DETECTING MALICIOUS PROCESSES THROUGH END-POINT DNS-MONITORING

D Rajendra Dev [1], P Amrutha [2], P Bandhavi [3], P S S lakshmi [4] and R Dharani [5]

Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India.

Abstract.

DNS is a foundational internet service which reduces the complexity of remembering random strings of numbers (IP addresses) by converting them to domain names and controls which server the end user will reach. This makes DNS service an enticing threat vector to perpetrators. Though monitoring network DNS has served as a useful way to counter such malicious attacks, changing scenarios demand contemporary solutions. As the attackers have upgraded their game to bypass the rigorous monitoring and furthermore with the recent support for encrypted DNS queries. This provides the attackers an easier means to hide from network-based DNS traffic monitoring. Therefore, we need an end-point DNS monitoring system. URLs are the access points in the user end. Attackers often use URLs to advertise scams or propagate malware. Miscreants often register these domains "just in time", right before an attack. Our analysis yields many findings that may ultimately be useful for early detection of malicious domains

Introduction

One of the core features of the Internet is the Domain Name System (DNS). A crucial function is provided by the Domain Name System (DNS), the Internet's lookup tool for mapping names to IP addresses. Unfortunately, this also makes it possible for attackers to lead users to websites hosting scams, malware, and other harmful content. Since malicious users can conceal their network footprint behind short-lived domain names and perform various attacks (like command and control or fast-flux), the security research community has consistently concentrated on how to identify between domains that are valid and those that are engaged in malicious activity such as fast-flux networks, bot networks, DGA domains, and spam networks.

To lessen these risks, network administrators seek to establish a reputation for each domain based on the possibility that the domain hosts malware, phishing, or other types of attacks. The velocity at which new names are created makes it difficult to establish a reputation for them rapidly. Over tens of thousands of new domains are registered daily. Current DNS reputation systems make use of a resolver's DNS lookup properties.

The Domain Name System (DNS), the Internet's lookup service for mapping names to IP addresses, is critical for many applications Unfortunately, it also allows attackers to direct victims to websites hosting scams, malware, and other malicious content. To mitigate these threats, network operators attempt to assign a reputation to each domain based on the likelihood that the domain is associated with a specific type of attack (e.g., scam, phishing, malware hosting). The rate at which new domains appear makes establishing a reputation for these domains particularly difficult Existing DNS reputation systems rely on the characteristics of DNS lookups performed by resolvers. Existing DNS reputation systems distinguish legitimate from malicious domains based on the characteristics of DNS lookups from resolvers that look up a domain. Unfortunately, these systems must observe many DNS lookups before determining a domain's reputation, which occurs only after a compromise has occurred. To aid in the detection of malicious domains prior to an attack, we examine and characterize the initial DNS activity for each domain.

How the observable behaviour of a malicious domain differs from that of a legitimate domain. We investigate two aspects of domain-related initial DNS behaviour (1) the DNS infrastructure used to resolve domains to IP addresses, and (2) the DNS lookup patterns from the networks that perform initial lookups to

the domain. Certain DNS infrastructure characteristics, such as IP address ranges and that host either the authoritative name servers for the sites or the sites themselves, may be unique to malicious domains.

Identifying infrastructure that is shared by malicious domains may provide hints for identifying malicious domains before attacks are launched. Early DNS lookup characteristics can help network operators discover valuable information about the nature of the domains being looked up. Notably, we discover that malicious domains are initially queried from a much more diverse set of subnets than legitimate domains. Early lookup patterns for newly registered malicious domains differ markedly from those for legitimate domains. Spam domains are initially looked up by a more diverse set of network address regions than legitimate domains. Newly registered spam domains gain "popularity" more quickly.

