-Sign is capable of automatic signature generation as part of the eDare framework [13] which offers "malware filtering as a service" and is targeted for Network Service Providers (NSP), Internet service Providers (ISP), small and large enterprises.

2 Related work on automatic signature generation (ASG)

Automated signature generation for new attacks of this type is extremely difficult due to the following reasons [10]. In order to create a malware signature, we must identify and isolate malicious traffic from benign traffic, which is not an easy task under all circumstances due to sophistication of hacking techniques. Signature generation for new attacks is also difficult since as soon as the signature generation methods are known to the attacker and especially the statistical ones, he or she may be able to defeat them by using statistical simulability as demonstrated in [14,15]. The signature must be general enough to capture all instances malicious traffic while at the same time specific enough to avoid overlapping with the content of benign traffic in order to reduce false positives. This problem has so far been handled in an ad-hoc way based on human judgment. As a case in point, current rule-based network Intrusion Detection Systems (IDS) can do little to stop zero-day worms [4]. They depend on content, protocol-anomaly and behavioral signatures which can only be generated in a delay after the malware has been launched and already created substantial damage.

Several approaches have been employed in order to expedite the process of signature generation for effective containment of worms. *Autograph* [3] stores source and destination addresses of each inbound unsuccessful TCP connection it observes. Once an external host has made unsuccessful connection attempts to more than s internal IP addresses, the flow classifier considers it to be a scanner. All successful connections from an IP address, flagged as a scanner, are classified as suspicious, and their inbound packets written to the suspicious flow pool. Autograph next selects the most frequent byte sequences across the flows in the suspicious flow pool as signatures. At the start of a worm's propagation, the aggregate rate at which all infected hosts scan the IP address space is quite low. Because Autograph relies on overhearing unsuccessful scans to identify suspicious source IP addresses, early in an epidemic an Autograph monitor will be slow to accumulate suspicious addresses, and in turn slow to accumulate worm payloads. To address this problem Autograph uses a tattler that, as its name suggests, shares suspicious source addresses among all monitors, toward the goal of accelerating the accumulation of worm payloads.

Honeycomb [8] tries to spot patterns in traffic previously seen on the honeypot. Honeycomb overlays parts of flows in the traffic and use a *Longest Common Substring* (LCS) [16] algorithm to spot similarities in packet payloads. Tang and Chen [10] followed-up this work by designing a double-honeypot system, deployed in a local network for automatic detection of worm attacks from the Internet. Two algorithms based on Expectation-Maximization [17] and Gibbs sampling [18] are proposed for efficient computation of Position Aware Distribution Signature (PADS). The *PAYL sensor* [4] employ anomaly detection which is based on the principle that "zero-day" attacks are delivered in packets whose data is unusual and distinct from all prior "normal content" flowing to or from the victim's site. The *Nemean architecture* [6] is a semantic-aware Network Intrusion Detection System (NIDS) which contains two components: a data abstraction component that normalizes packets from individual sessions and renders semantic context, and a signature generation component that clusters similar sessions and uses machine-learning techniques to generate signatures for each cluster. In a related study, the *Amd* algorithm generates semantic-aware code templates and specifies the conditions for a match between the templates and the programs being checked [7]. *Polygraph* [19] provides a content-based signature generation techniques for polymorphic worms. The underlying assumption is that possible to automatically generate

signatures that match many variants of polymorphic worms offering low false positives and low false negatives. Newsome et al. [19] propose and evaluate a system that expands to notion of single substring signatures (tokens) to conjuctions, ordered sets of multiple tokens and Bayesian (score) tokens. *EarlyBird* [5] sifts through the invariant portion of a worm's content will appear frequently on the network as it spreads or attempts to spread. In *Netspy* [4] the invariant portion of network traffic generated by a spyware program is used to derive a spyware signature. This is because a signature that has content related to specific user input will miss network activity generated by the program on other user input. Netspy uses a variant of the longest common subsequence (LCSeq) algorithm [20] to find such invariants sections.

Filiol [11] address the problem that commercially available anti-viruses are not resistant against black-box analysis. He suggested generating multiple sub-signatures that are randomly selected from a longer signature. Sub signatures are distributed such that "any sub-pattern is a fixed value which depends on the user/computer identification data".

The aforementioned automatic signature generation techniques focus on analyzing similarities and anomalies in executables' substrings, code tokens as well as statistical distribution of code across variants of malware. Consequently, testing of such signatures was performed on short, stream-based malware such as Nimda, Code Red/Code Red II, MS Blaster (1.8 KB long), Sober, Netsky and B[e]agle. Nevertheless, larger malware executable files, carrying full-fledged applications usually contain a significant portion of common code segments which are planted by software development platforms spawning the malware but are found also in benign executables. As a result, automatically selecting a signature that will be both sensitive and specific is a very challenging task in the case of these large files.

The goal of this research is to generalize the above work by proposing an automatic signature generation technique, termed Auto-Sign, capable of generating highly sensitive and highly specific signatures for malware of any size and type operating in any operating system environment (i.e., Trojan horses, spyware, adware, viruses, and worms). The technique is also capable of handling malware such as self-decrypted/self-executed files; or archive files (i.e., CAB, MSI, Zip). Of course a signature derived for unencrypted/uncompressed malware cannot be used for detecting the same malware in encrypted or compressed files. Short, stream-based malware (not a self-contained application) which does not include significant portions of common code is also not a typical candidate for Auto-sign.

3 The proposed automatic signature generation method

In order to create and employ signature for effective and efficient detection of malware in executables, our technique should generate a signature which complies with several requirements. First, we are interesting to find a signature σ of length s with a low probability p_0 to appear in a benign file. In classical signature-based detection, the number of appearances of a *contiguous* signature σ in a benign file of n bytes is distributed as:[1]

$$Z \sim N\left((n-2)p_0, \sqrt{(n-2)p_0(1-p_0)}\right) \quad (1)$$

The false positive rate of such detector is directly determined by p_0. Thus lowering the value of p_0 will result in a lower false positive. In order to apply with a certain level of false positive rate, one should require that the p_0 will not exceed a certain value e.

