Dazzle-attack: Anti-Forensic Server-side Attack via Fail-free Dynamic State Machine

Bora Lee*[1], Kyungchan Lim*[2], JiHo Lee[1], Chijung Jung[1], Doowon Kim[2], Kyu Hyung Lee[3], Haehyun Cho[4], and Yonghwi Kwon[1]

[1] University of Virginia, Virginia, USA
{boralee,jiholee,cj5kd,yongkwon}@virginia.edu
[2] The University of Tennessee, Tennessee, USA
klim7@vols.utk.edu, doowon@utk.edu
[3] University of Georgia, Georgia, USA
kyuhlee@uga.edu
[4] Soongsil University, Seoul, South Korea
haehyun@ssu.ac.kr

Abstract. Server-side malware is one of the prevalent threats that can affect a large number of clients who visit the compromised server. In this paper, we propose Dazzle-attack, a new advanced server-side attack that is resilient to forensic analysis such as reverse-engineering. Dazzle-attack retrieves typical (and non-suspicious) contents from benign and uncompromised websites to avoid detection and mislead the investigation to erroneously associate the attacks with benign websites. Dazzle-attack leverages a specialized state-machine that accepts any inputs and produces outputs with respect to the inputs, which substantially enlarges the input-output space and makes reverse-engineering effort significantly difficult. We develop a prototype of Dazzle-attack and conduct empirical evaluation of Dazzle-attack to show that it imposes significant challenges to forensic analysis.

1 Introduction

Malware analysis is a crucial task in revealing the real intentions and actors behind cyber attacks. Analyzing malware can lead to various forensic evidence, such as what sensitive information the malware wants to leak and to where (e.g., addresses of attacker-controlled servers). In recent years, malware gets more sophisticated in hiding its code using various techniques such as code obfuscation and remote code execution. In response, advanced malware analysis techniques have been proposed: (1) program analysis techniques including symbolic execution and forced execution [21,23,24,28,32,33,35,39,41,44,51,53,59–63,68,78,80] that can uncover hidden malicious logic in malware and (2) deep packet inspection techniques [19,25,46,69] that can see through malicious payloads delivered through network packets. While it is challenging to dissect malware completely,

* Bora Lee and Kyungchan Lim are co-first authors and are listed in alphabetical order.

analyzing behaviors of malware often results in critical hints for triaging the attacker (e.g., via network addresses they connect to).

In this paper, we explore the possibility of creating a forensically stealthy malware. Specifically, we present an anti-forensic attack, dubbed DAZZLE-ATTACK[5]. It collects inputs from multiple *benign and uncompromised websites* that are *not associated with cyber attackers* (e.g., www.npr.org). The input content is ordinary (i.e., not influenced by the attacker), avoiding detection from network packet inspection techniques and leaving no forensic evidence in the network trace. DAZZLE-ATTACK does not include executable malicious code itself, making static analysis based forensic analysis (e.g., anti-virus techniques) ineffective.

The inputs are later used to construct a malicious payload through a special state machine proposed by [37,38], which is carefully designed to make the analysis of DAZZLE-ATTACK inconclusive. Specifically, the state-machine can take any inputs and generate varying outputs depending on inputs, making the input-output space extremely large. As such, DAZZLE-ATTACK evades state-of-the-art malware detection techniques, including reverse-engineering and forensic triaging. We design and implement a set of tools that can create DAZZLE-ATTACK from the following two inputs: (1) malicious code snippet to deliver and (2) a set of benign contents. The created DAZZLE-ATTACK will run the specialized state-machine to convert the predetermined benign contents into the malicious code snippet when all the benign contents appear together.

Our contributions are summarized as follows:

- We propose DAZZLE-ATTACK, an anti-forensic technique that transforms ordinary contents from benign websites to malicious payloads.
- We leverage the concept of ambiguous translator [37,38] for translating input words to malicious payloads to impose challenges in reverse-engineering.
- We implement a set of automated tools to create DAZZLE-ATTACK, including a website crawler, statistical analyzer, and the ambiguous translator generator.
- Our evaluation result shows that DAZZLE-ATTACK is effective in delivering malicious payloads without being analyzed and detected.

Threat model. We assume a forensic/malware analysis scenario. Specifically, we assume that an attacker already compromised a victim server and placed the malware. While the malware might be executed, it did not deliver the malicious payload (i.e., attack) yet. Exploiting servers can be done by leveraging software vulnerabilities [26, 64] in Internet-facing server programs. The exploitation of web servers is out of the scope of this paper, but is typical in advanced cyber attack scenarios [12,29,40,48]. We assume that the victim server may log network requests, but may not know when the malware delivered the malicious payload.

2 Motivating Example

We show the effectiveness of the DAZZLE-ATTACK by following a forensic analyst's perspective. Assume that a forensic analyst finds an instance of DAZZLE-

[5] The name DAZZLE-ATTACK is originated from Dazzle camouflage which is a family of ship camouflage consisted of complex patterns of geometric shapes [67].

ATTACK in a compromised server[6], *before it launches the attack*. Then, he aims to understand (1) the purpose and (2) the actors behind the DAZZLE-ATTACK.

Forensic Analysis of Dazzle-attack. We present four different analysis attempts on DAZZLE-ATTACK, to demonstrate its resilience to forensic analyses.

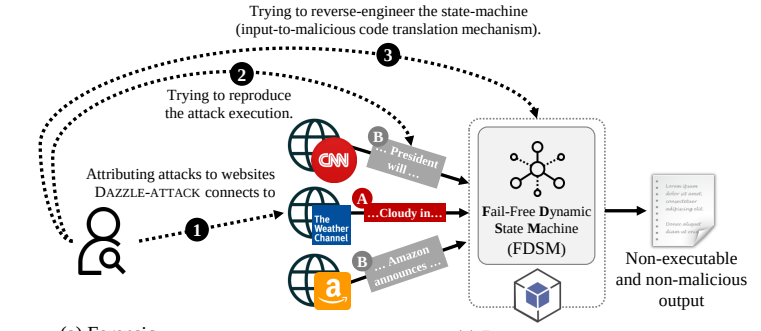
1) Analyzing inputs (i.e., network trace): The forensic investigator obtains available network logs including network packet headers and actual payloads *before the attack happens*. Unfortunately, from the domain names, IP addresses, and the content that DAZZLE-ATTACK has interacted with, the investigator cannot understand what it does. A naive way of attributing the attack to the benign websites, as shown in ❶ in Figure 1, is misleading the analysis.

