SA^3: Automatic Semantic Aware Attribution Analysis of Remote Exploits

Deguang Kong[1,2], Donghai Tian[1], Peng Liu[1], and Dinghao Wu[1]

[1] College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802,
{dkong,dtian,pliu,dwu}@ist.psu.edu
[2] Dept. of Computer Science and Engineering, University of Texas at Arlington, TX, 76013

Abstract. Web services have been greatly threatened by remote exploit code attacks, where maliciously crafted HTTP requests are used to inject binary code to compromise web servers and web applications. In practice, besides detection of such attacks, attack attribution analysis, i.e., to automatically categorize exploits or to determine whether an exploit is a variant of an attack from the past, is also very important. In this paper, we present SA^3, an exploit code attribution analysis which combines semantic analysis and statistical analysis to automatically categorize a given exploit code. SA^3 extracts semantic features from an exploit code through data anomaly analysis, and then attributes the exploit to an appropriate class based on our statistical model derived from a Markov model. We evaluate SA^3 over a comprehensive set of shellcode collected from Metasploit and other polymorphic engines. Experimental results show that SA^3 is effective and efficient. The attribution analysis accuracy can be over 90% in different parameter settings with false positive rate no more than 4.5%. To our knowledge, SA^3 is the first work combining semantic analysis with statistical analysis for exploit code attribution analysis.

Keywords: Remote Exploit, Shellcode, Attribution, Mixture of Markov Model.

1 Introduction

A great number of code injection attacks (e.g., buffer overflow attacks, format string attacks) are used by crafted HTTP requests to compromise different kinds of web services or web applications. From the CERT [1] and SecurityFocus [2] statistics, the remote code injection attack is still one of the major attacks these days. In (remote) code injection attacks, malicious HTTP requests/replies can be forged to inject malicious code by masquerading as normal requests/replies. Different kinds of shellcode are representatives of exploit code, which can be injected into target services or applications through network connections. Worms can take advantage of these exploit code for infections and propagations. In this paper, the exploit code we focus on is remote shellcode which can be used as the payload of a packet to spread via HTTP requests. Throughout the paper, we use the terms remote exploit code and shellcode interchangeably.

There are mainly two types of techniques used for shellcode analysis and detection: the emulation-based approach and statistics-based approach. The emulation based

M. Rajarajan et al. (Eds.): SecureComm 2011, LNICST 96, pp. 190–208, 2012.

approach (e.g., [3,4]) emulates the executions of instruction sequences, and thus shellcode's behaviors are exposed in the virtual running environment. However, it is antagonized by many kinds of anti-emulation techniques [5]. For example, drive-by downloads web attacks [6], which target memory corruption vulnerabilities, have to prepare the environment before their successful launch. Improper emulations of the execution context will lead to incorrect executions of instruction sequences, and thus fail to expose specific specific behaviors.

Statistical analysis is another promising method used in network intrusion detection systems including the remote shellcode detection and analysis [7,8]. The basic idea of the statistical approach is to extract the distinguished features to differentiate between the normal packets and various malicious packets. The payload of a packet and the payload header information (e.g., port number, protocol field) can be used as features for classification. The disadvantage of the statistical approach is that it usually lacks clear semantic information correlated with the packets whose contents may result in malicious behaviors, and therefore it can also be evaded by different kinds of anti-statistic approaches [7,8].

From the analysis above, we can see current exploit code detection and analysis approaches are still quite limited. Meanwhile, lots of shellcode variants appeared in the past several years according to AV-test's statistics [9]. Thus in this paper we present an automatic semantic aware attribution analysis of remote exploits. The significance of such analysis is that it provides more information about an attack in addition to detecting the attack. The attribution analysis can be used in, for example, a shellcode scanner to identify different types of shellcode variants. As far as we know, such shellcode attribution analysis is still lacking in the literature. Note that Hu et al [10] present a function-call graph based approach to index the large malware repositories, which can be viewed as a kind of malware attribution analysis. Our motivation is similar to theirs, but our work is more specific for shellcode attribution analysis. Compared with shellcode detection, our work focuses more on automatically categorizing exploits and determining whether an exploit is a variant of an attack from the past. We believe this is also important besides telling whether a piece of code is malicious or not.

Exploit code attribution poses several challenges. First, the emulation based approach cannot be directly applied to this problem because we need quantitative metrics to measure the distances of different exploit code. Second, we cannot fully rely on the statistical approach because it is deceptive once the statistical features (e.g., the number of specific instructions or system calls) fail to reflect the security-critical operations, which are probably highly related with the shellcode behaviors. Third, how to extract the semantics which determine the shellcode attribution remains an open question. The emulation based approach seems a good candidate for extracting the behaviors of different shellcode. However, it can miss trivial differences existed in the behaviors of different classes of shellcode. For example, self-contained exploit code [4] often exhibits same behaviors by following the routine of "decrypt-loop" mode. Furthermore, if specific behaviors are absent in the emulation environment, it could produce more false negatives. Also, the time cost for the emulation based approach is usually very high compared with static analysis.

Our Approach. We present SA^3, a novel automatic Semantic Aware Attribution Analysis of remote exploit code. SA^3 first makes semantic analysis on the payload of packets, and then a Markov-based model is used to model each type of shellcode. Specifically, for semantic analysis, we use static data anomaly analysis on the packet payload; for statistic analysis, we use a two-way of Mixture Markov model. The statistical model is based on the refined exploit code sequences, which are pruned from the whole code sequences in the framework of static analysis. Once the model is built, any new code can be fed into the model to get an attribution analysis result.

One important characteristic of our work is that we use the "features" acquired from semantic analysis for attribution analysis. We present SA^3 based on an observation that the attribution for a piece of exploit code has great correlations with the exploit code's semantic characteristics (e.g., the opcode sequence, the instruction sequence) and also its statistical characteristics (e.g., the number of instructions, the out-degree of control flow graph). The changes of the semantics also cause the changes of the statistic exposure in shellcode instructions. This observation motivates us to consider about the integration of semantic analysis with statistical analysis by taking advantages of both of them.

Our work stands between the semantic analysis and statistical analysis. Instead of using dynamic emulation techniques introduced before, our work uses the static data anomaly analysis by making static analysis on the instruction sequences. The advantage of this approach is to capture the semantics of the exploit code with moderate time cost. Also, it will not be attacked by anti-emulation techniques [5]. Compared with only emulation based approach, our work can also overcome some inherent defects (e.g., different shellcode may expose similar behaviors) by introducing the statistical analysis. Compared with only statistical based approach, our analysis is more robust by incorporating the semantics to avoid "black-box" learning.