Lookup patterns DNS query patterns Danzig et al. and Jung et al. conducted the first studies of DNS lookup behaviours at a local resolver; both of these studies examined lookup behaviour from the perspective of lookups to a single local resolver, and did not attempt to characterize how these lookup patterns differed for malicious domains. To build the domains' reputation, No to's and Exposure studied DNS lookup behaviour within a local domain below the DNS resolvers. This perspective on DNS lookup behaviour is useful, but it cannot reveal coordinated behaviour across multiple networks, and it must first detect an attack or compromised hosts before it can detect malicious domains.

Literature Survey

Cho Do Xuan1 and Hoa Dinh Nguyen1 [1] This paper shows that the URL attributes and can help ameliorate the capability to descry vicious URL significantly which gives the system, that may be considered as an optimized and friendly habituated result for vicious URL discovery

Shantanu and Janet B [2]. In this paper address the discovery of vicious URLs as a double classification problem and estimate the performance of several well- known machine literacy classifiers.

Anand Desai, Janvi Jatakia and Rohit Naik et al [3] This paper discusses the development of an extension for Chrome that will serve as middleware between users and harmful websites , reducing the likelihood that users may fall victim to those websites. Furthermore, it is impossible to compile an entire list of all dangerous content because even it is subject to constant change. This paper also discusses the use of machine learning to train the tool and classify new content so that appropriate action may be made after each observation into the relevant categories.

Frank Vanhoenshoven and Gonzalo Napoles [4] This paper analyze the performance of many popular classifiers, including Naive Bayes, Support Vector Machines, Multi-Layer Perceptron, Decision Trees, Random Forest, and k-Nearest Neighbors, in the detection of dangerous URLs as a binary classification problem.

Bc Andrea Turiaková [5] In this paper, they concentrate on the issue of identifying harmful URLs using data from URLs using technology for machine learning. In order to enable machine learning models to be trained on the URL features, they designed and constructed a system that extracts, transforms, and stores the URL data and they used a real-world dataset of malicious and benign URLs given by the ESET organisation to train and assess models built with the SVM light.

Hyunsang Choi and Bin B. Zhu et al [6] This paper presents a method using machine learning to determine the type of attack a malicious URL seeks to launch and detect malicious URLs of all the common attack types.

Proposed System

The majority of DNS-based malicious activity detection research focuses on identifying victim hosts by examining and simulating various DNS traffic characteristics, such as the variety of resolved IPs, geographic information name string structure, and DNS Time-to live (TTL) values. Yet, recent successful assaults show that attackers use genuine web and cloud storage services to conceal their command and control (C&C) connections in order to obfuscate the communication route and make it unidentifiable at the network level. Due to their apparent imitation of typical user behaviour, malware is difficult to detect by conventional DNS traffic-based detection techniques.

The recent increase in such attacks suggest the need for a comprehensive process-level DNS monitoring solution, PDNS. A back-end analytic server and a number of DNS sensors deployed on end hosts make up PDNS. The DNS sensors keep track of and record data on host DNS activity as well as related applications and activities. The back-end server gathers sensor data and creates machine learning models to identify malicious DNS activity and its related process. PDNS gathers and examines two different kinds of DNS activity information in order to create a model to identify malicious conduct. On the one hand, it gathers recognised and previously suggested network-based DNS capabilities, like IP and domain location diversity. On the other hand, it introduces additional process-based data, such as the quantity of requested domains, that describe the interaction between processes and DNS actions.

PROPOSED SYSTEM ARCHITECTURE

The use of encryption over DNS queries has recently increased, giving attackers another way to create a conduit to their C&C while evading the network-level detection mechanisms already in place. One would need to broaden the observed context around the host's DNS activity and look at the programmes that start such activity in order to detect such attacks.