2) Dynamic analysis: The analyst tries to execute the DAZZLE-ATTACK sample, hoping that it can exercise intended malicious behaviors so that they can be analyzed (❷). However, without knowing the particular input that can trigger the intended attack (which we call *attack triggering input*), the DAZZLE-ATTACK instance does not expose its real intention (e.g., malicious code). Note that DAZZLE-ATTACK will only generate malicious payloads when *the inputs from benign websites are presented as the attacker expected*. If not, it will generate a non-malicious output (i.e., a string that does not look any malicious or another malicious code for obscuring the real objective). For example, in Figure 1, DAZZLE-ATTACK launches an attack (i.e., translates inputs to a malicious payload) when it receives "...Airlines say..." from *www.cnn.com*, "...Cloudy in..." from *www.weather.com*, and "...Amazon recommends..." from *www.amazon.com* as shown in the second row of Figure 1-(d). However, when the analyst executes the program, it obtains "...President will..." (Ⓑ), "...Cloudy in..." (Ⓐ), and "...Amazon announces..." (Ⓑ), where Ⓐ means that the input is a part of attack delivering input and Ⓑ represents a non-attack delivering the input. As a result, the malicious payload (the third row of Figure 1-(d)) is not generated.

3) Static analysis: The analyst tries to use static analysis techniques including symbolic execution to learn the real intention of the DAZZLE-ATTACK instance. However, they suffer from over-approximation. They may obtain a set of all possible inputs and outputs *without a particular order*, which cannot provide a concrete malicious code snippet. Moreover, among the identified outputs, there are no suspicious outputs (e.g., those look like code such as `unlink()`). This is because the state-machine can generate outputs that are different from the annotated outputs of the state machine. In other words, it can generate a malicious code snippet 'fwrite' without having the exact word 'fwrite' annotated in the state machine (Details are elaborated in Section 3.2). As a result, even after exploring millions of possible paths and inputs via symbolic execution tools [15, 28, 34, 57], the attack delivering inputs were not found (Details are described in Section 4).

4) Manual analysis: The analyst manually reads the source code to understand how the input words are translated and what the hidden malicious behaviors are (❸). The analyst collects all possible inputs that can be processed by the state machine and tries to construct inputs hoping it can reveal mali-

[6] The assumption on the compromised servers in cyber attacks is typical [12, 29, 48].



Input	Output
"**Airlines say**", "**Cloudy in**", "**Amazon recommends**", ...	@fopen("tmp");@fwrite(...);...
"President will", "**Cloudy in**", "Amazon announces", ...	Invested @amazonnews more information ...

(d) Input Translation by FDSM

A Content that will deliver the attack B Content that will ***not*** deliver the attack

Fig. 1. Forensic analysis on DAZZLE-ATTACK. Malicious output is not generated without the attack triggering input.

cious payloads. However, he observes it is not possible to establish a one-to-one mapping because the same state transition can be triggered by multiple inputs, which implies he may need to test almost every possible word (due to *dynamic output translation* in Section 3.2).

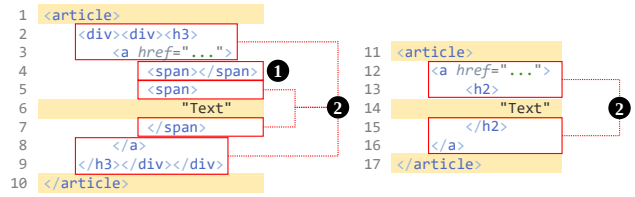
3 Design

To create DAZZLE-ATTACK, we first profile websites to identify candidate input contents (Section 3.1) and then construct the ambiguous translator (Section 3.2).

3.1 Identifying input words via profiling

DAZZLE-ATTACK operates on the contents obtained from benign websites that are *not controllable* by attackers (e.g., a headline news title on *www.cnn.com*). This design choice is crucial to deceive forensic investigators (i.e., hide the identity of attackers). However, uncontrollable inputs might be unreliable because they may have changed when the attack is launched. We mitigate this issue by choosing inputs that are *statistically reliable* via website content profiling. In addition, we propose a DOM path update resilient parsing technique.

3.1.1 Profiling. The website profiler takes a set of web pages as input and generates multiple *InputVector* candidates, consisting of *Input Words* and *Statistics*. Input words are essentially the chosen inputs that make DAZZLE-ATTACK launch the attack (i.e., deliver the malicious payload). It crawls the web pages regularly for a specified period (e.g., every hour for one month by default) so we have multiple snapshots for each page.



(a) Part of the persistent DOM tree of www.cnn.com (July 2019)

(b) Part of the persistent DOM tree of www.cnn.com (December 2020)

Fig. 2. Comparing two persistent DOM trees.

• **Obtaining a persistent DOM tree.** From the crawled website snapshots, we compute a *persistent* DOM tree that only includes elements that appear in *all crawled instances* so that unreliable contents will be excluded.

• **DOM parsing resilient to updates.** We leverage a parsing technique that is resilient to updates in websites' DOM structures. Specifically, we eliminate nodes that do not contain any text. In addition, for an element only contain another element (i.e., a holder element), we consider it can be reduced when it is compared. For instance, " <h3> Text </h3> " and " Text " are considered equivalent in the profiler and DAZZLE-ATTACK. In addition, nodes with empty content will also be ignored as well (e.g., " "). These two techniques handle cosmetic changes in websites. Figure 2 shows an example of two persistent DOM trees obtained from www.cnn.com from (a) July 2019 and (b) December 2020. Observe that they have different DOM trees. However, after we remove tags that do not have any content (❶) and consider placeholders (❷) between the two DOM trees are equivalent, we identify the two are semantically identical (as the highlighted parts are same). This parsing technique is used by all the components of DAZZLE-ATTACK.