Contributions. The main merits of SA^3 are listed as follows. To our knowledge, our approach is the first work to make exploit code attribution analysis by combining semantic analysis with statistical analysis. Semantic analysis is used for extracting the semantic-binding code with certain malicious intent. Statistical features can help to capture the "whole" view of a packet from macroscopic point. These two different views complement each other. Our evaluation shows that our analysis result is better than purely statistical approach, which also refutes the conclusion of "impossibility of modeling polymorphic shellcode [11]" in some degree.

The rest of this paper is organized as follows. First of all, we formalize the problem in Section 2. Next we show our approach SA^3 in Section 3, followed by evaluation in Section 4. Then we introduce the related work in Section 5. Finally, we conclude the paper in Section 6.

2 Problem Statement and Analysis

2.1 Problem Formalization

Let $I(i \in I)$ be a set of different classes of exploit code; and $D(1 \leq j \leq D)$ be the total number of instances (variants) generated from a certain class. We use S_{ij} to denote

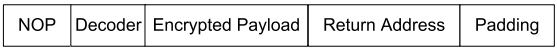

Fig. 1. A demonstration of polymorphic shellcode instance

the jth exploit code instance generated from class w_i, i.e., $S_{ij} \in w_i$. For example, in reality, a set of different types of exploit code can be generated from different polymorphic shellcode engines $I = \{\text{Clet}, \text{CountDown}, \text{Pex}, \text{Tapion}, ...\}$. Different instances of the same type of exploit code can be generated using complicated obfuscation techniques like polymorphism and metamorphism [12].

Definition 1. Exploit Code Attribution Problem

(1) For lots of exploit code instances S_{ij}, how to generate a profiling for each category w_i;

(2) For an unknown exploit code s, what is the attribution of s? That is, find i such that $s \in w_i$.

2.2 The Challenge of the Problem

According to the definition of exploit code attribution problem, Problem (1) is a training problem in shellcode classification and Problem (2) is an recognition problem after a profiling for each category of exploit code is built. Problem (1) is the key step while Problem (2) can be easily solved after the learning model is built in Problem (1). These two problems match well with the standard machine learning problem. Naturally, we refer to machine learning techniques for a solution.

From above analysis, it seems any statistical approach can work in the context of exploit code attribution analysis. However, the statistics-based approach may not produce promising results. Song et al. [11] conclude that it is impossible to model the polymorphic shellcode (See Fig. 1 for an example). Polymorphic shellcode accounts the largest part of the exploit code, and therefore, modeling all of the exploit code (e.g., for attribution analysis) is much more difficult. Next, we will briefly explain why modeling polymorphic shellcode instances is difficult. The contents of the polymorphic shellcode instances usually consist of several parts: NOP part (sled), decoder part, encrypted payload part, return address part and padding part (Fig. 1). Modeling the NOP part may amount to modeling random instructions because many instructions are semantically equivalent to "NOP." For example, for the shellcode generated by CLET [13], there are 55 kinds of sled used in the "NOP" part. In the return address part, there are also many variations of the target address by adding padding bytes before it. In the padding part, the binary code can be filled in, without influencing the execution results. For obfuscation purpose, the padding bytes may have similar distribution to the normal traffic distribution. In the decoding part, different encryption keys can generate different encrypted exploit code. Clearly, due to great varieties in each part of polymorphic exploit code, the variations for a whole exploit code packet can be even larger. These great variations may result in, (R_1) no fix patterns exposed in a whole packet; (R_2) the attribution analysis process misguided by padding bytes and noisy bytes.

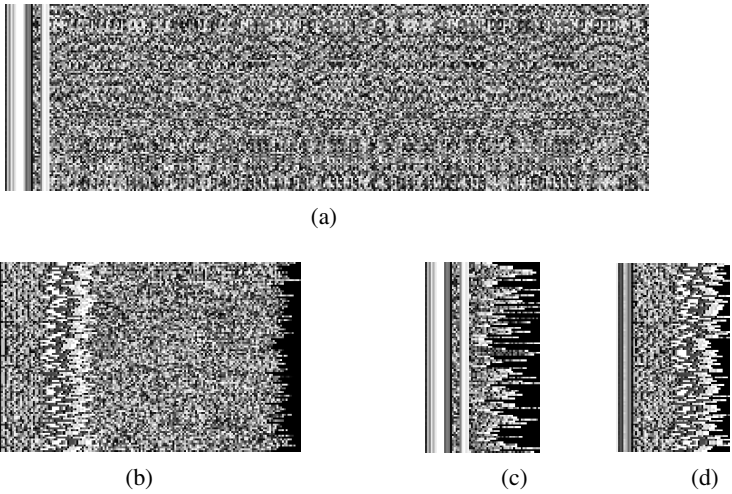

Fig. 2. Varieties of shellcode instances. (a) Pex (each of 100 instances is 344 bytes); (b) CLET (each of 100 instances is 168 bytes); (c) Pex refined shellcode after semantic analysis (corresponding to (a), each is 60 bytes); (d) CLET refined shellcode after semantic analysis (corresponding to (b), each is 60 bytes). Each pixel of the image represents a byte obtained from a shellcode instance

Fig. 2 shows two examples of shellcode varieties. Fig. 2(a) shows a spectral image, where each pixel represents a byte from a shellcode sequence generated from polymorphic engine Pex [14]. Each row is corresponding to a shellcode sequence with 344 bytes in length, and totally 100 instances form the image. Similarly, Fig. 2(b) shows the spectral image formed by 100 sequences generated from polymorphic engine CLET [13], where each row is a shellcode sequence of 168-byte in length. Clearly, these images demonstrate great varieties of different bytes in exploit code, which imply that the shellcode attribution analysis is a challenging problem.

3 Approach

In Fig. 3, we describe the framework of SA^3 . The core modules of SA^3 are *Semantic Analysis Module*, and *Statistical Analysis Module*. More detailedly, we use *Data Anomaly Analysis* in the Semantic Analysis Module and a *Two-way Mixture of Markov Model* in the Statistical Analysis Module.

The whole workflow of SA^3 can be divided into training stage (*with the real line*) and the recognition stage (*with the dashed line*). First of all, the same type of exploit code instances are fed into the semantic analysis module, and data anomaly analysis are conducted on them. We get the refined exploit code instances, which are actually the instruction sequences pruned of useless instructions. Next, a two-way Mixture of Markov Model is built on the refined input instruction sequences. We construct a mixture of Markov Model corresponding to each category of exploit code. When a new

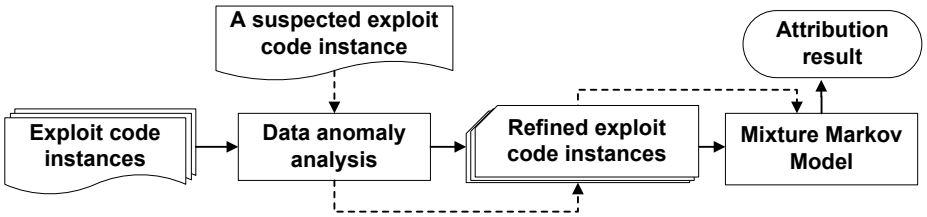

Fig. 3. SA^3 flow graph (real line for training stage, dashed line for recognition stage)

exploit code instance comes, it will be first analyzed through the data anomaly analysis module. Thus the refined code sequences are distilled as the input to the two-way mixture of Markov Model. The decision result is obtained by attributing the exploit code sequence to the one with the most fitting value.