Data on DNS activity is gathered by PDNS at the process and network levels. shows the general direction of the PDNS data collecting. The main data gathering component, PDNS sensor, is placed on a host and tracks DNS requests made by all processes. The sensor monitors each process' DNS activity. Before reporting to our DNS backend, it also gets information about each process from the kernel, such as loaded dynamic-link libraries and binary signatures. By referring to additional information sources for DNS WHOIS, IP WHOIS, and IP Geo-location, the PDNS backend enhances the network and DNS-related information by aggregating process-level DNS activity reports from all PDNS sensors deployed on each host. PDNS also collects DNS activity history from the local DNS server and cross-checks with the DNS records reported to PDNS backend (Figure 2 5). This final component acts as a fail-safe in case the attacker takes over the end-host and launches an attack against

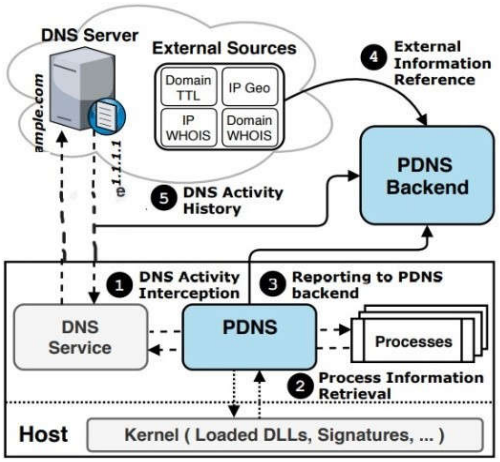

Fig 1: PDNS Architecture

The proposed system is based on the Network architecture.

There are two different parts mentioned in the architecture as below:

PDNS Application: PDNS consists of several DNS Application, deployed on end-hosts. It also obtains each process information from the kernel, such as loaded dynamic-link libraries and binary signatures. The PDNS application collects process-level DNS activity reports.

Backend Analytical Server: It collects data from sensors and builds machine learning models to detect malicious DNS behavior and its associated process. The PDNS backend aggregates process-level DNS activity reports from all PDNS sensors installed on each host and further extends the network and DNS related information by referring to other information sources for DNS WHOIS, IP WHOIS, IP Geo-location, etc.

Dataset Description

We now provide a dataset that is been used by us. where we built our customized dataset which contains all kinds of Urls(Malicious, benign, phishing) in huge number. Those particular dataset has been trained in such a way that it find out malicious and safe websites.

150	linkedin.com/pub/dir/elizabeth/scarborough	benign
151	cqgcr.ca/ang/producer01.php	benign
152	huffingtonpost.com/2011/10/24/south-carolina-primary-case_n_1029599.html	benign
153	http://qz.com/403774/quartz-daily-brief-japanese-military-verizon-aol-sporty-branson-saucy-swedes/	benign
154	select.nytimes.com/gst/abstract.html?res=F70814F93B581A7A93CBA9178BD95F4C8385F9	benign
155	startrekcostumes.net/	benign
156	en.wikipedia.org/wiki/Neil_Curtis	benign
157	music.carowinds.com/	benign
158	http://9779.info/%E6%A0%91%E5%8F%B6%E7%B2%98%E8%B4%B4%E7%94%BB/	malware
159	newjersey.craigslist.org/	benign
160	roverslands.net	phishing
161	familypedia.wikia.com/wiki/Joseph_Philippe_Roi_de_Villere_%281727-1769%29	benign
162	http://mylust.com/videos/232790/hentai-slut-with-big-juicy-tits-gets-fucked-doggy-style/	benign
163	acronymfinder.com/John-Burroughs-High-school-(Burbank%2c-CA)-(JBHS).html	benign

Fig 2: Dataset

OUTPUT SCREENS

```
# predicting sample raw URLs

urls = ['titaniumcorporate.co.za', 'en.wikipedia.org/wiki/North_Dakota']

for url in urls:
    print(get_prediction_from_url(url))
```

