Detection of drive by download based on URL and domain information

Agbefu Ralph Edem ,[†1] Yoshiaki Hori [†1] and Kouichi Sakurai[†1]

Web pages that host drive by download exploits have become a popular means by which an attacker delivers malicious contents onto computers across the internet. In a drive by download attack, an attacker embeds a malicious script into a web page. When a user visits this webpage, the malicious code is executed and attempts to exploit any browser or plug-in vulnerability. To deal with the problem of drive by download attacks, we propose a rule based scoring method for detecting and proactively blacklisting such websites based on the domain information. In this paper we analyze URL and domain information characteristics of drive by download pages, this include: IP address, registrant, country, domain creation date, domain update date and domain expiration date.

1. Introduction

1.1 Background

As the Internet continues to provides services such as entertainment and for communicating, it is also attracting an increasing number of attacks. The user's computer seems to be the weakest link in these kinds of transactions. Usually personal computers contain a number of applications that are rarely updated. The goal of an attacker will be to identify such vulnerabilities and exploit the vulnerabilities. This is achieved by the use of the so called Drive by download attack technique. In a drive by download attack, an attacker embeds a malicious script into a web page. When a user visits this webpage, the malicious code is executed and attempts to exploit any browser or plug-in vulnerability. Malicious execution and exploiting of vulnerabilities are done without the users consent. Drive by download attacks have been on the rise in the last few years.

†1 Department of Informatics, Kyushu University

1.2 Drive by download attack

Drive by download attacks have been pervasive over the last years. Recently antivirus vendor Trend Micro has recently detected a drive-by download attack on well known benign pages such as Facebook that uses malicious advertisements to infect users with malware[1)] . According to Sophos[2)], drive by download is the favorite web threat being used cybercriminals. Websites that are well maintained have also been compromised and injected with malicious content[3)]. The goals of a drive by download attack include: taking effective, temporary control of the victim's web browser

The following are the steps involved in a drive by download attack

(1) Initially an attacker injects malicious code into web server.
(2) The victim visits compromised website. The web server sends the requested page along with injected malicious scripts.
(3) The malicious script causes redirection from one web server to another .
(4) After a number of n redirects. The victim is finally directed to the exploit server, which sends the exploit.
(5) On execution of such exploit, the attacker gains control of the victims browser
(6) Browser is then instructed to visit malware distribution sites
(7) Malwares are downloaded download and executed

The processes involved is further illustrated in the **Fig. 1**:

2. Related works

Researchers have developed a number of detection mechanisms for drive by download attacks. In this section we discuss some of the available detection methods. The detection methods can be grouped into static analysis system and dynamic analysis system.

Static analysis as the name suggests, approaches detection of drive by download attack by analyzing static aspects of a web page. This uses information obtained from the HTML content, URL information and static analysis of JavaScript to detect attacks. We look at a number of proposed static analysis systems. Prophiler[4)], static based systems uses a total of 77 features extracted from web pages to detect drive by download launching sites. These features were obtained

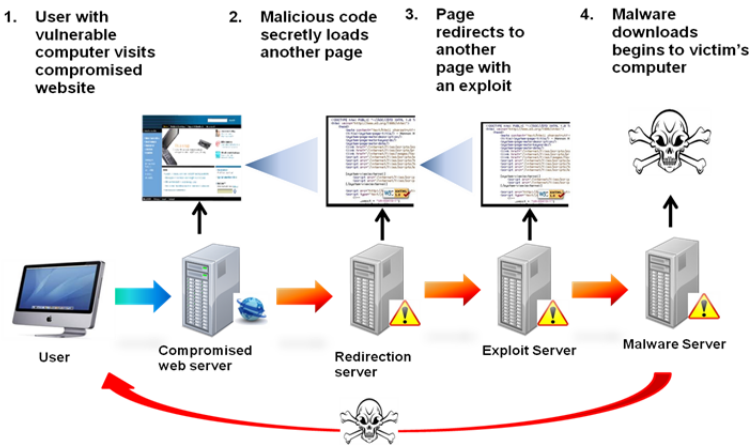

Fig. 1 Drive by download attack

from a web pages URL, HTML content and JavaScript information.Our proposal is different from Prophiler in the sense we consider strictly the URL and domain information in our system. Fukushima et al[5] proposed a way of blacklisting drive by download attacks based on IP address block and the registrar information. Our proposal although similar to[5], is quite different in that we include extra URL and domain information such as domain registration date to detect and blacklist drive by download attacks. Hence our proposal seeks to improve on the existing works.