• **Reliability of a persistent DOM tree.** To better understand the reliability of the persistent DOM tree, we obtain 5 persistent DOM trees collected from July 2019 (1st to 31st), November 2019 (1st to 30th), March 2020 (1st to 31st), July 2020 (1st to 31st), and November 2020 (1st to 30th) for 10 websites[7]. Then, we compare the five persistent DOM trees of each website. The result shows that 94.7% of the persistent DOM tree's elements reliably appear during the first 4 months. The percentage of reliable elements becomes smaller as time goes: 85.5% for 8 months, 76.3% for 12 months, and 71.3% for 16 months. This suggests that DAZZLE-ATTACK is quite reliable in its first four months. Moreover, to improve the reliability further, we use multiple DOM elements and use them as alternative elements (i.e., backup options) as explained in Section 14.

• **Extracting input words.** For each DOM element in the persistent DOM tree, we extract all words that are longer than three characters. Shorter words (e.g., "for" and "or") are not good candidates in general because they usually do not have much meaning and are too frequently seen across different pages.

[7] www.cnn.com, www.npr.org, www.gnu.org, 19hz.info, techtonic.fm, earthquaketrack.com, news.ycombinator.com, www.kimbellart.org, lite.poandpo.com, chromereleases.googleblog.com

The extracted words are annotated with their positions found in the text of the DOM element. For example, a text "***Episode 976: Terms of Service***" is annotated as follows: "Episode" = 1st, "976:" = 2nd, "Terms" = 3rd, and "Service" = 4th. Observe that the word "of" is *not considered* as its length is *not longer than 3*. At runtime, the position will be used to extract input words.

• **Computing input word statistics.** Given the extracted words, the TextRank algorithm [56] is used to identify frequently observed words among them. The top ten (the number of words is configurable) words from the result are chosen. Then, we compute the statistics of the chosen words. Specifically, for each input word, we calculate (1) coverage, (2) regularity, and (3) distribution of the word during the profiling period.

1) Coverage: This represents the percentage of an input word in a DOM element appearing during the profiling. For example, suppose that we crawl every hour for 5 days, resulting in 120 ($=24*5$) data points. If an input word appears 100 out of 120 data points, its coverage is $\frac{100}{120}$. Intuitively, an input word with a higher coverage has a higher probability of being observed in the future.

2) Regularity: The regularity represents how regularly the content will appear. To compute the regularity, for each input word, we measure the variance of time distances between the two adjacent appearances. Specifically, given N data points, the distances between the adjacent two points, n and n+1, are calculated, resulting N-1 distances: $d_1, d_2, ..., d_{N-1}$. Then, we count the number of unique distances, denoted as $CNT_{\text{unique\_distances}}$. If all the distances are equal (i.e., they regularly appear), the value (i.e., regularity) will be 1.0 (i.e., 100%). We compute the regularity as follows: $1.0 - \frac{CNT_{\text{unique\_distances}}}{N-1}$. Intuitively, an input word with a higher regularity will appear in a more predictable way than an input word with a lower regularity.

3) Distribution: It represents how an input word *appeared evenly* during the profiling period. This is complementary to the regularity because some input words with high regularity may not be evenly distributed. For instance, if a particular input word appears only three times but at the beginning, in the middle, and at the end of the profiling period, it would have a high regularity score. However, it is not well distributed over the period. A higher distribution value means an input word has been observed evenly and frequently during the profiling period. Algorithm 1 shows how we compute the distribution value (DIST). First, given a sequence of N data points (the input D), we divide them into G groups (G = 24 in this paper, line 2) so that each group includes N/G data points (line 8 as D_{sub} represents the group). Then, we count whether the content appears in each group (line 10) and divide it by the value of G (lines 11-12). After that, we multiply G by 2 and repeat the above process (line 13). After we repeat this R times (R = 5 in this paper), we add the computed value for different Gs and divide by R (line 14). To this end, an input word that appears in more groups will have a higher distribution value.

Algorithm 1: Computing distribution

```
Input  : D: a set of data including all the profiled data points,
         N: the number of data points. W: input word.
         R: the number of iterations for distribution computation.
Output: DIST: distribution value for the input word W.
procedure Distribution(D, N, W)
    G ← 24, DIST ← 0, r ← 1
    while r ≠ R do
        i ← 0, D_c ← 0
        while i ≠ G do
            r ← N/G
            n ← i + 1
            D_sub ← D_[i*r, n*r]
            i ← i + 1
            if W ∈ D_sub then
                D_c ← D_c + 1
        DIST ← DIST + D_c/G
        G ← G * 2, r ← r + 1
    return DIST / R
```

3.1.2 Handling unexpected DOM changes. A website may change its DOM structure. As DAZZLE-ATTACK walks on a DOM tree to locate the input words, changes in the DOM structure can affect the input word collecting process.

To handle this problem, we *chain alternative input words* from multiple DOM elements so that DAZZLE-ATTACK can reliably deliver the payload even with unexpected DOM changes. Specifically, for each selected input word, we identify input words from *other DOM elements that always appear together* with the selected input word. The selected alternative DOM elements should *not share* many DOM elements in their DOM paths (e.g., no more than 20%) so that the alternative input words can work when a significant portion of DOM structure (e.g., 80% of the DOM structure) is changed. If there is no such alternative input word within the same webpage, we use input words from other pages. In practice, a website often has many duplicated contents across multiple sub-webpages. For instance, *www.npr.org*'s front page [5] and it's National News Section [6] have identical contents because top news in the "National News" section also appear in the front page. Such contents can be potential alternative input words to tolerate unexpected DOM changes. At runtime, if DAZZLE-ATTACK fails to extract an input word from a webpage (e.g., from the front page) because the DOM element containing the input word cannot be located, it tries an alternative input word on another page (e.g., from the National News Section sub-page). If it fails again, it continues trying the next alternative input words until it succeeds. In this paper, we chain *4 alternative DOM elements* for an input word. The number of alternative DOM elements is configurable.