Semantic Module. For each category of the input instruction sequences, we prune semantic-unrelated code existed in the code sequences. Data anomaly analysis is used to capture the "semantics" of the exploit code through preserving the useful instructions while pruning the useless ones which are probably containing padding and noisy bytes in the packets. This module is used for solution of R_2 presented in Section §2.2.

Statistical Module. For the pruned instruction sequences, a two-way Mixture of Markov Model is built for the solution of R_1 (Section §2.2). On one hand, it is not very clear what kind of relationship exists in the instruction sequences. On the other hand, Markov model is very suitable to model the uncertainty existed in different context. Thus, we refer to a two-way Mixture Markov model, to model relationships between the instruction sequences. The property of "two-way mixture model" makes it more robust and powerful to represent the varieties of different categories of code.

3.1 What Is the Semantics Used in Exploit Code?

Data Anomaly Analysis. We observe that certain control and data flow information remain invariable to implement certain functions in the exploit code. We call those control and data flow information as "semantic." The data anomaly analysis is used to capture those semantics, because it can preserve the useful instructions by pruning the useless ones which contain padding and noisy bytes.

First of all, we use disassemble analysis to analyze the input binary instruction sequences. In this paper, what we focus on is HTTP message flows. In an HTTP request message, malicious payload only exists in Request-URI and Request-Body of the whole flow [15]. We extract these two parts from the HTTP flows for further semantic analysis. Then we make disassemble analysis on these input sequences. If the disassemble module finds consecutive instructions in the input sequences, it generates the disassemble instruction sequences as output. An instruction sequence is a sequence of CPU instructions which has only one entry point. A valid instruction sequence should have at least one execution path from the entry point to another instruction within the sequence. Since we do not know the entry point of the code when the code is present in the byte sequences, we explore an improved recursive traversal disassemble algorithm

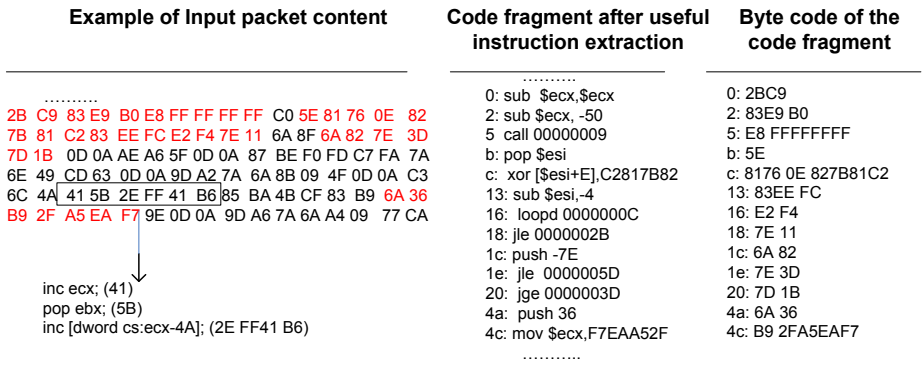

Fig. 4. A motivating example to show the procedure of semantic analysis on the input code sequence

introduced by Wang et al. [15] to disassemble the input instruction sequences. For an N-byte sequence, the time complexity of disassemble algorithm is $O(N)$.

After disassemble analysis, it may generate zero, one, or multiple instruction sequences, which do not necessarily correspond to real code. Next, we distill useful instructions by pruning useless instructions using the technique introduced in SigFree [15]. Useless instructions are those illegal and redundant byte sequences. By using the code abstraction, a static analysis technique, we can emulate the executions of instruction sequences. There are possibly 6 states in the state transition graph generated from the code sequences. State *U* represents undefined variable state; state *D* represents defined but not referenced variable state and state *R* represents defined and referenced variable state. The other three abnormal states are defined as follows: state *DD* represents abnormal state *define-define*, state *UR* represents abnormal state *undefine-reference*, and state *DU* represents abnormal state *define-undefine*. Basically, the pruned useless byte sequences correspond to three kinds of dataflow anomalies: *UR*, *DD*, *DU*. When there is an undefine-reference anomaly (i.e., a variable is referenced before it is ever assigned with a value) in an execution path, the instruction which causes the "reference" is a useless instruction. When there is a define-define anomaly (i.e., a variable is assigned a value twice) or define-undefine anomaly (i.e., a defined variable is later set by an undefined variable), the instruction that caused the former "define" is also considered as a useless instruction. Since crafted noisy bytes in the packets typically do not contain useful instructions, such irrelevant bytes in the packets are filtered out after the useful instruction extraction phase. The remaining instructions are likely to be related to the semantics of the code kept in the exploit code sequences.

Here, we further explain our motivation for useful instruction extraction. From our observations, lots of "useful instructoins" are left invariant across different shellcode instances even after complicated obfuscations (e.g., "junk insertion," "instruction replacement"). For padding and noisy bytes, they still can be assembled into code sequences. However, usually it lacks clear meanings and correlated relations for those coincidental instruction sequences. Thus, they will be pruned after rigorous data flow anomaly analysis. Moreover, we note that the remaining useful code sequences are more likely to be similar to those from the same category instead of those from the other categories.

Motivating Example. An example of polymorphic code analysis is shown in Fig. 4. Here the leftmost part is the original packet content in binary, the middle part and the right part are the disassemble code and its corresponding binary code of the useful instructions after removing useless ones, respectively. For example, the disassembly code *inc ecx* appeared in address 42 is pruned because *ecx* is defined again in address 4*c* to produce a *define-define* anomaly. In address 44, the contents in the memory cell with address *ecx*-4*A* is referenced without being defined beforehand. Thus we prune this instruction because it produces an *undefine-reference* anomaly.

Figs. 2(c) and 2(d) show two other examples. Fig. 2(c) gives the spectral image formed by the remaining instructions of 100 instances corresponding to Fig. 2(a). Similarly, Fig. 2(d) gives the spectral image formed by the remaining instructions of 100 instances corresponding to Fig. 2(b). In both images, each pixel represents a byte from the remaining instructions. Clearly, the lengths of the preserved code sequences are decreased. More importantly, the fixed patterns in the original code sequences are preserved while the bytes located in different positions with large varieties are cut off.