Dynamic analysis uses client honey pot to detect whether a page is malicious or not. Client honey pot are divided into two kinds: high-interaction client honey pot and low-interaction client honeypot. High-interaction client honey pots[6] uses traditional browsers that run in a controlled or monitored environment. However, low-interaction client[7] emulates the browsers. Dynamic analysis, although, an effective way of detecting drive by download attacks do require a lot of resource. Our proposal which is based on static analysis although effective does not use up as much resource compared to its dynamic counterparts.

3. Motivation and Contribution

Detection and subsequent blacklisting of malicious pages including drive by download links are usually based on the traditional way, which is to add known to malicious web pages to a blacklist database. This way of blacklisting is known as reactive blacklisting. With reactive blacklisting however, unknown malicious won't be detected and hence wont be blacklisted. There is the need for a proactive way of detecting malicious pages, as attackers tend to regularly change URLs and domain information to make it difficult to with a reactive kind of blacklisting. This limits the capability of reactive blacklisting.

In this paper, we propose a rule based scoring system for proactively detecting and blacklisting of malicious pages based on IP address and domain information. Our approach is based on the fact that certain IP address range and domain are constantly abused by attackers.

4. Our Approach

The goal of this paper is to detect and blacklist drive by download attacks proactively. To this end, we observe IP address and domain information. We obtain information about drive by download URLs, such information include:

(1) IP Address
(2) Registrant
(3) Country
(4) Domain registration and Expiry dates
(5) AS Number

Whereby registrant refers to the owner of the malicious page, country refers to the origin of the malicious page. The country of origin is determined by looking up the domain information in GeoIP[13]. Domain registration and expiry dates refer to the day the domain was started and when the lease for it dies out. We evaluate the obtained features and determine which to be included as candidates for detecting of drive by download attacks. We devise a rule-based scoring system similar to the one implemented in SpamAssasin[8], for the detection and blacklisting of drive by download pages. In such a system, rules are applied to certain input under examination. A rule here refers to what action the system

is to take on encountering a malicious feature. For each feature that is determined to be malicious a numerical value is assigned,called the score value. A match of a rule result in an increase of the score value for that associated . If the sum of the score value passes a predetermined threshold, the URL under consideration is flagged and subsequently blacklisted. For our work we chose a threshold value of 5.0 at which a time a web page is marked as malicious. In this system we ensure no one feature leads to the threshold value. That is more than one feature is needed to for a page to flagged as malicious.

4.1 Dataset

The dataset for our proposal were obtained from three different sources. The sources are: MDL[9], malware domains[10], malware URL[11]. We use three sources for our work compared to previous work that used a single source[5]. These sites provide a database of malicious pages and are updated daily. From these sources we obtained a 1000 malicious. However, since our work considers pages related only to launch drive by download attacks, we proceed to prune the database. Description of exploit, redirect or driveby are used as filter URLs to group links as performing a drive by download attack. As a result of the pruning of the database we ended up with 110 malicious URL that launches drive by download attacks

4.2 Feature analysis

Based on the collected URLs we get to know how frequently or not a feature is used. The objective here is to determine properties that are frequently used by attackers as they seek to hide their pages from detection systems.

4.2.1 IP address

We discuss the characteristics of the IP addresses of the malicious links. From our results, we observed cases of a number of IP addresses assigned to one host name and vice versa. For privacy reasons, we avoid the usage of IP addresses, and replaced them with letters. That is a letter 'A' in our result might for example refer to the IP address: 1.1.1.1. Below is a table showing the breakdown of the IP addresses.

4.2.2 AS Number

The Autonomous System (AS) numbers are used by various routing protocols. AS numbers are assigned to the regional registries by the IANA[12]. For the same reasons of privacy, we present AS Numbers with letters. In Table 2. is a sample result which shows a some AS numbers and their frequency of usage.