However estimating p_0 is not an easy task. Assuming we are using a relatively small sample of benign files, then simply measuring the frequency of signature σ in the sample might be unreliable. As proposed in [11], we can use a first-order Markov model to estimate the probability p_0 of a signature σ containing s bytes as:

$$\hat{p}_0(\sigma) = \Pr(b_1) \cdot \Pr(b_2|b_1) \cdot \Pr(b_3|b_2) \cdots \Pr(b_s|b_{s-1})$$

where b_i is the ith bytes in σ.

Let $f^*(x)$ represents a function that returns a signature of length x that appears in the malware code and has lowest $\hat{p}_0$. However if $f^*(s)$ is not sufficient, i.e. $\hat{p}_0(f^*(s)) < e$, then we are compelled to search for a larger signature $f^*(s+1)$. Note that $\hat{p}_0(f^*(x))$ is monotonically decreasing in x because there are only two cases:

1. If $f^*(x) \subset f^*(x+1)$ then $\hat{p}_0(f^*(x+1)) = \hat{p}_0(f^*(x)) \cdot \Pr(b_{x+1}|b_x) \leq \hat{p}_0(f^*(x))$ (which satisfies monotony)
2. If $f^*(x) \not\subset f^*(x+1)$ then $\hat{p}_0(f^*(x+1)) \leq \hat{p}_0(f^*(x))$ because by definition of $f^*\hat{p}_0(f^*(x+1)) \leq \hat{p}_0(f^*(x)\&b)$ where & indicates the concatenation operation.

Other requirements from the signature are: the signature should be sufficiently short to decrease the problems of Internet packet splits and detection hardware storage limitations (i.e., various IDS/IPS devices such as DefensePro [21]); Third, the signature should comply to the limitations of high-speed deep packet inspection devices that can detect and remove malware in real-time in high-speed data stream. Finally, it should be well-defined to enable totally automatic generation. The major challenge in conforming to the aforementioned requirements was to develop a methodology that can locate code segment or segments highly unique to a given malware instance and can therefore serve as a powerful and unique signature meeting the stringent requirement set forth by commercial high-speed malware filtering devices.

[1] This is slightly different from the probabilistic law presented in [14] because here we are not referring to the general case of which the signature bytes are not necessary contiguous.

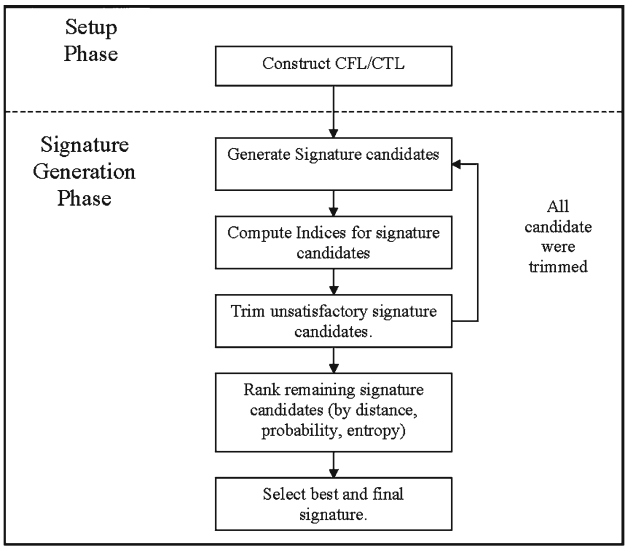

Fig. 1 The Auto-Sign methodology

Since many malware executables nowadays are in fact developed using 4th Generation development platforms (e.g. MSVC, J2EE, Delphi), the binary code of malware nowadays contains a large portion of code placed by these development platform. These portions of common code are identical or almost identical except address references. Such common segments were not developed by the authoring hacker, but were linked to the malware as part of the underlying code generator's library and are termed Common Function Code (CFC). To significantly decrease the risk of selecting such abundant segments as a signature that may lead to high false positive rate, we must first identify and disregard the CFC part. The CFC of a malware file can be identified by analyzing the malware content against a repository of CFC which is termed a Common Function Library (CFL). The CFL can be derived based on a collection of benign executables and should be regularly updated in order to take into account the evolution of benign (and potentially malicious) files.

To meet the requirements stated above, we developed the Auto-Sign methodology which is schematically described in Fig. 1.

3.1 Setup

During the setup phase a data structure (library) representing a collection of benign executables is constructed. The data structure is termed CFL.

The CFL is comprises of the following data structures.

1. **3-gram-frequency:** A vector of 2^{24} ($\sim$16 million) entries where each cell represents the number of occurrences of each 3-gram (three consecutive bytes) in the collection of benign files. The n-gram size was chosen to be 3 in order to comply with memory constraints. N-Gram was also used in [22] for malware detection and in [23,24] for malware phylogeny. The collection of benign files used for constructing the CFL is scanned sequentially in order to record the number of occurrences of each 3-gram.
2. **3-gram-Files-association:** A $2^{24} \times 64$ bit-map where a '1' Binary value in a cell (i, j) indicates the appearance of a specific 3-gram i in the jth group of files. The CFL files are divided into 64 groups.
3. **3-gram-relative-position-within-file:** A $2^{24} \times 64$ bit-map where a '1' Binary value in a cell (i, j) indicates the appearance of 3-gram i in the jth internal segment of a file (assuming the file is divided into 64 equal length segments).

The common threat library termed CTL can be constructed as a separate data structure in a similar manner.

Maintaining lookup tables which capture the appearance of each possible 3-gram in a file (or group of files for large CTL/CFLs), as well as the relative position of 3-grams in each file promotes the scalability of Auto-Sign. This enables handling large repositories of CFL/CTL files representing many heterogeneous software platforms, when other techniques such as LcSeq [20] are not feasible anymore.

3.2 Signature generation

The aim of this process is to auto generate an efficient signature. The process is repeated for each malware M. First, a set of signature candidates is extracted from different positions within the malware M. Each signature candidate C_M looks for a fixed contiguous sequence of bytes $C_M \in \{0, 1, 2, \ldots, 255\}^s$ with $s = |C_M|$. Using our 3-grams representation C_M comprises a number of 3-grams depending on the length of the signature (e.g. a signature of length 4 bytes is comprised from two 3-grams that overlap by two bytes). Employing the 3 data structures prepared during the setup stage (3.1), the following indices are calculated for each signature candidate extracted from the malware.