From the above polymorphic shellcode example (Fig. 4) and other instances, we find that the remaining code sequences usually consist of the following features: (F_1) getPC: the code to get the current program counter, usually contains opcode "*call*" or "*fstenv*"; (F_2) Iteration: a polymorphic exploit code usually performs iterations over encrypted shellcode using the operations like *loop*, *rep* and the variants of such instructions (e.g., *loopz*, *loope*, *loopnz*); (F_3) Jump: a polymorphic exploit code is probable to contain conditional/unconditional branch statements (e.g., *jmp*, *jnz*, *je*); (F_4) Decryption: for the encrypted shellcode, certain machine instructions (e.g., *or*, *xor*) are more often to be found in decryption routines since decryption needs to decrypt the shellcode before execution. These features are preserved after semantic analysis, which can be further used for statistical modeling. We believe these features help to capture the category of shellcode, and they may exist in most of the self-contained exploit code.

It may be attempting to use (F_2, F_4) as the only feature for category analysis. Fortunately, we also have other useful instructions preserved except for the features (F_1, F_2, F_3, F_4). This motivates us to use the statistical model for capturing the differences across various exploit codes as much as possible. For non self-contained code, not all features (e.g., F_2, F_1) exist in the shellcode (e.g, code generated from Avoid UTF8/tolower [4]) because of the absence of GetPC and self-reference operations. In these cases, the remaining instruction sequences still can be taken as good indicators for shellcode category analysis since noisy bytes are filtered. The pruned bytes are more likely to mislead the state-of-the-art statistics-based learning approaches (e.g., N-gram based learning [16], Markov Chain [8], Support Vector Machine [6]) for category analysis or detection. Note the length of code sequence can be viewed as the dimensions for the training code sequences. To prune useless instruction also means to reduce the dimension of training data This makes the machine learning module much easier and more accurate by alleviating the difficulty of "curse of dimensionality [17]".

3.2 What Is the Statistics Used for Modeling?

Why Use Markov Model? Let Y be the set of single bytes and Y^i denote the set of i-byte sequences. $X = Y \cup Y^2 \cup Y^3 \cup Y^4$ is the token set in our system because a

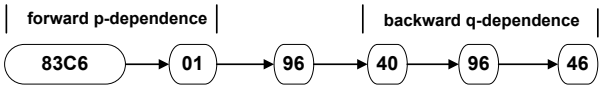

Fig. 5. Explanation of dependence in Markov Model

token in a useful instruction contains at most four bytes (e.g., "*AAFFFFFF*"), which corresponds to the word size of 32-bit systems. A Markov chain [18] is a sequence of random variables X_1, X_2, X_3, ..., satisfying the Markov property: given the present state, the future and past states are independent. More formally, probability $\Pr(X_{n+1} = x|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \Pr(X_{n+1} = x|X_n = x_n)$, where x_i is the value for each state X_i, and $\Pr(X_{n+1} = x|X_n = x_n)$ is the conditional probability for transition from state X_n to X_{n+1}. The possible values of X_i form a countable set S called the state space of the chain. We observe that there are close relations among the code tokens in the refined instruction sequences. Markov Chain [18] is a good candidate to model uncertain dependencies in different contexts. In the context of code sequence analysis, each token in a sequence can be viewed as a state in a Markov Chain. We assume a token in a sequence is dependent on the token in front and also the token next because of the great dependencies existed in the code sequences of the nearest neighbors. To be exact, the dependency of token x_j on x_i is the co-occurance of token x_j and x_i. If x_i appeared in front of x_j in the same sequence, we call x_j is forward dependent on x_i. Otherwise, if x_i appeared after x_j in the same sequence, we call x_j is backward dependent on x_i. 1-order Markov chain requires the nth token in a chain is only dependent on the $(n-1)$th token. However, in real code segment, the nth token can be dependent on the $(n-1)$th, $(n-2)$th, ..., $(n-p)$th tokens in a sequence, and also related to $(n+1)$th, $(n+2)$th, ..., $(n+q)$th tokens. We do not know what is the value of p and q beforehand.

In our model we define two kinds of relationships to represent those bidirectional dependence. We call our Markov-derived model as a Two Way Mixture Markov (TWMM) Model. First, we define the forward dependence, i.e., nth token is depended on consecutive p tokens in front. Next, we define the backward dependence, i.e., nth token is dependent on the next consecutive q tokens. Then parameters $\pi_i (i = 1, 2)$ are used to make a balance between them, where $\pi_1 + \pi_2 = 1$. Fig. 5 shows an example, where token 96 is forward dependent on p ($p = 2$) tokens ($83C6, 01$) in front, and also backward dependent on next q ($q = 3$) tokens ($40, 96, 46$).

Model Construction. First, we construct a TWMM model for each category of code sequences. Second, after a new code sequence is fed into the model, we attribute it to the class with the highest fitting value. However, if the highest fitting value is still less than a certain threshold, we will attribute it to the normal sequence. Here the fitting value is the accumulation of probabilities, which reflects the matching score from a code sequence to the model.

Next we show how to compute the probability for a code sequence. The probability of a code sequence can be decomposed into the product of the probability of each token in a sequence. For different tokens appeared, there is a transition matrix to label the probability from one token to another. Hence, p-forward tokens' transition

probability to a specific token is the probability from front p tokens' transition probability to this token. Similarly, q-backward tokens' transition probability to a specific token is the probability from next q tokens' transition to this token. In forward model, the ith token's probability is computed through product of the p-forward tokens' transition probability to this token. Similarly, in backward model, the ith token's probability is computed through the product of the q-backward tokens' transition probability to this token. Since the probability is a product of $p(L_n - p)$ values in the forward model, and a product of $q(L_n - q)$ values in the backward model, where L_n is the length for the nth code sequence. Therefore, the $p(L_n - p)$ root is needed for computing the sequence probability in forward model and $q(L_n - q)$ root is needed in backward model.