Table 1 Distribution of IP addresses

IP address	Count
A	15
B	8
C	5
D	4
E	3
F	3
Others	65

Table 2 Distribution of AS Number

AS Number	Frequency
U	11
V3	6
W	4
X	3
Y	2
Z	2

4.2.3 Country

Country in this sense refers to the country where the domain is hosted.We can determine such information with the help of GeoIP[13]. Table below shows the distribution of drive by download sites. The results indicate that most of the URLs are located in Russia and USA.

Table 3 Geographical Distribution

Country	Drive by download URLs	Percentage
Russia	20	18%
USA	20	18%
DominicanRep.	18	16%
Germany	14	13%
Ukraine	8	7%
Australia	7	6%
Others	25	22%

4.2.4 Domain Registration and expiry date

Our results show that most of the drive by download sites have recent domain registration dates. This is due to the fact that once a drive by download site is detected and blocked, attackers create new sites and hence the reason for a greater number of the sites having recent registration dates. We use the day of domain creation compared with the detection date to determine how recent the page is. Attackers realizing that there is a a high likelihood of them been blocked register domain for very short periods. Hence a greater number of the URLs have duration of 1 year.

Table 4 Domain Duration

Domain Duration	Percentage
$1year$	81%
$2-5years$	6%
$6-10years$	9%
$>10years$	4%

Table 5 Domain Registration Date

Domain Registration Date	Percentage
$Lessthan1week$	57%
$1week-amonth$	17%
$1month-1year$	9%
$>1years$	7%

4.2.5 Registrant

The registrant refers to the owner of the domain. That is the one who registers the domain in his or her name. Our result on registrant, shows attackers usually have more than one malicious site registered in their names.

4.3 Candidate for evaluation

On examining the features , we decide on which particular features can be used in our proposal. The features as stated previously are: AS number, IP address, Registrant, Country, Domain registration date and domain expiration date. AS number has been found not be good way of evaluating malicious URL[5]. This is due to the fact that AS contains different number of IP address blocks. Thus simply considering AS as a factor will result in a number of benign web pages been classified as malicious leading to a large false positive. Similarly using only IP address will severely limit the detection capability of our proposal, we decide to use the IP address block as the factor to be used. Since certain parts of the world seem to be renowned as contributing significantly to drive by download URLs, we also decide to use the geographical location as a candidate. Based on the results obtained, we notice certain registrant seem to be quite common to the drive by download sites, hence registrant is also considered as a candidate for evaluation in our proposal.

5. Design of our proposal

In order to implement our proposal, we devise a rule-based scoring system for the detection. In such a system, rules are applied to certain input under our consideration. A match of the rule result in an increase of the score for that associated URL. If the sum of the scores passes a predetermined threshold, the input under consideration is flagged and subsequently blocked. For the rule-based system we chose a threshold value of 5.0 at which a time a web page is marked as malicious. In this system we ensure no one feature leads to the threshold value. That is more than one feature is needed to for a page to flagged as malicious.

5.1 Score values of features

We assign score values to the candidate features and give reasons for such an assignment.

5.1.1 Geographical Distribution

Our results show that countries such as USA, Russia and Germany are fairly likely to have a drive by download site hosted there. We assign a score of 1.5 for countries that contribute more than 15% of the malicious sites collected, else a score of 1.0 is assigned.

5.1.2 Domain registration date and expiry date

We have 81% of domains that have a registration duration of less than a year. This is a strong indication that this is a very popular feature used by attackers, hence we assigned a value of 3.0 to this feature. Furthermore, by observing the date of domain registration, we notice most of these sites were created during the last six months. As a result we assign a value of 2.0 to this feature.

5.1.3 IP address block

Our analysis shows a certain range of IP address blocks are utilised frequently by attackers. Address ranges such as 195.162.a.b and 217.107.x.y are examples of such IP address ranges. For this feature we assign a score of 2.0.

5.1.4 Registrant

The registrant seems to be a very common element in our database. That means for an encountered malicious page, the registrant of that page is also likely to be owner of another malicious page. Hence we assign a score value of 2.0 this feature.