Spread

In Sect. 3.1 we have divided each file into 64 segments. The "Spread" measure represents the spread of the signature's 3-grams along the various segments for all the files in the CFL. For example, Spread = 1 indicates that the signature is located in only one segment in all the files of the CFL.

The match of a certain 3-gram k in a benign subset S of three contiguous bytes may be described as a Bernoulli

variable:

$$X_k^S = \begin{cases} 1 & p_k^{\text{gram}} \\ 0 & 1 - p_k^{\text{gram}} \end{cases}$$

The value p_k^{gram} is determined by the efficiency of the k's 3-gram.

We define the following variable with respect to a segment j:

$$Y_j = \begin{cases} 1, & \text{iff } \forall 3\text{gram} \subset C_M, \\ & \quad \forall \text{file} \in \text{CFL} : 3\text{gram} \subset \text{segment}_j(\text{file}) \\ 0, & \text{else} \end{cases} \quad (2)$$

where segment$_j$ (file) denote the jth internal segment of file.

Let $p_{j,k,i}$ represent the probability of k's 3 gram to appear at least one time in segment j of a benign file i. $p_{j,k,i}$ can be estimated the complementary probability, i.e. that the probability of the event that k is not been found in the segment:

$$p_{j,k,i} = 1 - \left(1 - p_k^{\text{gram}}\right)^{|\text{segment}_j(i)|-2} \quad (3)$$

Thus Yj is distributed as Bernoulli with success probability of

$$p_j^{\text{seg}} = \prod_{k \subset C_M} \prod_{i \in \text{CFL}} p_{j,k,i} \quad (4)$$

Note that because the segments are equally sized then $p_j^{\text{seg}} = p_i^{\text{seg}}; \quad \forall i \neq j$

The following formula specifies the spread measure:

$$\text{Spread}_{\text{CFL}}(C_M) = \sum_{j=1}^{64} Y_j \quad (5)$$

Spread is distributed as the sum of 64 identically distributed Bernoulli random variables which can be approximated to Normal distribution $N(64 \cdot p^{\text{seg}}, \sqrt{64 \cdot p^{\text{seg}} \cdot (1 - p^{\text{seg}})})$.

Freq

The average frequency of all 3-grams that comprise a candidate signature (computed using the 3-gram-frequency data structure). Higher "Freq" may indicate a bad signature candidate. The following formula specifies the probability measure:

$$\text{Freq}_{\text{CFL}}(C_M) = \frac{\sum_{\forall k \subset C_M; |S|=3} \sum_{\forall \text{file} \in \text{CFL}} \sum_{\forall S \subset \text{file}; |S|=3} X_k^S}{|\text{CFL}|(|C_M|-2)} \quad (6)$$

Note that $|C_M| - 2$ represent the number of 3 gram in the signature C_M.

Freq is a sum of distributed Bernoulli random variables. Note that the central limit theorem can not be applied here directly as the X's are not identically distributed (different success probabilities). Thus we are using Poisson approximation as proposed in [25]:

$$\Pr(\text{Freq} = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \quad (7)$$

where

$$\lambda = \frac{\sum_{\forall k \subset C_M; |S|=3} \sum_{\forall \text{file} \in \text{CFL}} |\text{file} - 2| \cdot p_k^{\text{gram}}}{(|C_M| - 2)} \quad (8)$$

The cumulative distribution function (CDF) is:

$$\Pr(\text{Freq} \leq k) = \frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor} \quad (9)$$

where $\Gamma(x, y)$ is the Incomplete gamma function and $\lfloor k \rfloor$ is the floor function.

R = Freq/Spread

By dividing the aforementioned freq and spread we are able to further increase and normalize the crude probability metric for a candidate signature. For example, when all the comprising 3-grams of the candidate signature are concentrated in the same area within an executable (most likely indicating an area of common code) the normalized score will yield a higher value compared to a situation (in a different executable) where the 3-grams are found spread in numerous relative positions within the executable.

We hypothesize that those m candidates are less likely to appear within CFL files in consistent areas of a file thus less likely to belong to chunks common code. This hypothesis indicates that low ratio candidate signatures are associated with lower signature probabilities in benign files (signature *efficiency*).

Proposition 1 mathematically examines this hypothesis under limited circumstances. In Sect. 4.2 we empirically examine this hypothesis in general circumstances.

Proposition 1 *Let R_1 and R_2 represent the ratios of two candidate signatures C_1 and C_2 both of length $s = 3$ test on single file. If $P_0(C_1) < P_0(C_2)$ then $E(R_1) < E(R_2)$.*

Proof R_1 and R_2 are ratio of two Poisson variables. Thus

$$E(R_1) = E\left(\frac{\text{Probability}_1}{\text{Spread}_1}\right) = E(\text{Probability}_1) \cdot E\left(\frac{1}{\text{Spread}_1}\right)$$

Using the approximation of the mean of the inverse of a Normal distribution [26]

$$E(R_1) \approx \lambda_1 \cdot \frac{1}{\mu_1}$$

According to the theorem condition (s = 3, |CFL| = 1) and using Eq. (8) we obtain:

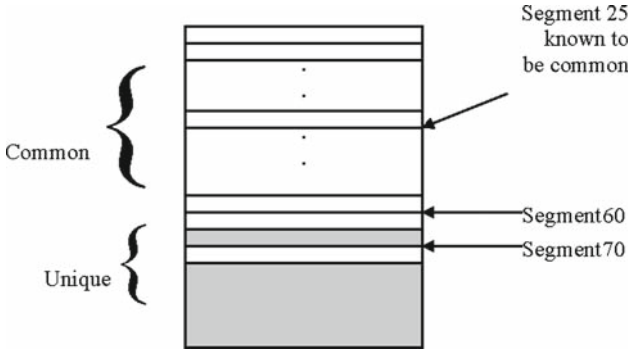

Fig. 2 Illustrating the motivation of the distance estimator

$$\lambda_1 = |\text{file} - 2| \cdot p_0(C_1)$$
$$\mu_1 = 64\left(1-(1-p_0(C_1))^{|\text{file}|/64-2}\right)$$

i.e.,

$$E(R_1) = \frac{|\text{file} - 2| \cdot p_0(C_1)}{64\left(1-\left(1-p_0(C_1)\right)^{|\text{file}|/64-2}\right)}$$

We assume the file is large enough (i.e. larger than 64×4 which is reasonable assumptions as usually files contains thousands of bytes). In this case moving from $P_0(C_1)$ to $P_0(C_2)$, i.e. the probability is increased $\frac{P_0(C_2)}{P_0(C_1)}$ times will result in increasing the nominator by $\frac{P_0(C_2)}{P_0(C_1)}$ times but the denominator to be increased in less than $\frac{P_0(C_2)}{P_0(C_1)}$ times. Thus the entire rate is increased as Proposition 1 argues.