3.2 Dazzle-attack Creator

Two inputs are required to create DAZZLE-ATTACK: (1) chosen input words from the profiler and (2) payloads to deliver (e.g., source code of existing malware).

3.2.1 Fail-free Dynamic State Machine. The core of DAZZLE-ATTACK is the fail-free dynamic state machine (FDSM). The technique is borrowed from an existing work [37]. It has two distinctive anti-forensic characteristics. First, it is a *fail-free* state machine that *always transits states* regardless of the current state and input, even if the input is *not annotated with the state transitions* (**C1**). In a typical state machine, a state transition only happens when there is a transition that can accept the current input. If not, the state machine will be stuck and fail to make a transition which can be traced by a forensic analyst to infer that the provided input is *not valid*. Second, the output generation rule of FDSM during state transition is *dynamic* (**C2**). This means that the output is changing based on a concrete input at runtime. This significantly enlarges the search space of the possible inputs and outputs.

• **Making transitions on any inputs (C1).** FDSM is designed to make transitions from any state on any inputs. If an input does not match with any possible transitions from the current state, it makes a transition to a state which has a transition condition most similar to the provided input. Specifically, for all next states from the current state, it calculates the distance (by subtracting values from each byte offset) between the current input and the transition conditions. Then, it selects a transition with the smallest distance.

• **Dynamic output translation (C2).** FDSM takes any inputs and generates outputs where each input leads to a *unique output*. When FDSM makes a transition on an input that is not exactly matched with the input of the transition, it changes the output translation rule by applying the differences between the current input and the input annotated on the transition. This makes the output space very large as the output can vary as much as the input varies.

Consider FDSM taking I as input and making a transition T^x where the transition's annotated input and output are denoted as T^x_{IN} and T^x_{OUT} respectively. Now, assume a scenario when an input with no matching transitions from the current state is given. In this case, FDSM makes a transition T^x if the distance (i.e., the sum of the distance between characters) between I and T^x_{IN} is the smallest compare to other transitions' annotated inputs (i.e., T^{others}_{IN}). Moreover, we extend the output space by dynamically changing T^x_{OUT} based on the current input I. Specifically, given I that is different from T^x_{IN}, and assume that the state machine makes T^x transition, instead of generating T^x_{OUT} according to the state machine, we generate an output computed by $T^x_{OUT} - (T^x_{IN} - I)$ on each byte of T^x_{OUT}, T^x_{IN}, and I and '$-$' operator represents subtraction on each byte between the two operands (with the same byte offsets).

3.2.2 Constructing FDSM. First, we create states and transitions for translating the chosen input to the given malicious payload, so that it can generate malicious payload when the predefined attack delivering inputs are provided. We then add dummy states and transitions to connect all states. Note that the dummy states and transitions can also be used to create decoy (i.e., fake) payloads so that it can mislead the forensic analysis. Inputs/outputs of the transitions to the dummy states are chosen in a way that the inputs of all transitions

look similar, making it challenging to know which transitions are for malicious payload generations. Specifically, for each newly added transition, its input is derived by choosing a similar word (i.e., synonyms/antonyms in dictionaries [2, 7]) to its neighboring transition's input.

4 Evaluation

4.1 Reliability of Dazzle-attack

Dazzle-attack takes input from webpages that are not under the control of the attacker, meaning that the reliability of Dazzle-attack's attack is probabilistic. To understand the reliability of Dazzle-attack in practice, we create a mock attack with real-world websites and show the result.

Experiment with real-world websites. To understand how reliably input words will show during an attack, we conduct an experiment from May 2019 to June 2019 (40 days). In particular, we profile 5 websites (Twitter: Houston Rockets, Trinity Church Boston, NASA Image of the Day, eBay, and Oracle Arena)[8] for the first 20 days, and observe the websites for the next 20 days to check whether the input words appear. We select an input word from each of the websites, resulting in 5 input words in total. As shown in Figure 3-(a), during the profiling period, there are about *19 hours that all desired input words appear together* (highlighted). Figure 3-(b) shows the input words that appeared on the websites during the observation period (May 28th, 2019 ~ June 17th, 2019). There are about 4 hours all the input words appeared together. We present two more such experiments on our website [11], showing that creating reliable and stealthy attacks is possible.

4.2 Anti-forensic capability of Dazzle-attack

Datasets for payloads. We collect 573 server-side malware from known malware collection repositories [9, 17, 18, 20, 58, 66, 73, 75–77, 79]. The samples consist of eight types: webshells, backdoors, bypassers, uploaders, spammers, SQL-Shells, reverse shells, and flooders. For each category, we collect a similar number of samples (e.g., 61~79). Details of each type of the samples can be found in Appendix A.1.

Statistics of Dazzle-attack. We generate Dazzle-attack instances for all 573 collected malware samples as shown in Figure 4. We categorize them by the samples' sizes. As shown in Figure 4-(a), the sizes of Dazzle-attack are significantly larger than the original samples (from to 26 to 67 times roughly). To mitigate this significant size increase, we apply compression, e.g., gzip, to reduce the size of Dazzle-attack. Figure 4-(b) shows the sizes after the compression. Except for the first group, the size of Dazzle-attack is about 5 times larger than the original sample. Dazzle-attack in the smallest group is 10 times

[8] https://twitter.com/houstonrockets, https://www.trinitychurchboston.org, https://www.nasa.gov/multimedia/imagegallery/iotd.html, https://www.ebay.com, https://www.theoaklandarena.com