6. Evaluation

For a webpage with an IP address within the range identified to be malicious we assign a socre value of 2.0 to it. If it belongs to a registrant enlisted a malicious, a further value of 2.0 is assigned. If the webpage is found to originate from a country known to be popular for drive by downloads, that is to specific, if the country contributes more than 15% of total malicious URLs, a score value of 2.0. Otherwise no score value is assigned. Continuing, if the domain registration date of the webpage has a duration of a year, a score value of 3.0 is assigned. Depending on how recent the webpage is score values are also assigned. If the webpage happens to have been created within the past 6 months, a score value of 2.0 is assigned. A total value of 5.0 means the will be treated as a drive by download page and blocked, else it is a benign page.

To evaluate the effectiveness and performance of our proposal, we collect a set of malicious ULR known to trigger URL to find out how much of it will be detected and blacklisted based on our approach. We collect a database of 50 malicious URL from the sources previously mentioned. This database of 50 malicious are URLs that are known to launch drive by download attacks. We then apply our rule-based scoring system on them. Using our rule-based system, we were able to detect 42 URLs to be malicious out of the total 50 to be malicious. This represents 84% of the total database. From this results we argue that our proposal is capable of detecting and blacklisting drive by download sites.

7. Conclusion and Future work

We discuss the limitations and future work. We use a database of 110 malicious URLs to come up with the features we used in our work, this is however not near enough to discover characteristics of malicious sites that launch drive by download attack. Our work is also susceptible to evasion attack. That is an attacker can decide not use one or two of the stated features and hence resulting in a malicious page not detected. For example just by changing the duration of a domain, it is possible for an attacker to evade our detection system. We did not implement our detection system to known benign pages, hence we cannot determine the false positive generated on benign pages using our proposed system. As future work , we should look at increasing our database to reflect the real world scenario. This will result in adjusting of the currently assigned score values. It is also necessary for the system to implemented on known benign pages.

In conclusion, threats that use drive by download as a way of infecting usersź is on the rise. There is the need for an effective system to be developed to tackle this problem. We propose a rule-based scoring system which detects and blacklists malicious pages based on URL and domain information.

Acknowledgement

This work is partially supported by Grants-in-Aid for Scientific Research (C) (21500078), Japan Society for the Promotion of Science (JSPS). The first author of this research is supported by MEXT scholarship.

References

1) TrendMicro, Facebook Malvertisement Leads to Exploits, Oct. 2011 Available: http://blog.trendmicro.com/facebook-malvertisement-leads-to-exploits/
2) Ericka Chickowski, Mass SQL Injection Attack Hits 1 Million Sites, Oct. 2011 Available: http://www.darkreading.com/database-security/167901020/security/news/231901236/mass-sql-injection-attack-hits-1-million-sites.html/
3) ZDNet Australia, The death of trusted websites, http://www.zdnet.com.au/insight/security/soa/
4) D. Canali, M. Cova, G. Vigna and C. Kruegel. 'Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages,' *In proceedings of WWW 2011*,2011, pp. 197-206.

5) Y. Fukushima, Y. Hori and K. Sakurai. 'Proactive Blacklisting for Malicious Web Sites by Reputation Evaluation Based on Domain and IP Address Registration,' *In Proceedings of International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11*, 2011, pp. 352-361.
6) T.H Project. (2008, September) Capture-hpc.
Available: https://projects.honeynet.org/capture-hpc
7) Honeyc Project. 2008.
Available: https:
projects.honeynet.org/honeyc
8) The Apache Foundation, *SpamAssasin*, June 2009.
Available: http://spamassasin.apache.org
9) Malware Domain List, Jan. 2012.
Available: http://www.malwaredomainlist.com/mdl.php
10) Malware Domains, Jan. 2012.
Available: http://www.malwaredomains.com/
11) Malware URL, Jan. 2012
Available: http://www.malwareurl.com/
12) Internet Assigned Numbers Authority, *IANA*
Available: http://www.iana.org/
13) GeoIP,
Available: http://www.geoiptool.com/