Distance

We examine the distance of each candidate signature that does not appear in the common library to its nearest signature that does appear in the common library.

The structure of executable is built from a continuous partition of common library functions and then a unique code partition. For the sake of clarity we assume that there is only one partition of each type as illustrated in Fig. 2. Each line represents a segment of 64 bytes which is also the length of the candidates that we examine. Let assume for the example that we know that the 25th segment is part of the common library (by looking into many other files). However we are not aware where the unique partition begins. However segments that are located in higher places (i.e. are located far way from the 25th segment) have larger chance to be part of the unique block (and because of that can be good candidate for a signature). Specifically in Fig. 2, the 70th segment is more likely to be part of the unique partition than the 60th segment. We formulate the last intuition in the following proposition.

Proposition 2 *Assuming the executable consists of two contiguous partitions of common and unique code, then the expected squared distance of an unique segment to the closet known common segment is grater than the expected squared distance of a common segment to the closet known common segment.*

Proof Without loss of generality the common code is located in the first n segments and the unique code is located in the remaining m segments. A unique segment located in position b ($b > n$) and a common segment is located in position a ($a \leq n$).

The location of the closet known common segment is distributed according to *some unknown*discrete distribution:

$$\Pr(\text{location} = i) = \beta_i; \sum_{i=1}^{n} \beta_i = 1 :$$

The expected squared distance from location to a and b are:

$$\begin{aligned}\text{SD}_{\text{common}} &= E\left[(\text{location} - a)^2\right] = E\left[\text{location}^2\right] \\ &\quad -2aE\left[\text{location}\right] + a^2 \\ \text{SD}_{\text{unique}} &= E\left[(\text{location} - b)^2\right] = E\left[\text{location}^2\right] \\ &\quad -2bE\left[\text{location}\right] + b^2\end{aligned}$$

Thus:

$$\text{SD}_{\text{unique}} - \text{SD}_{\text{common}} = (b-a)\left(E\left[\text{location}\right] + b + a\right) > 0$$

Entropy

In addition the new estimators presented above, we also use entropy measure which has known to be useful for selecting effective signatures [11].

Bits

the total number of files (or file groups) that contain the signature.

Measures 1–6 were calculated for the CFL. Measures 2 and 6 were calculated for the CTL.

The signature generation process can therefore be summarized as follows:

1. **Generate signature candidates** of length L by splitting the examined malware executable to segments of equal length L.
2. **Calculate the following measures** (for each signature candidate):
   Spread (based on CFL), Freq/Spread (based on CFL), Bits (Based on CFL), Entropy (you should define it with all others), Freq (based on CTL), Bits (based on CTL). This stage is done using the data structures created in the setup stage.

3. **Mark each candidate appearing in the CFL (bits > 0):** Only signatures that do not have any incident where all their comprising 3-grams appear in a file within the CFL repository are considered viable candidates.
4. **Trim candidates** appearing in the CFL (CFL-bits > 0) or appearing in the CTL. Note that we trim candidates that appear in the CTL because common functions of malware are also not good candidates for identifying a certain malware.
5. **Rank remaining candidates (marked during stage 3) by distance from CFL:** The n candidates with the highest physical distance from the CFL/CTL areas (as calculated from the 3-gram-relative-position-within-file matrix) are selected. Our assumption is that a high distance of a signature candidate from the closest chunk of common code within the executable indicates an area of code which represents the malware more uniquely.
6. **Rank remaining candidates using CFL–Freq/CFL–Spread:** Out of the n candidates selected in phase 5, m candidates with the lowest CFL–Freq/CFL–Spread ratios are selected.
7. **Rank remaining candidates based on entropy:** Out of the m candidates selected in phase 6, the w signature candidates with the highest degree of entropy are ranked as best and final. It is customary to assume that a signature candidate with a higher degree of entropy is less likely to be associated with areas common to different executables (i.e., constants, text with repeating characters) and is therefore a more unique identifier of the malware.
8. **Select actual signature** Out of the w remaining candidates the one with the highest entropy is selected as the actual signature.

Table 1 provides a pseudo-code specification of the aforementioned signature selection process.

Proposition 3 *The computational complexity of the algorithm in Table 1 is* $O(CR \cdot (CL - 2))$ *where CR denote the number of requested candidates and CL is the candidate length.*

Proof The computational complexity of the algorithm in Table 1 is computed as follows: the GenerateSignatureCandidates complexity is $O(CR)$, the complexity of the first loop (lines 3–13) is $O(CR \cdot (CL - 2))$, the complexity of the second loop (lines 15–19) is $O(CR)$, the complexity of the third loop (lines 21–26) is $O(CR)$, the complexity of Candidates.SortBY_DistanceFromCFL is $CR \cdot \log(CR)$, the complexity of the fourth loop on lines 29–30 is $O(CR)$ (worth case when $N = 0$), the complexity of the fifth and sixth loop in lines 32–34 and 37–39 is also $O(CR)$, and the complexity of Candidates.SortBY_Entropy is $CR \cdot \log(CR)$. Thus, the overall complexity is $O(CR \cdot (CL - 2))$.

4 Evaluation

4.1 Examining the effectiveness of Auto-Sign

The evaluation described in this section comprises of a set of experiments we conducted to test the effectiveness of Auto-Sign. The experimental design was aimed at assessing the impact of various independent variables on the quality of the signature which is characterized by the number of false positive appearances of a signature in a set of clean (benign) files. The independent variables used by this study are the proportion of CFL versus test files and the size of a signature (32, 64, 128 bytes).