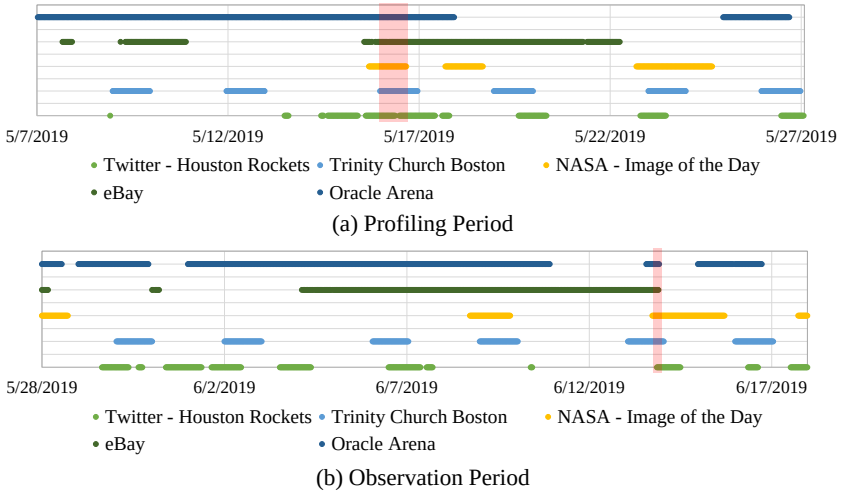


Fig. 3. Input words appearing during the experiment. X-axis represents the date and Y-axis represent the appearance of input words from websites.

bigger than the original sample. However, their sizes are less than 30 KB, which is commonly observed in real-world PHP applications.

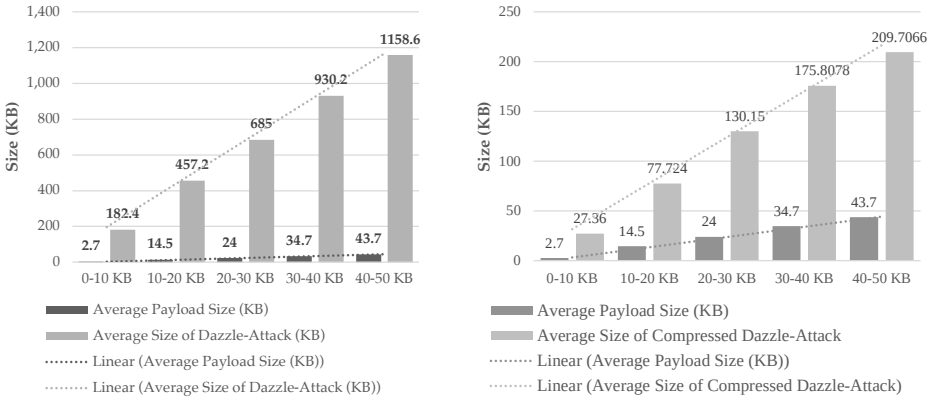
4.3 Comparison with existing obfuscators

We compare DAZZLE-ATTACK with state-of-the-art obfuscation techniques. We prepare two sets of samples: benign samples and malicious samples. We apply existing obfuscators to obtain obfuscated versions of samples. We also create DAZZLE-ATTACK of the samples. Then, we run existing malware detectors to see whether the obfuscated samples and DAZZLE-ATTACK instances are detected.

Obfuscator selection. Four state-of-the-art obfuscators are chosen based on their popularity: PHP Obfuscator [30], YAK Pro [43], Best PHP Obfuscator [1], and Simple Online PHP Obfuscator [47].

Malware detector selection. We use three widely used malware detectors (PHP Malware Finder [72], Linux Malware Detector [4], and Shellray [8]) and a recently released PHP malware scanning tool called MalMax [55] that handles multiple layers of obfuscations and exposes all hidden malicious behaviors of malware. We do not use popular anti-virus software [52,71] because they perform worse than the malware detectors we selected as mentioned in [55].

Malicious sample selection and methodology. From the 573 malware we collected, malware samples that are *not detected by existing malware detectors* are excluded from this experiment. Specifically, PHP Malware Finder identifies 413 samples, Linux Malware Detector flags 185 samples as malware, and Shellray detects 524 samples. To this end, we use the different numbers of samples for experiments with each malware detector. Then, each obfuscator is applied to the samples and obtains obfuscated malicious payloads. Finally, the four malware



(a) Size of Dazzle-Attack (Uncompressed) (b) Size of Dazzle-Attack (Compressed)

Fig. 4. Size of DAZZLE-ATTACK instances.

Table 1. Detection results on malicious and benign samples.

Obfuscator	PHP Mal. Finder		Linux Mal. Detect		Shellray		MalMax	
	Mal.	Benign	Mal.	Benign	Mal.	Benign	Mal.	Benign
PHP Obfuscator [30]	399/413	161/573	98/185	0/573	479/524	0/573	573/573	0/573
YAK Pro [43]	264/413	139/573	16/185	0/573	239/524	1/573	573/573	0/573
Best PHP Obfuscator [1]	412/413	573/573	25/185	0/573	505/524	557/573	573/573	0/573
Simple PHP Obfuscator [47]	413/413	573/573	0/185	0/573	524/524	573/573	573/573	0/573
Dazzle-attack	0/413	0/573	0/185	0/573	0/524	0/573	0/573	0/573

Green cells on 'Mal.' columns indicate that techniques are effective against malware detectors (Detected less than 5%, and lighter green if 5%~50%) while green cells on 'Benign' columns mean that they have no false positives (Lighter green if 5%~50%). Red cells represent the opposite (undesirable) results.

detectors scan all the samples obfuscated by existing obfuscators and DAZZLE-ATTACK instances.

Result for malicious samples. Table 1 shows that all existing malware detection tools are unable to detect DAZZLE-ATTACK (0%), while most of the malware samples obfuscated by the other tools can be detected by at least three detectors: PHP Malware Finder (averagely 89.5%), Shellray (83%), and MalMax (100%). The detection rate of Linux Malware Detector (LMD) is relatively lower than others as LMD is not specifically designed for PHP server-side malware.

Understanding false alarms. Some malware detectors often consider *any obfuscated programs as malicious*, causing high false positive rates. To understand false positive, 573 benign PHP program files from popular PHP programs' codebases (including WordPress [10], Joomla [3], phpMyAdmin [27], and CakePHP [50]) are collected. Initially, none of the 573 benign files are flagged as malware by the existing detectors. However, once they are obfuscated by PHP Malware Finder and ShellRay, we observe many of them are detected as malware (i.e., high false positive rates) by Best PHP Obfuscator [1] and Simple Online PHP Obfuscator [47]. Linux Malware Detector has no false positives, while it misses many PHP malware samples in general (i.e., low true positive rate).