Fig 3: Raw URLs

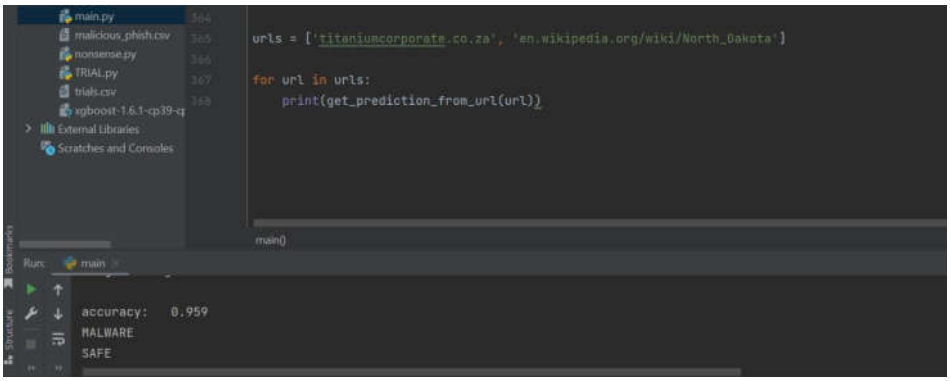

Fig 4 : Classifying the URLs

Fig 5: Accuracy

CONCLUSION & FUTURE SCOPE

PDNS, a novel end-point DNS monitoring system consisting of end-point DNS sensors and a centralized backend. PDNS utilizes extensive monitoring of DNS activities and set of host-based features which narrow down our analysis of host scope to a process-level. It enhances the visibility and context of monitoring, thus provides a capability to cope with various enormous detecting techniques and hence it detects the unseen malware. PDNS demonstrates its capability in detecting a malfeasant process with robust nature.

We expect that our approach would provide effective stable foundation and aid future research. As an enhancement to this work apart from detecting a malicious activity we can as well integrate PDNS into the system to also take preventive actions once the malicious activity is detected. There is also a scope to optimize the code to reduce computational time and make PDNS reactive to real time data.

7. REFERENCES

[1] Cho Do Xuan1 and Hoa Dinh Nguyen This study demonstrates how URL properties may considerably improve the system's capacity to detect malicious URLs, providing an optimal and user-friendly result for malicious URL detection.
https://thesai.org/Downloads/Volume11No1/Paper_19Malicious_URL_Detection_based_on_Machine_Learning.pdf
[2] Shantanu d Janet B The discovery of vicious URLs as a double classification problem and estimate the performance of several well- known machine literacy classifiers.
https://www.researchgate.net/publication/350931304_Malicious_URL_Detection_A_Comparative_Study
[3] Anand Desai, Janvi Jatakia, Rohit Naik et al The use of machine learning to train the tool and categorise fresh content is also discussed in this study in order to enable suitable actions to be taken following each observation into the right categories.
https://ieeexplore.ieee.org/document/8256834
[4] Frank Vanhoenshoven, Gonzalo Napoles ,In order to identify potentially unsafe URLs, this article compares the performance of several well-known classifiers.
https://ieeexplore.ieee.org/document
[5] Hyunsang Choi, Bin B. Zhu et al ,The method described in this work uses machine learning to identify the kind of attack a malicious URL intends to conduct and find harmful URLs for all popular attack types.
https://is.muni.cz/th/cfw06/Master_s_thesis.pdf/7850079
[6] CHOI, Hyunsang; ZHU, Bin B.; LEE, Heejo. Detecting Malicious Web Links and Identifying Their Attack Types. In: Proceedings of the 2Nd USENIX Conference on Web Application Development. 2011.
[7] SAHOO, Doyen; LIU, Chenghao; HOI and Steven C H Malicious URL Detection using Machine Learning: A Survey. 2017.
[8] GOOGLE. Google Safe Browsing - Blacklist service provided by Google. Available also from: https : / / safe browsing . google .com/.
[9] URLHAUS. Malware blacklist operated by abuse.ch. Available also from: https://urlhaus.abuse.ch/.
[10] D Sahoo and C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”. CoRR, abs/1701.07179, 2017.