The first task in the evaluation (Fig. 3) is to assemble two repositories: one comprising benign and the other comprising malicious (threat) executables. The benign repository is randomly split ten times into training- and testing-sets. A distinct CFL is constructed ten times for each of these sampled training sets where the size of the training set is increased in a linear fashion during each of these ten iterations. The threat repository, on the other hand, is held fixed during the evaluation and therefore the CTL is constructed once before being used by the signature generation algorithm.

In each such configuration we generated three signatures with lengths of 32, 64, 128 bytes for all 849 malware files. The false positive count of the best signatures generated by Auto-sign was calculated by performing a cross-validation of the selected signature against the test set.

The repository of benign files included 5494 files with lengths ranging from 3 KB to 8 MB. The repository of malware instances comprised of 849 files with lengths ranging from 6 KB to 4.25 MB (11 executables where above 1 MB and 200 above 300 KB). The distribution of malware file types is depicted in Fig. 4. Generating the signatures for the malware repository took 5.7 s using one 3 GHz Pentium processor. Figure 5 depict a sample of Auto-Sign's output with a list of files, their signatures and the indices calculated for these signatures.

Table 2 depicts the average false positive counts for monotonically increasing proportions of the CFL in the following two configurations: 33% training set (CFL with CTL)/67% testing set; and 50% training (CFL with CTL)/50% testing set. Table 3 compares the average false positive counts for 25% training–75% testing set once with the CTL and once without. For each such configuration we used a 10 cross-fold validation over the 849 files for the 3 signature lengths.

The results indicate the fact that even a large CFL of up to 50% of the clean files repository cannot compensate for short

Table 1 Pseudo-code of signature selection process

```
1  Candidates=GenerateSignatureCandidates(Maleware,CandidatesRequired,CandLen);
2
3  ForEach Candidate in Candidates do
4        {
5        Candidate.CFLspread   = CFL.CalcSpread (Candidate);
6        Candidate.CFLNormProb = CFL.Average3gramProb (Candidate) / Candidate.CFLspread;
7        Candidate.CFL_bits    = CFL.Count_Files_With_All_3gram (Candidate)
8        Candidate.Entropy     = Entropy(Candidate.string);
9        Candidate.InCFL       = (Candidate.CFL_bits > 0);
10
11       Candidate.CTL_bits    = CTL.Count_Files_With_All_3gram (Candidate)
12       Candidate.InCTL       = (Candidate.CTL_bits > T);
13       }
14
15 ForEach Candidate in Candidates do
16       {
17       If Candidate.InCFL then continue; // disregard
18       Candidate.DistanceFromCFL = DistanceFrom(Candidate.offset, Candidates);
19       }
20
21 ForEach Candidate in Candidates do
22       {
23       If Candidate.InCFL or
24         Candidate.InCTL then
25         Candidates.delete(Candidate);
26       }
27
28 Candidates.SortBY_DistanceFromCFL();
29 For i=Candidates.count downto N+1 do
30       Candidates.delete(i);
31
32 Candidates.SortBY_CFLNormProb();
33 For i=Candidates.count downto M+1 do
34       Candidates.delete(i);
35
36 Candidates.SortBY_Entropy();
37 For i=Candidates.count downto 2 do
38       Candidates.delete(i);
39 Signature = Candidates[1];
```

signatures of 32 bytes. The size of the signature is optimal at 64 bytes as the improvement from 64 to 128 bytes is not substantial. Moreover, the important factor with regards to the CFL selection is the "proximity" of the common code to the code generated by coding platforms. This is evident from the fact that in the case of the 25–75%, removing the