Result from MalMax. MalMax [55] detects all the malicious code hidden by the four existing obfuscators without flagging any benign obfuscated samples. However, MalMax detects none of the Dazzle-attack instances. This is because MalMax focuses on executing all statements without precisely identifying attack triggering inputs. Simply executing all statements of a target is sufficient for analyzing the existing obfuscators but not sufficient for Dazzle-attack.

5 Discussion

Mitigation. To effectively analyze Dazzle-attack, an automated analysis technique specialized for FDSM is needed. In other words, an analysis engine that can understand and explore a finite state machine may reveal malicious payloads hidden in Dazzle-attack. However, due to a large number of states and transitions coupled with the fail-free transition and dynamic output translation (Section 3.2), it still requires significant effort to reverse-engineer Dazzle-attack even with an analysis specialized to a state machine. A real-time detector and monitoring tool can prevent Dazzle-attack from damaging the victim's system. However, it is limited to preventing damages. Investigating Dazzle-attack is still challenging.

Availability of attack delivering inputs for analysts. We assume that an analyst does not know the attack delivering input, and the goal of the analyst is to identify it by analyzing the malware. If an analyst knows the attack delivering inputs, malicious payload hidden in Dazzle-attack can be exposed by executing it directly. Note that, in practice, server-side network logs only include HTTP requests (without contents of HTTP responses) because logging contents will increase the log size significantly.

6 Related Work

Advanced malware analysis. A group of research [13, 14, 16, 22, 28, 36, 39, 42, 45, 49, 54, 65, 68, 70, 74] tries to detect and analyze malware. In particular, a dynamic analysis based forced execution technique [41] aims to handle evasive JavaScript malware. They forcibly drive execution into every branch even if the branch condition is not satisfied. While they are effective in detecting malware that hides malicious code behind sophisticated predicates, it is not effective in exposing malicious payload in Dazzle-attack because it is encoded as states and transitions of the FDSM. Moreover, there are static, symbolic [31], and fuzzing tools for malware analysis that can reveal malicious behaviors. As discussed in Section 4, Dazzle-attack is resilient to such malware analysis techniques.

Network traffic based analysis. There are also forensic analysis and network traffic analysis approaches that analyze the causal relationship between network and system events [14, 74]. For such techniques, Dazzle-attack is difficult to analyze as it gets all inputs from common benign websites where many other applications and systems may access them when Dazzle-attack is active. As a result, understanding who are the actors behind the attack is particularly

challenging. There are approaches that detect common patterns of malware [39, 42]. While they are effective in traditional malware, DAZZLE-ATTACK can evade such techniques as DAZZLE-ATTACK can be implanted into existing programs.

7 Conclusion

In this paper, we present DAZZLE-ATTACK, a new type of attack that secretly delivers malicious payloads while imposing fundamental challenges to post-mortem forensic analysis. We leverage FDSM that effectively thwarts various forensic analysis attempts. Our evaluation shows that DAZZLE-ATTACK is highly effective in preventing forensic analysis.

Acknowledgement. We thank the anonymous referees for their constructive feedback. The authors gratefully acknowledge the support of NSF 1916499, 1908021, 1850392, 2145616, and 2210137. This research was partially supported by Science Alliance's StART program, National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A4A102 9650), and gifts from Cisco Systems and Google exploreCSR. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

References

1. Best PHP Obfuscator (2018), http://www.pipsomania.com/best_php_obfuscator.do
2. A text file containing 479k English words (2019), https://github.com/dwyl/english-words
3. Joomla: Content Management System (CMS) (2019), https://www.joomla.org/
4. Linux Malware Detect (2019), https://www.rfxn.com/projects/linux-malware-detect/
5. NPR: National Public Radio (2019), https://npr.org/
6. NPR: News and National Top Stories (2019), https://npr.org/sections/national/
7. PHP: Pspell Functions (2019), https://www.php.net/manual/en/ref.pspell.php
8. Shellray: A PHP webshell detector (2019), https://shellray.com/
9. VirusShare (2019), https://virusshare.com/
10. WordPress (2019), https://wordpress.com/
11. Dazzle-Attack: Supplementary Materials (2020), https://sites.google.com/view/dazzle-attack-additional/home
12. Agency, C..I.S.: Russian State-Sponsored Advanced Persistent Threat Actor Compromises U.S. Government Targets (2020), https://us-cert.cisa.gov/ncas/alerts/aa20-296a
13. Anderson, H.S., Kharkar, A., Filar, B., Evans, D., Roth, P.: Learning to evade static pe machine learning malware models via reinforcement learning. arXiv preprint arXiv:1801.08917 (2018)