More formally, let $x_{n,i}$ denote the ith token in the nth sequence, $A_1(x_{n,i}|x_{n,j};\theta_1)$ denote the transition probability from token $x_{n,j}$ to token $x_{n,i}$ in forward model θ_1, $A_2(x_{n,i}|x_{n,j};\theta_2)$ denote the transition probability from token $x_{n,j}$ to token $x_{n,i}$ in backward model θ_2. Since the same token can be transferred to different tokens, the sum of such transition probability should be normalized to 1, i.e.,

$$\sum_{x_{n,i}} A_k(x_{n,i}|x_{n,j};\theta_k) = 1 \ (k = 1, 2). \tag{1}$$

Let $g(x_n|\theta_1)$ and $g(x_n|\theta_2)$ denote the probability for the nth sequence's matching scores in the forward model and backward model, respectively. Thus we have

$$g(x_n|\theta_1) = \left(\prod_{i=p+1}^{L_n} \prod_{j=i-p}^{L_n-p} A_1(x_{n,i}|x_{n,j};\theta_1) \right)^{\frac{1}{(L_n-p)p}} \tag{2}$$

$$g(x_n|\theta_2) = \left(\prod_{i=1}^{L_n-q} \prod_{j=i+1}^{i+q} A_2(x_{n,i}|x_{n,j};\theta_2) \right)^{\frac{1}{(L_n-q)q}} \tag{3}$$

Next, by combing $g(x_n|\theta_1)$ and $g(x_n|\theta_2)$ in a balanced way, we have G_n to denote the matching score for nth sequence, i.e.,

$$G_n = \sum_{k=1}^{2} \pi_k g(x_n|\theta_k), \tag{4}$$

where $\pi_1 + \pi_2 = 1$. To obtain the solution for this model means to estimate the parameters in Eq. (4). Suppose we have N sequences for each category, thus the object function G to be optimized is the product of the likelihood for each sequence G_n, i.e.,

$$G = \prod_{n=1}^{N} \sum_{k=1}^{2} \pi_k g(x_n|\theta_k). \tag{5}$$

Model Solution. Here we show how to solve Eq.(5). The object function G is to be maximized to fit the model according to the principle of maximum likelihood estimation [19]. From the point view of optimization techniques, the object function is not

concave in terms of the mixture of two different Markov chains, thus directly setting the first order derivatives on the likelihood does not work. This model is also different from the standard mixture model which requires the same format of sub-models in a mixture model. Thus we use the Expectation Maximum (EM) algorithm [20] to iteratively maximize the likelihood function with a gradient descent algorithm. The EM algorithm usually takes two steps, Expectation Step and Maximization Step. At each step, the model's likelihood function is updated in the direction of gradient ascent, and this process is iterated until the likelihood converges. The monotonic property makes this approach effective for the solution of many non-convex optimization problems. Next we show how to train our model with the EM algorithm.

First, we construct the affiliated function [20]

$$W(\Theta, Q) = \sum_{n=1}^{N} \sum_{k=1}^{2} Q_{nk} \log \frac{\pi_k g(x_n|\theta_k)}{Q_{nk}}, \tag{6}$$

where Q_{nk} works as the hidden variable to denote the weight of data point n in terms of model k, and $\sum_{k=1}^{2} Q_{nk} = 1$. Since the log function is a concave function, according to the Jensen's inequality,[1] we have $\log(\sum x) \geq \sum \log x$. Thus $\log G \geq W(\Theta, Q)$. The maximization of the object function G in Eq. (5) is equivalent to the maximization of Eq. (6) because Eq. (6) is the new lower bound of the likelihood function to be maximized. Let Θ denote the parameters in the transition probability matrix $A_k(x_{n,i}|x_{n,j};\theta_k)(1 \leq k \leq 2)$, Q denote the hidden variable set Q_{nk}. Let Θ^t and Q^t denote each group of parameters used in the tth iteration in the parameter estimation process. During the maximization step, the object function of Eq. (6) is required to be monotonically increased. Based on this, we obtain

$$W(\Theta^t, Q^t) \leq W(\Theta^{t+1}, Q^t) \leq W(\Theta^{t+1}, Q^{t+1}), \tag{7}$$

which can be solved by using the Lagrange Multipliers [21] to find the stationary points with $\arg\max_{\Theta} W(\Theta, Q^t)$ and $\arg\max_{Q} W(\Theta^{t+1}, Q)$ satisfied in each step.

Let $C(x_{n,i}|x_{n,j})$ denote the frequency of token transition from $x_{n,j}$ to $x_{n,i}$ in nth sequence. Naturally, we use $C(\cdot|x_{n,j})$ to denote the frequency of the token transition from $x_{n,j}$ to any tokens in the nth sequence of the model. To solve Eq.(7), we obtain solutions in Eqs. (8–9). The complete training algorithm is shown in the table below.

$$Q_{nk} = \frac{\pi_k g(x_n|\theta_k)}{\sum_{k=1}^{2} \pi_k g(x_n|\theta_k)}, \pi_k = \frac{\sum_{n=1}^{N} Q_{nk}}{N}, \tag{8}$$

[1] For any concave function $f(x)$, if the balanced parameter t satisfies $0 < t < 1$, we have $f(tx_1 + (1-t)x_2) \geqslant tf(x_1) + (1-t)f(x_2)$.

$$A_k(x_i|x_j;\theta_k) = \frac{\sum_{n=1}^{N} \frac{Q_{nk}}{L_n(L_n-\lambda_k)} C(x_{n,i}|x_{n,j})}{\sum_{n=1}^{N} \frac{Q_{nk}}{L_n(L_n-\lambda_k)} C(\cdot|x_{n,j})}, \lambda_1 = p, \lambda_2 = q. \quad (9)$$

Algorithm 1. EM training Algorithm

Input: Instruction sequences $I_0, I_1, I_2, ..., I_n$ of each category, ε is the parameter used for convergence decision.

Output: Parameters (Θ, Q) for each category.

Procedure:

1: Initialize $\pi_k, A_k(x_{n,i}|x_{n,j};\theta_k), Q_{nk}(1 \le k \le 2)$
2: Compute the probability for each sequence to obtain $W(\Theta^t, Q^t)$ with Eq.(6)
3: update Q_{nk},π_k with Eq.(8); update $A_k(x_{n,i}|x_{n,j})$ with Eq.(9)
4: **if** $W(\Theta^{t+1}, Q^{t+1}) - W(\Theta^t, Q^t) < \varepsilon$ **then**
5: The algorithm converges, stop training
6: **else**
7: goto Step 2
8: **end if**

The above Markov-derived model has a large state space (2^{32}), and thus it seems impractical for code sequence recognition. Fortunately, lots of tokens never or seldom appear in the state space, and this gives us the opportunity to greatly decrease the state space. First, we ignore never appeared tokens and prune seldom appeared tokens by setting a threshold. It leads to sparse items in the whole state space and very sparse transition matrices. Second, we use the data structure of hash table for storage of state transition probabilities in order to reduce the computation cost.

4 Evaluation

We test our system offline on massive polymorphic exploit code packets and on HTTP normal reply/request traces. First of all, we evaluate our approach on different kinds of exploit code in terms of false negatives and false positives, and then we compare our approach with the approach free of any semantic analysis before attribution analysis. Next, we evaluate our approach in terms of computation time cost. Finally, we discuss the advantages and limitations of our approach.