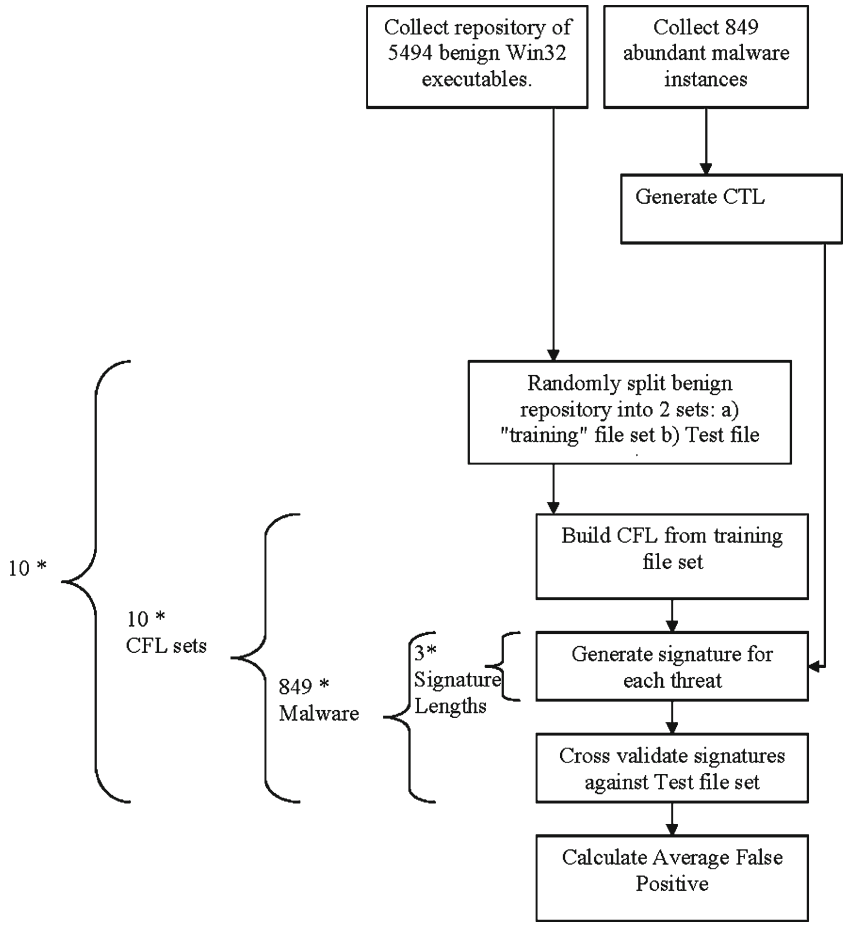

Fig. 3 The Auto-Sign evaluation plan

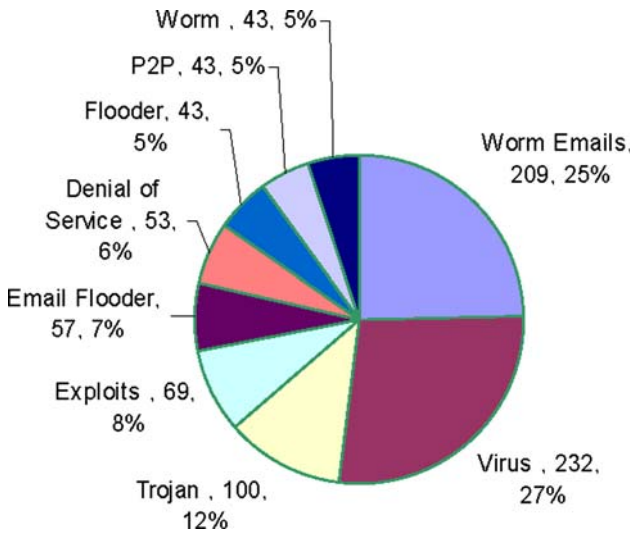

Fig. 4 Distribution of Malware Types

files of the CTL hampers the precision of the signatures and increases the amount of false positives. Moreover, the results also indicate that there is no point increasing the CFL beyond some optimal threshold. Once the CFL represents a critical mass of the common code (in the case of choosing a 25% CFL randomly, approximately 7.5% of the clean repository), adding files to the CFL does not yield any marginal decrease in the false positive rates of candidates.

Finally, we also generated signatures for the malware using random selection and entropy-maximization for the various signature lengths. Under the entropy-maximization approach, the signature was picked from an area within the malware which has the highest entropy score. This was done in order to minimize the likelihood of the signature's appearance in benign executables which would yield frequent false-positive detections.

Table 4, depicts the number of malware files whose signature was found in the benign repository of 5,494 files. The results indicate that the entropy maximization technique is superior to random selection in all signature lengths however both techniques are far from being feasible for meeting the quality of the CFL and especially CFL + CTL performance.

4.2 Are the estimators' good indicators for the signature efficiency?

Our method filters the signature candidate list by employing different estimators in a cascade manner. In order to obtain

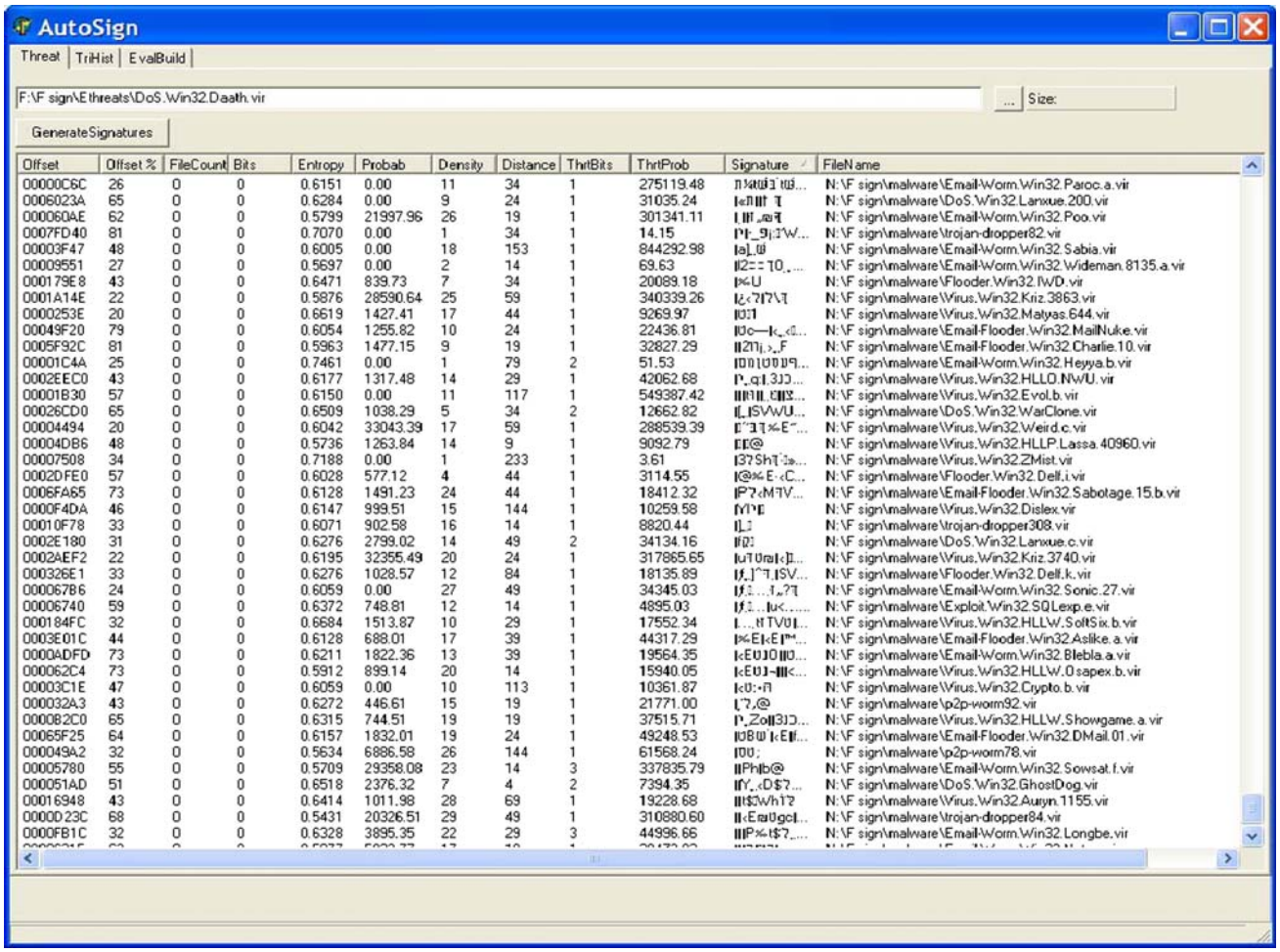

Fig. 5 Sample of signature data with indices

Table 2 False positive counts as a function of the CFL size

	50–50% with CTL				33–67% with CTL		
	32 byte	64 byte	128 byte		32 byte	64 byte	128 byte
CFL = 5%	32	0	0	CFL = 3.3%	54	9	10
CFL = 10%	28	1	0	CFL = 6.6%	31	1	1
CFL = 15%	19	0	0	CFL = 9.9%	25	1	0
CFL = 20%	17	0	0	CFL = 13.2%	30	0	0
CFL = 25%	21	0	0	CFL = 16.5%	25	0	0
CFL = 30%	20	0	0	CFL = 19.8%	22	0	0
CFL = 35%	22	0	0	CFL = 23.1%	22	0	0
CFL = 40%	26	0	0	CFL = 26.4%	19	0	0
CFL = 45%	23	0	0	CFL = 29.7%	17	0	0
CFL = 50%	25	0	0	CFL = 33%	21	0	0