14. Aqil, A., Atya, A.O.F., Jaeger, T., Krishnamurthy, S.V., Levitt, K., McDaniel, P.D., Rowe, J., Swami, A.: Detection of stealthy tcp-based dos attacks. In: MILCOM 2015-2015 IEEE Military Communications Conference. pp. 348–353. IEEE (2015)
15. van Arnhem, B.: phpscan: Symbolic execution inspired PHP application scanner for code-path discovery (2017), https://github.com/bartvanarnhem/phpscan
16. Balzarotti, D., Cova, M., Felmetsger, V., Jovanovic, N., Kirda, E., Kruegel, C., Vigna, G.: Saner: Composing static and dynamic analysis to validate sanitization in web applications. In: 2008 IEEE Symposium on Security and Privacy (S&P). pp. 387–401. IEEE (2008)
17. Bart, P.: Php-backdoors: A collection of php backdoors
18. BDLeet: public-shell: Some Public Shell (2016), https://github.com/BDLeet/public-shell
19. Becchi, M., Crowley, P.: A hybrid finite automaton for practical deep packet inspection. In: Proceedings of the 2007 ACM CoNEXT conference. p. 1. ACM (2007)
20. BlackArch: webshells: Various webshells (2019), https://github.com/BlackArch/webshells
21. Cadar, C., Dunbar, D., Engler, D.R., et al.: Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs. In: OSDI. vol. 8, pp. 209–224 (2008)
22. Christodorescu, M., Jha, S., Seshia, S.A., Song, D., Bryant, R.E.: Semantics-aware malware detection. In: 2005 IEEE Symposium on Security and Privacy (S&P). pp. 32–46. IEEE (2005)
23. Dahse, J., Schwenk, J.: Rips-a static source code analyser for vulnerabilities in php scripts. Retrieved: February **28**, 2012 (2010)
24. designsecurity: progpilot: A static analysis tool for security (2016), https://github.com/designsecurity/progpilot
25. Dharmapurikar, S., Krishnamurthy, P., Sproull, T., Lockwood, J.: Deep packet inspection using parallel bloom filters. In: 11th Symposium on High Performance Interconnects, 2003. Proceedings. pp. 44–51. IEEE (2003)
26. Erdődi, L., Jøsang, A.: Exploitation vs. prevention: The ongoing saga of software vulnerabilities. Acta Polytechnica Hungarica **17**(7) (2020)
27. Fauth, M.M.: phpmyadmin: A web interface for MySQL and MariaDB (2019), https://github.com/phpmyadmin/phpmyadmin
28. Filaretti, D., Maffeis, S.: An executable formal semantics of php. In: European Conference on Object-Oriented Programming. Springer (2014)
29. FIREEYE: APT41: Double Dragon, a dual espionage and cyber crime operation (2019), https://content.fireeye.com/apt-41/rpt-apt41
30. Fonk, M.: php-obfuscator: A parsing PHP obfuscator (2019), https://github.com/naneau/php-obfuscator
31. Fratantonio, Y., Bianchi, A., Robertson, W., Kirda, E., Kruegel, C., Vigna, G.: Triggerscope: Towards detecting logic bombs in android applications. In: 2016 IEEE symposium on security and privacy (SP). pp. 377–396. IEEE (2016)
32. Grimes, H.Y.: Eir - static vulnerability detection in php applications (2015)
33. Hauzar, D., Kofroň, J.: Weverca: Web applications verification for php. In: International Conference on Software Engineering and Formal Methods. pp. 296–301. Springer (2014)
34. Jensen, T., Pedersen, H., Olesen, M.C., Hansen, R.R.: Thaps: automated vulnerability scanning of php applications. In: Nordic conference on secure IT systems. pp. 31–46. Springer (2012)

35. Jovanovic, N., Kruegel, C., Kirda, E.: Pixy: A static analysis tool for detecting web application vulnerabilities. In: 2006 IEEE Symposium on Security and Privacy (S&P). pp. 6–pp. IEEE (2006)
36. Jovanovic, N., Kruegel, C., Kirda, E.: Static analysis for detecting taint-style vulnerabilities in web applications. Journal of Computer Security **18**(5), 861–907 (2010)
37. Jung, C., Kim, D., Chen, A., Wang, W., Zheng, Y., Lee, K.H., Kwon, Y.: Hiding critical program components via ambiguous translations. In: 2022 IEEE/ACM 44rd International Conference on Software Engineering (ICSE). IEEE (2022)
38. Jung, C., Kim, D., Wang, W., Zheng, Y., Lee, K.H., Kwon, Y.: Defeating program analysis techniques via ambiguous translation. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 1382–1387. IEEE (2021)
39. Kapravelos, A., Shoshitaishvili, Y., Cova, M., Kruegel, C., Vigna, G.: Revolver: An automated approach to the detection of evasive web-based malware. In: Presented as part of the 22nd USENIX Security Symposium. pp. 637–652 (2013)
40. Kasturi, R.P., Sun, Y., Duan, R., Alrawi, O., Asdar, E., Zhu, V., Kwon, Y., Saltaformaggio, B.: TARDIS: rolling back the clock on cms-targeting cyber attacks. In: 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020. pp. 1156–1171. IEEE (2020). https://doi.org/10.1109/SP40000.2020.00116, https://doi.org/10.1109/SP40000.2020.00116
41. Kim, K., Kim, I.L., Kim, C.H., Kwon, Y., Zheng, Y., Zhang, X., Xu, D.: J-force: Forced execution on javascript. In: Proceedings of the 26th international conference on World Wide Web. pp. 897–906. International World Wide Web Conferences Steering Committee (2017)
42. Kinder, J., Katzenbeisser, S., Schallhart, C., Veith, H.: Detecting malicious code by model checking. In: International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. pp. 174–187. Springer (2005)
43. Kissian, P.: YAK Pro: Php Obfuscator (2019), https://www.php-obfuscator.com/
44. Kneuss, E., Suter, P., Kuncak, V.: Phantm: Php analyzer for type mismatch. In: FSE'10 Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering. No. CONF (2010)
45. Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C., Roli, F.: Adversarial malware binaries: Evading deep learning for malware detection in executables. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 533–537. IEEE (2018)
46. Kumar, S., Dharmapurikar, S., Yu, F., Crowley, P., Turner, J.: Algorithms to accelerate multiple regular expressions matching for deep packet inspection. In: ACM SIGCOMM Computer Communication Review. vol. 36, pp. 339–350. ACM (2006)
47. Lie, R.: Simple online PHP obfuscator: encodes PHP code into random letters, numbers and/or characters (2019), https://www.mobilefish.com/services/php_obfuscator/php_obfuscator.php
48. Magazine, C.: New Report Reveals Chinese APT Groups May Have Been Entrenched in Some Servers for Nearly a Decade Using Little-Known Linux Exploits, CPO Magazine (2020), https://www.cpomagazine.com/cyber-security/new-report-reveals-chinese-apt-groups-may-have-been-entrenched-in-some-servers-for-nearly-a-decade-using-little-known-linux-exploits/