The massive polymorphic exploit code packets are generated by the metasploit [14] framework (e.g., PexFnstenvSub, Pex, ShikataGaNai), and also from polymorphic engines (e.g., CLET [13], ADMmutate [22], JempiScodes [23]). CLET, ADMmutate, JempiScodes and ShikataGaNai are advanced polymorphic engines which obfuscate the decryption routines by metamorphism such as instruction replacement and garbage insertion. CLET uses spectrum analysis to counterattack the byte distribution analysis. Opcodes of the "*xor*" and "*fnstenv*" instruction are frequently found in the decryption routine of PexFnstenvSub and also in getting the values of register of the program counter (GetPC). Pex uses *xor* decoders and relative call to get PC. The normal

HTTP traffic contains 300,000 messages collected for three weeks at seven workstations owned by seven different individuals in our lab's computers. To collect the traffic, a client-side proxy monitoring incoming and outgoing HTTP traffic is deployed underneath the web server. Those 300,000 messages contain various types of non-attack data including JavaScript, HTML, XML, PDF, Flash and multimedia data, which render diverse and realistic traffic typically found in the wild. We run our experiments on a 2.4GHz Intel Quad-Core machine with 2GB RAM, running Windows XP SP2.

4.1 Attribution Analysis Results

First, we evaluate our approach in different parameter settings in terms of different combinations of p and q. Second, we compare our approach with the approach free of making any semantic analysis beforehand. Here we do not discuss much about data anomaly analysis, since they have been well studied in previous researches [15,24].

Exploit Code Attribution. For each category of exploit code, we generate a corresponding TWMM Model, and then the new packets are fed into the model to evaluate the false positives and false negatives. We use 5-fold cross validation to train the model and get the false negatives by matching the packet with the corresponding model. During the packet attribution phase, a threshold is set to decide the attribution for this packet. The threshold will both influence the false positives and false negatives in the ROC curve. As is shown in Fig. 6, for different combinations of p and q, we can get different results by setting different thresholds. Another factor to influence the attribution analysis result is the setting of the parameters p and q. There are many choices of (p, q) combinations since p and q can be freely selected if we do not know any prior knowledge of the structures of code sequences. It is not realistic to brute-force search all possible (p, q) combinations. From our observations, for each token, the tokens close in distance have much more influential power on it. That means p and q can be set to small numbers. We do not know exactly which is the best to achieve the optimal results. In our evaluation, tentatively, we select $p, q \in \{2, 4\}$. From the results on different datasets in Fig. 6, we can infer that token relevances are different on different datasets. Besides the parameters which influence the attribution results, the attribution analysis accuracy varies depending on the "nature" of the exploit code. On all six datasets, the detection accuracy can reach to above 90% in different parameter settings. This is a good indictor to show the effectiveness of our approach. The false positive rate is up to 4.5% at most. We may further bootstrap the misclassified packets to increase the analysis accuracy in our future work. .

Comparison with Approach *without* Semantic Analysis. We compare our approach with the approach free of any semantic analysis beforehand. The same TWMM model is constructed for the original packets but *without* any semantic analysis before the attribution analysis. In the approach without any semantic analysis, the tokens used are all one-byte tokens because we do not have any prior knowledge about the minimum semantic cell used in the whole code sequence. Note that the changes of combinations of (p, q) do not make much difference for detection accuracy and false negative rate in our attribution analysis, thus we set $p = 2, q = 2$ when making a comparison with the approach without semantic analysis. The results are also shown in Fig. 6. Our semantic

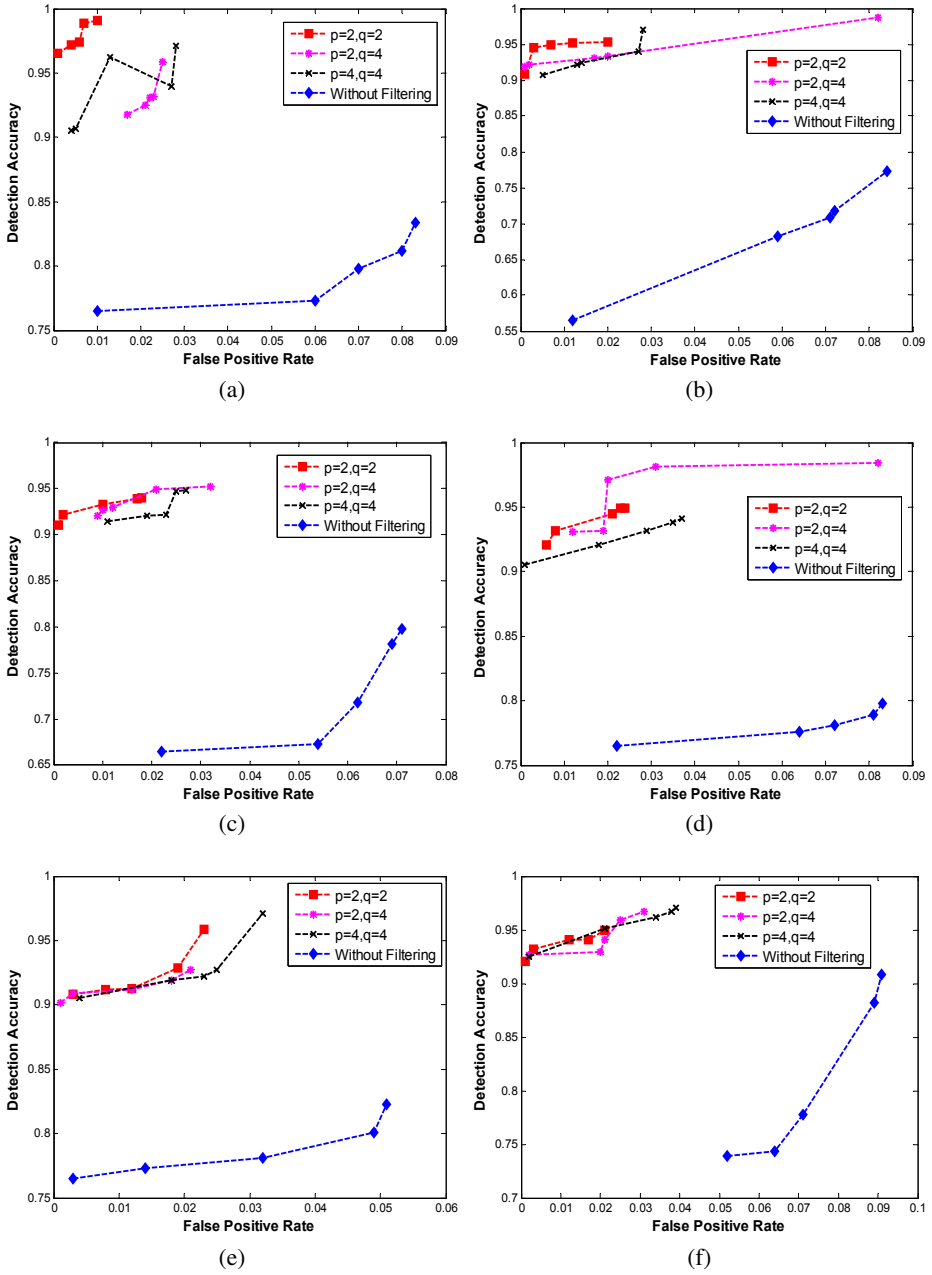

Fig. 6. Comparisons of semantic aware approach with the approach without filtering noises on six data sets: (a) CLET; (b) ADMutate; (c) PexFnstenvSub; (d) JemipiScodes; (e) Pex; (f) Shikata-GaNai