a good signature (i.e. which does not appear in the benign file), the estimators need to be an informative discriminator whether the signature might appear or not in a benign file. In this section we examine how indicative the proposed estimators are.

For this purpose we performed the following experiment. We randomly sampled 2,000 64 byte signatures from various malware instances and calculated their estimators. Then, we searched for the signatures in a benign corpus to verify whether they appear in benign files.

Table 3 False positive counts as function of the CFL size—with/without CTL

	25–75% with CTL				25–75% without CTL		
	32 byte	64 byte	128 byte		32 byte	64 byte	128 byte
CFL = 2.5%	43	10	7	CFL = 2.5%	58	20	10
CFL = 5%	43	6	5	CFL = 5%	55	14	10
CFL = 7.5%	46	10	9	CFL = 7.5%	52	11	10
CFL = 10%	34	0	0	CFL = 10%	39	2	2
CFL = 12.5%	25	0	0	CFL = 12.5%	36	1	1
CFL = 15%	24	0	0	CFL = 15%	32	2	1
CFL = 17.5%	35	0	0	CFL = 17.5%	40	2	1
CFL = 20%	35	0	0	CFL = 20%	38	1	1
CFL = 22.5%	25	0	0	CFL = 22.5%	36	3	2
CFL = 25%	27	0	0	CFL = 25%	32	2	1

Table 4 False positive counts: random versus entropy-maximization

Random			Entropy		
32 bytes	64 bytes	128 bytes	32 bytes	64 bytes	128 bytes
297	249	227	197	124	91

Table 5 summarizes the t-test results of two-samples (assuming unequal variances) which compare the estimators values in a case that the signature appeared in the benign corpus or not. The results are very encouraging. For all estimators the null hypothesis are rejected with a significance level of $\alpha = 5\%$. Thus, the proposed estimators are good indicators for predicting if a signature would appear in benign files.

4.3 Are the estimators' statistically independence?

In addition to the fact that estimators should be indicative in selecting the best signature, we would like that the estimators will be diverse in the way they rank signatures. Otherwise, there is no point in using multiple estimators. This requirement is similar to the diversity requirement in mixture-of-experts techniques in AI [27], and specifically in malware detection [28,29]. These methods are very effective, mainly due to the phenomenon that various types of models have different "inductive biases". Diversity has therefore been applied to increase the predictive performance of classifier systems. In order to examine the diversity of the estimators, we used the data described in the previous section and checked whether the estimators were statistically independent using chi-square test (Table 6).

Unfortunately, our results indicate that the estimators are statistically dependent with $\alpha = 5\%$. However, Auto-Sign does not simply weigh these estimators. Instead, it uses it in a cascade fashion: first, it employs a distance estimator, then a ratio estimator, and finally an entropy estimator. Thus, it is more reasonable to examine the statistics dependence which reflects this cascading procedure. Assuming that each estimator filters half of the candidates to be used by the following cascade, we examine the statistical dependence between estimators' values for the chosen (unfiltered) candidates and the corresponding values of the subsequent estimator. The results presented in Table 7, indicate that the estimators are statistically independent with $\alpha = 5\%$ when they are applied in cascade.

Table 5 Results of *t*-test of two-sample assuming unequal variances

Estimator	Mean value of the estimator for signatures that appeared in benign corpus	Mean value of the estimator for signatures that did not appear in benign corpus	*p* value on two-tail	Conclusion
Entropy	0.55 ± 0.008833	0.32 ± 0.075756	4.05E-37	Significant
Ratio	1493.06 ± 2441	3728.99 ± 3306	1.7E-25	Significant
Distance	13.28 ± 24.22	$0.586 + 3.963629$	3.62E-82	Significant
Number of observations	1695	305		

Table 6 Results of chi-square test for independent

Test	p value	Conclusion
Distance versus entropy	0.037	Statistically dependent
Ratio versus entropy	1.53636E-52	Statistically dependent
Ratio versus distance	0.005	Statistically dependent

Table 7 Results of chi-square test for independence in cascade manner

Test	p value	Conclusion
Distance versus ratio	0.363	Statistically independent
Ratio versus entropy	0.070	Statistically independent

5 Conclusion

This paper proposes a new approach for automatic generation of signatures for malware executable of all sizes with an intention to be used by high-speed malware filtering devices. We consider the fact that large executables are comprised of substantial amounts of code that originates from the underlying standard development platforms and is thus replicated across various instances of both benign and malware developed by these platforms. In order to minimize the risk of false positive classification of benign executables as malware, we propose and evaluate a method to discard signature candidates that contain such replicated chunks of code.

The empirical findings presented in Sect. 4 support the viability of the general approach proposed by this research and suggest that eliminating signature candidates belonging to common code segments has a more profound effect on restricting the level of false positives than increasing the length of a signature. N-grams, however, are not the only measure which can be used to realize the general approach posed by Auto-Sign and alternative ones can be used as a substitute for representing signatures (i.e., segment prefix/suffix, hash values, etc.)

The main benefit of the proposed method is that it enables analysis at the binary level and does not require a semantic interpretation of code into function blocks using techniques such as code markers, disassembly, state-machines, etc. This benefit means that the methodology is generic and is not affected by changes in CPU or introduction of new development platforms.