49. Mao, J., Bian, J., Bai, G., Wang, R., Chen, Y., Xiao, Y., Liang, Z.: Detecting malicious behaviors in javascript applications. IEEE Access **6**, 12284–12294 (2018)
50. Masters, L.: CakePHP: The Rapid Development Framework for PHP (2019), https://cakephp.org/
51. Medeiros, I., Neves, N.F., Correia, M.: Automatic detection and correction of web application vulnerabilities using data mining to predict false positives. In: Proceedings of the 23rd international conference on World wide web. pp. 63–74. ACM (2014)
52. Microsoft: Microsoft Defender Advanced Threat Protection (2019), https://docs.microsoft.com/en-us/windows/security/threat-protection/microsoft-defender-atp/microsoft-defender-advanced-threat-protection
53. Mirtes, O.: phpstan: PHP Static Analysis Tool (2019), https://github.com/phpstan/phpstan
54. Moser, A., Kruegel, C., Kirda, E.: Exploring multiple execution paths for malware analysis. In: 2007 IEEE Symposium on Security and Privacy. pp. 231–245. IEEE (2007)
55. Naderi-Afooshteh, A., Kwon, Y., Nguyen-Tuong, A., Razmjoo-Qalaei, A., Zamiri-Gourabi, M.R., Davidson, J.W.: Malmax: Multi-aspect execution for automated dynamic web server malware analysis. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. pp. 1849–1866 (2019)
56. Nathan, P.: Pytextrank, a python implementation of textrank for text document nlp parsing and summarization. https://github.com/ceteri/pytextrank/ (2016)
57. Nguyen, H.V., Nguyen, H.A., Nguyen, T.T., Nguyen, T.N.: Auto-locating and fix-propagating for html validation errors to php server-side code. In: Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering. pp. 13–22. IEEE Computer Society (2011)
58. nixawk: fuzzdb: Web Fuzzing Discovery and Attack Pattern Database (2018), https://github.com/nixawk/fuzzdb
59. Nunes, P.J.C., Fonseca, J., Vieira, M.: phpsafe: A security analysis tool for oop web application plugins. In: 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (2015)
60. Olivo, O.: TaintPHP: Static Taint Analysis for PHP web applications (2016), https://github.com/olivo/TaintPHP
61. OneSourceCat: phpvulhunter: A tool that can scan php vulnerabilities automatically using static analysis methods (2015), https://github.com/OneSourceCat/phpvulhunter
62. Papagiannis, I., Migliavacca, M., Pietzuch, P.: Php aspis: using partial taint tracking to protect against injection attacks. In: 2nd USENIX Conference on Web Application Development. vol. 13 (2011)
63. Peng, F., Deng, Z., Zhang, X., Xu, D., Lin, Z., Su, Z.: X-force: force-executing binary programs for security applications. In: 23rd USENIX Security Symposium. pp. 829–844 (2014)
64. Piantadosi, V., Scalabrino, S., Oliveto, R.: Fixing of security vulnerabilities in open source projects: A case study of apache http server and apache tomcat. In: 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST). pp. 68–78. IEEE (2019)
65. Preda, M.D., Christodorescu, M., Jha, S., Debray, S.: A semantics-based approach to malware detection. ACM SIGPLAN Notices **42**(1), 377–388 (2007)
66. Ridter: Pentest (2019), https://github.com/Ridter/Pentest
67. Ruslan Budnik: The Fantastic Idea of Dazzle Camouflage (2019), https://www.warhistoryonline.com/instant-articles/dazzle-camouflage.html

68. Saxena, P., Akhawe, D., Hanna, S., Mao, F., McCamant, S., Song, D.: A symbolic execution framework for javascript. In: 2010 IEEE Symposium on Security and Privacy. pp. 513–528. IEEE (2010)
69. Sherry, J., Lan, C., Popa, R.A., Ratnasamy, S.: Blindbox: Deep packet inspection over encrypted traffic. ACM SIGCOMM Computer communication review **45**(4), 213–226 (2015)
70. Shu, X., Yao, D., Ramakrishnan, N.: Unearthing stealthy program attacks buried in extremely long execution paths. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. pp. 401–413. ACM (2015)
71. Symantec: Norton™ - Antivirus & Anti-Malware Software (2019), https://us.norton.com/
72. Systems, N.: GitHub - nbs-system/php-malware-finder: Detect potentially malicious PHP files (2019), https://github.com/nbs-system/php-malware-finder/
73. tanjiti: webshellSample: Webshell sample for WebShell Log Analysis (2018), https://github.com/tanjiti/webshellSample
74. Taylor, T., Hu, X., Wang, T., Jang, J., Stoecklin, M.P., Monrose, F., Sailer, R.: Detecting malicious exploit kits using tree-based similarity searches. In: proceedings of the Sixth ACM Conference on Data and Application Security and Privacy. pp. 255–266. ACM (2016)
75. tennc: webshell: A webshell open source project (2019), https://github.com/tennc/webshell
76. Troon, J.: php-webshells: Common php webshells (2016), https://github.com/JohnTroony/php-webshells
77. tutorial0: WebShell: WebShell Collect (2016), https://github.com/tdifg/WebShell
78. vimeo: psalm: A static analysis tool for finding errors in PHP applications (2019), https://github.com/vimeo/psalm
79. xl7dev: WebShell: Webshell & Backdoor Collection (2017), https://github.com/xl7dev/WebShell
80. Yang, Q.: Taint-em-All: A taint analysis tool for the PHP language (2019), https://github.com/quanyang/Taint-em-All

A Appendix

A.1 Payload types

A webshell is malware that enables attackers to access a compromised server via a web browser that acts like a command-line interface. Backdoor is used to provide remote access to an infected machine for attackers. Bypassers are used to avoid detections of local or remote security mechanisms (e.g., firewalls). Uploaders are used to remotely inject additional malware into victim machines. Spammers compose and send spoof/spam emails. SQLShells allows remote attackers to access databases of compromised servers, similar to webshells. A reverse shell is a type of shell that communicates back to the attacker's machine from a victim's machine. Flooders are used to launch Denial of Service (DoS) attacks by sending an excessive number of network packets.