Table 1. Average Time Cost for Each Packet (millisecond)

Polymorphic engine	Training time	Decision time
CLet	5,213	7.2
Admutate	4,142	3.2
PexFnstenvMov	3,829	4.5
JempiScodes	2,487	6.8
Pex	3,152	4.1
ShiKataGaNai	7,650	4.3

aware approach outperforms the approach without semantic analysis on all six data sets, and the detection accuracy can be boosted more than 10% on all six data sets with nearly the same false positives. The promising results show that our semantic aware attribution analysis is effective and much better than the approach free of semantic analysis.

4.2 Performance Evaluation

Table 1 shows the time cost during the training phase and decision phase. The training time is the average time cost for each packet used in training, which includes the time of semantic analysis and also the time used for the training process of statistical model. The decision time is the average time cost for packet recognition, also including the time for semantic analysis. The training time cost is high due to the EM algorithm used in mixture Markov model. The EM algorithm usually needs hundreds times of iterations before convergence especially when data do not fit a model very well (e.g., exploit code instances generated from ShiKataGaNai). It also takes time in the semantic analysis module, but the time cost for semantic analysis is negligible compared with the EM algorithm in the training process. Fortunately, in order to reduce the time cost, we can conduct the training process offline before the recognition phase.

4.3 Discussion

Here we further discuss the strengths and limitations of our approach.

Strengths. First of all, our approach can filter noises through semantic analysis in the code sequences, and thus it has very good noise tolerance. Second, our approach is very robust to many different kinds of attacks (e.g., coincidental-pattern attacks [25], the token-fit attacks [26], allergy attacks [27]) due to the semantic analysis module applied. Moreover, our approach explores the semantic features to the classification process which leverages the "semantics" to increase attribution analysis accuracy. This opens a door to combine the semantic analysis with statistical analysis for practical tasks.

Limitations. First of all, since our semantic module is based on static analysis, we cannot handle some state-of-the-art code obfuscation techniques (e.g., branch-function obfuscation) in the semantic module, which may mislead the feature generations before statistical analysis. This can be solved by referring to more complicated semantic aware static/dynamic analysis techniques (e.g., symbolic execution, type inferences). Secondly, for non-self contained exploit code [4], sometimes we fail to capture the features of such code before statistical analysis. The code may mislead the classifier to

make the wrong decision results. This is also a problem that state-of-the-art statistical learning techniques cannot handle. Finally, during the training phase, it may be difficult to get many (e.g., 300, 400) training data for each category in a real deployment environment. The attribution results may decay due to lack of training instances. Fortunately, compared with other models (e.g., Support Vector Machine), Markov model has stronger recognition ability even with scare training data (e.g., 10, 20). That is why we use Markov-derived model in our statistical modeling module.

5 Related Work

There is a large body of work in the area of exploit code analysis and detection. We focus on two areas most related to our work: semantics-based approaches for malware especially exploit code analysis, and statistics-based approaches for those analysis.

Semantics-Based Approaches. Malware including exploit code analysis has received considerable attention from different research views. Various kinds of semantic techniques have been explored by making static or dynamic analysis on the binary code for malware detection. Emulation-based approaches [4,28] can be used to detect polymorphic shellcode by emulating the code execution to recognize specific behaviors (e.g., decryption routines) through dynamic analysis. Libemu [3] is another attempt to achieve shellcode analysis through code emulations. Gu et al. [29] present a new malicious shellcode detection methodology by analyzing snapshots of the processs virtual memory before input data are consumed. However, these emulation-based techniques can be antagonized by many anti-emulation techniques [5]. In our work, we use the static data anomaly techniques introduced in SigFree [15] to extract the semantics from the malicious code sequences. Another similar work to the semantic module we use is STIIL [24], which uses static taint and initialization analysis to detect exploit code embedded in data streams/requests targeting web services. Christodorescu et al. [30] present a dependency-graph based approach to mining the malicious behaviors present in a known malware that are not present in a set of benign programs, which can be used by malware detectors to detect malware variants. Also, Christodorescu et al. [31] use a trace semantics to characterize the behaviors of malware as well as the program being checked for infection, and use abstract interpretation to "hide" irrelevant aspects of these behaviors for malware detection/classification. The motivation of our work is very similar to these works, but ours is specific to exploit code category analysis, and more importantly, we present a novel approach for attribution analysis which combines the semantic analysis with statistical analysis. Spector [32] is a shellcode analysis system that uses symbolic execution to extract the sequence of library calls and low-level execution traces generated by shellcode. TaintCheck [33] exploits dynamic dataflow and taint analysis techniques to help find the malicious input and infer the properties of worms. Kruegel et al. [34] present a technique based on the control flow structural information to identify the structural similarities between different worm mutations. This work is close to our technique in that it analyzes the variants of worms, but they target worms, not exploit code.

Statistics-Based Approaches. Song et al. [23] study the possibility of deriving a model for representing the general class of code that corresponds to all possible decryption routines, and conclude that it is infeasible. Our work combines the semantic analysis and statistical analysis for exploit code attribution analysis, making it robust to many noise-injection attacks (e.g., allergy attack [27]). Different statistical model have been explored for intrusion detection systems, e.g., n-gram model [16] used in traffic anomaly detection, Markov chain model [8] used for web traffic anomaly detection and support vector machine [6] used for detection of drive-by-downloads attacks. A game-theoretical analysis on how a detection algorithm and an adversary could adapt to each other in an adversarial environment is introduced by Pedro et al. [35]. For exploit code attribution analysis, pure statistical approach may not produce very good results due to lack of semantic information. Recent work SAS [36] has looked at the combinations of semantic and statistical analysis to generate signatures for polymorphic worm detection. In contrast, our work is motivated for exploit code attribution analysis instead of for polymorphic worm detection, and the statistical model is also different, leading to different strategies used for classification and detection.

6 Conclusion

In this paper, we present SA^3, an automatic exploit code attribution analysis system. On the testing datasets, our approach outperforms the pure statistics-based approach with much better accuracy. To our knowledge, this is the first work that combines semantics and statistics for exploit code attribution analysis.