Nevertheless, enterprises which would like realize Auto-Sign in generating signatures for high-throughput network security appliances need to follow a more exhaustive and systematic methodology for building their CFL repository. Considering the global variety of development platforms and the mobility of threats facilitated by the Internet, ensuring the external validity of this study relies substantially on reaching a critical mass of CFL files which represents abundant development platforms. Furthermore, it often does not suffice for a signature to be available—deployed signatures must be managed, distributed and kept up-to-date by security administrators [30].

We plan to repeat the evaluation Auto-Sign on a larger scale with much more malware files and CFLs generated for different development environments. We also plan to evaluate additional methods for trimming, ranking and choosing the best signature out of the collection of candidates. In addition, in order to further strengthen the signatures and minimize the risks of false positives we propose to use "composite signatures" which are generated by using two or more distinct signatures for each malware. This activity addresses the biggest challenge of Auto-Sign which is the need to reduce to zero the amount of false positives before being deployed for generating signatures in high-speed malware filtering devices. In the future we plan to use Auto-Sign to generate multiple signatures in order to increase the resistance against black box analysis as described in [11].

Acknowledgments The authors gratefully thank the action editors and the anonymous reviewers whose constructive comments significantly helped in improving the quality and accuracy of this paper.

References

1. Brumley, D., Newsome, J., Song, D., Wang, H., Jha, S.: Towards automatic generation of vulnerability-based signatures. In: Proceedings of the 2006 IEEE Symposium on Security and Privacy (2006)
2. Szor, P.: The Art of Computer Virus Research and Defense, Addison–Wesley, Reading (2005)
3. Kim, H.A., Karp, B.: Autograph: Toward automated, distributed worm Signature detection. In: Proceedings of the 13th Usenix Security Symposium (Security 2004), San Diego, CA, August (2004)
4. Wang, K., Stolfo, S.J.: Anomalous payload-based network intrusion detection. In: Recent Advance in Intrusion Detection (RAID), September (2004)
5. Singh, S., Eitan, C., Varghese, G., Savage, S.: Automated worm fingerprinting. In: 6th Symposium on Operating Systems Design and Implementation (OSDI), December (2004)
6. Yegneswaran, V., Giffin, J.T., Barford, P., Jha, S.: An architecture for generating semantics-aware signatures. In: 14th USENIX Security Symposium. Baltimore, Maryland, August (2005)
7. Christodorescu, M., Jha, S., Seshia, S., Song, D., Bryant, R.E.: Semantics-aware malware detection. In: IEEE Symposium on Security and Privacy. Oakland, California, May (2005)
8. Kreibich, C., Crowcroft, J.: Honeycomb: creating intrusion detection signatures using honeypots. SIGCOMM Comput. Commun. Rev. **34**(1), 51–56 (2004)
9. Provos, N.: A virtual honeypot framework. CITI Technical Report 03-1, Center for Information Technology Integration, University of Michigan, Ann Arbor, Michigan, USA, October (2003)
10. Tang, Y., Chen, S.: Defending against Internet worms: a signature-based approach. In: Proceedings of IEEE INFOCOM'05, Miami, Florida, USA, May (2005)
11. Filiol, E.: Malware pattern scanning schemes secure against black-box analysis. J. Comput. Virol. **2**(1), 35–50 (2006)

12. Morin, B., Mé, L.: Intrusion detection and virology: an analysis of differences, similarities and complementariness. J. Comput. Virol. **3**(1), 39–49 (2007)
13. Elovici, Y., Shabtai, A., Moskovitch, R., Tahan, G., Glezer, C.: Applying Machine Learning Techniques for Detection of Malicious Code in Network Traffic. In: The 30th Annual German Conference on Artificial Intelligence (KI-2007), Lecture Notes in Computer Science, vol. 4667, pp. 44–50. Springer, Osnabrück (2007)
14. Filiol, E., Josse, S.: A statistical model for viral detection undecidability. J. Comput. Virol. **3**(2), 65–74 (2007)
15. Filiol, E., Raynal, F.: Malicioux, Malicious Cryptography … Reloaded and also Malicious Statistics. CanSecWest 2008 Vancouver, pp. 26–28 Mars (2008)
16. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, London (2001)
17. Lawrence, C.E., Reilly, A.A.: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins Struct. Funct. Genet. **7**, 41–51 (1990)
18. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262**, 208–214 (1993)
19. Newsome, J., Karp, B., Song, D.: Polygraph: automatically generating signatures for polymorphic worms. In: 2005 IEEE Symposium on Security and Privacy (S&P'05), pp. 226–241 (2005)
20. Hirschberg, D.S.: Algorithms for the longest common subsequence problem. J. ACM **24**4, 664–675 (1977)
21. DefensePro, Radware. http://www.radware.com/
22. Abou-Assaleh, T., Cercone, N., Kešelj, V., Sweidan, R.: NGram Based Detection of New Malicious Code. In: 28th Annual International Computer Software and Applications Conference Workshops and Fast Abstracts (COMPSAC'04), pp. 41–42 (2004)
23. Goldberg L.A., Goldberg, P.W., Phillips, C.A., Sorkin, G.: Constructing Computer virus phylogenies. J. Algorithms **26**(1), pp. 188–208
24. Karim, M.E., Walenstein, A., Lakhotia, A.: Malware Phylogeny Using Maximal πPatterns. In: EICAR 2005 Conference: Best Paper Proceedings, pp. 167–174 (2005)
25. Le Cam, L.: An approximation theorem for Poisson binomial distribution. Pac. J. Math. **10**, 1181–1197 (1960)
26. Lai, C.D., Wood, G.R., Qiao, C.G.: The mean of the inverse of a punctured normal distribution and its application. Biom. J. **46**(4), 420–429 (2004)
27. Rokach, L.: Collective-agreement-based pruning of ensembles. Comput. Stat. Data Anal. **53**(4), 1015–1026 (2009)
28. Menahem, E., Shabtai, A., Rokach, L., Elovici, Y.: Improving malware detection by applying multi-inducer ensemble. Comput. Stat. Data Anal. **53**(4), 1483–1494 (2009)
29. Moskovitch R., Elovici Y., Rokach L.: Detection of unknown computer worms based on behavioral classification of the host. Comput. Stat. Data Anal. **52**(9), 4544–4566 (2008)
30. Rieck, K., Laskov, P.: Language models for detection of unknown attacks in network traffic. J. Comput. Virol. **2**(4), 243–256 (2007)