Acknowledgments. This work was partially supported by AFOSR FA9550-07-1-0527 (MURI), ARO W911NF-09-1-0525 (MURI), and NSF CNS-0905131.

References

1. CRET: Computer emergency response team, `http://www.cret.org/`
2. Securityfocus, `http://www.securityfocus.com/`
3. Baecher, P., Koetter, M.: Getting around non-executable stack (and fix), `http://libemu.carnivore.it/`
4. Polychronakis, M., Anagnostakis, K.G., Markatos, E.P.: Emulation-Based Detection of Non-self-contained Polymorphic Shellcode. In: Kruegel, C., Lippmann, R., Clark, A. (eds.) RAID 2007. LNCS, vol. 4637, pp. 87–106. Springer, Heidelberg (2007)
5. Bania, P.: Evading network-level emulation, `http://packetstormsecurity.org/papers/bypass/`
6. Konrad Rieck, T.K., Dewald, A.: Cujo: Efficient detection and prevention of drive-by-download attacks. In: Proc. of 26th Annual Computer Security Applications Conference, ACSAC (2010)
7. Wang, K., Cretu, G.F., Stolfo, S.J.: Anomalous Payload-Based Worm Detection and Signature Generation. In: Valdes, A., Zamboni, D. (eds.) RAID 2005. LNCS, vol. 3858, pp. 227–246. Springer, Heidelberg (2006)
8. Song, Y., Keromytis, A.D., Stolfo, S.J.: Spectrogram: A mixture of Markov chains model for anomaly detection in web traffic. In: Proceedings of the Network and Distributed System Security Symposium (2009)

9. AV-test, http://www.av-test.org/
10. Hu, X., Chiueh, T.-C, Shin, K.G.: Large-scale malware indexing using function-call graphs. In: ACM Conference on Computer and Communications Security, pp. 611–620 (2009)
11. Song, Y., Locasto, M.E., Stavrou, A., Keromytis, A.D., Stolfo, S.J.: On the infeasibility of modeling polymorphic shellcode. In: Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS), pp. 541–551 (2007)
12. Collberg, C., Thomborson, C., Low, D.: A taxonomy of obfuscating transformations. In: Technical Report 148. University of Auckland (1997)
13. Detristan, T., Ulenspiegel, T., Malcom, Y., Superbus, M., Underduk, V.: Polymorphic shellcode engine using spectrum analysis, http://www.phrack.org/show.php?p=61&a=9
14. Moore, H.: The metasploit project, http://www.metasploit.com
15. Wang, X., Pan, C.C., Liu, P., Zhu, S.: SigFree: A signature-free buffer overflow attack blocker. In: 15th Usenix Security Symposium (2006)
16. Wang, K., Parekh, J.J., Stolfo, S.J.: Anagram: A Content Anomaly Detector Resistant to Mimicry Attack. In: Zamboni, D., Krügel, C. (eds.) RAID 2006. LNCS, vol. 4219, pp. 226–248. Springer, Heidelberg (2006)
17. Bellman, R.E.: Adaptive Control Processes: A Guided Tour. Princeton University Press (1961)
18. Meyn, S.P., Tweedie, R.: Markov Chains and Stochastic Stability. Cambridge University Press (2005)
19. Aldrich, J.: R.A. Fisher and the making of maximum likelihood 1912-1922. Statistical Science 12, 162–176 (1997)
20. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1–38 (1977)
21. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Cambridge (1999)
22. Macaulay, S.: Admmutate: Polymorphic shellcode engine, http://www.ktwo.ca/security.html
23. Jemiscode: Jemiscodes - a polymorphic shellcode generator, http://www.shellcode.com.ar/en/proyectos.html
24. Wang, X., Jhi, Y.C., Zhu, S., Liu, P.: STILL: Exploit code detection via static taint and initialization analyses. In: Proceedings of Annual Computer Security Applications Conference, ACSAC (2008)
25. Li, Z., Sanghi, M., Chen, Y., Kao, M.Y., Chavez, B.: Hamsa: Fast signature generation for zero-day polymorphic worms with provable attack resilience. In: IEEE Symposium on Security and Privacy (2006)
26. Newsome, J., Karp, B., Song, D.: Polygraph: Automatic signature generation for polymorphic worms. In: IEEE Symposium on Security and Privacy (2005)
27. Chung, S.P., Mok, A.K.: Advanced Allergy Attacks: Does a Corpus Really Help? In: Kruegel, C., Lippmann, R., Clark, A. (eds.) RAID 2007. LNCS, vol. 4637, pp. 236–255. Springer, Heidelberg (2007)
28. Polychronakis, M., Anagnostakis, K.G., Markatos, E.P.: Network–Level Polymorphic Shellcode Detection Using Emulation. In: Büschkes, R., Laskov, P. (eds.) DIMVA 2006. LNCS, vol. 4064, pp. 54–73. Springer, Heidelberg (2006)
29. Gu, B., Bai, X., Yang, Z., Champion, A.C., Xuan, D.: Malicious shellcode detection with virtual memory snapshots. In: INFOCOM, pp. 974–982 (2010)
30. Christodorescu, M., Kruegel, C., Jha, S.: Mining specifications of malicious behavior. In: Proceedings of the 6th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE 2007), pp. 5–14. ACM Press, New York (2007)

31. Preda, M.D., Christodorescu, M., Jha, S., Debray, S.: A semantics-based approach to malware detection. In: Proceedings of the 34th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL 2007), pp. 377–388. ACM Press, New York (2007)
32. Borders, K., Prakash, A., Zielinski, M.: Spector: Automatically analyzing shell code. In: Proceedings of the 23rd Annual Computer Security Applications Conference, pp. 501–514 (2007)
33. Newsome, J., Song, D.: Dynamic taint analysis for automatic detection, analysis, and signature generation of exploits on commodity software. In: Proceedings of Network and Distributed System Security Symposium (2005)
34. Krugel, C., Kirda, E., Mutz, D., Robertson, W., Vigna, G.: Polymorphic Worm Detection Using Structural Information of Executables. In: Valdes, A., Zamboni, D. (eds.) RAID 2005. LNCS, vol. 3858, pp. 207–226. Springer, Heidelberg (2006)
35. Pedro, N.D., Domingos, P., Sumit, M., Verma, S.D.: Adversarial classification. In: 10th ACM SIGKDD Conference On Knowledge Discovery and Data Mining, pp. 99–108 (2004)
36. Kong, D., Jhi, Y.-C., Gong, T., Zhu, S., Liu, P., Xi, H.: SAS: Semantics Aware Signature Generation for Polymorphic Worm Detection. In: Jajodia, S., Zhou, J. (eds.) SecureComm 2010. LNICST, vol. 50, pp. 1–19. Springer, Heidelberg (2010)