Journal of Applied Mathematics and Bioinformatics, Vol. 11, No. 1, 2021, 25-31
ISSN: 1792-6602(print), 1792- 6939(online)
https://doi.org/10.47260/jamb/1112
Scientific Press International Limited

Entropy-Based Evaluation of DNS Activity for Threat Hunting

Argyrios (Argi) Alexopoulos[1]

Abstract

The paper documents, based mainly on [1]-[5] published papers where a consistent mathematical description of cyberspace and various types of Cyber-Attacks and protection measures are presented, a mathematical approach for Cyber Threat Hunting[2] using Domain Name System (DNS)[3] observations. After referring [5] to the various Advanced Persistent Threat (APT)[4] hunting techniques we propose a high level, mainly, entropy-based technique for detecting the existence of various threat vectors in our networks, demystifying DNS Anomalies.

Keywords: Domain Name System (DNS), Advanced Persistent Threat (APT) actors, Entropy, Anomaly Detection.

[1] Cyberspace Analyst staff in International Organization, Belgium.

[2] https://en.wikipedia.org/wiki/Cyber_threat_hunting
[3] https://en.wikipedia.org/wiki/Domain_Name_System
[4] https://en.wikipedia.org/wiki/Advanced_persistent_threat

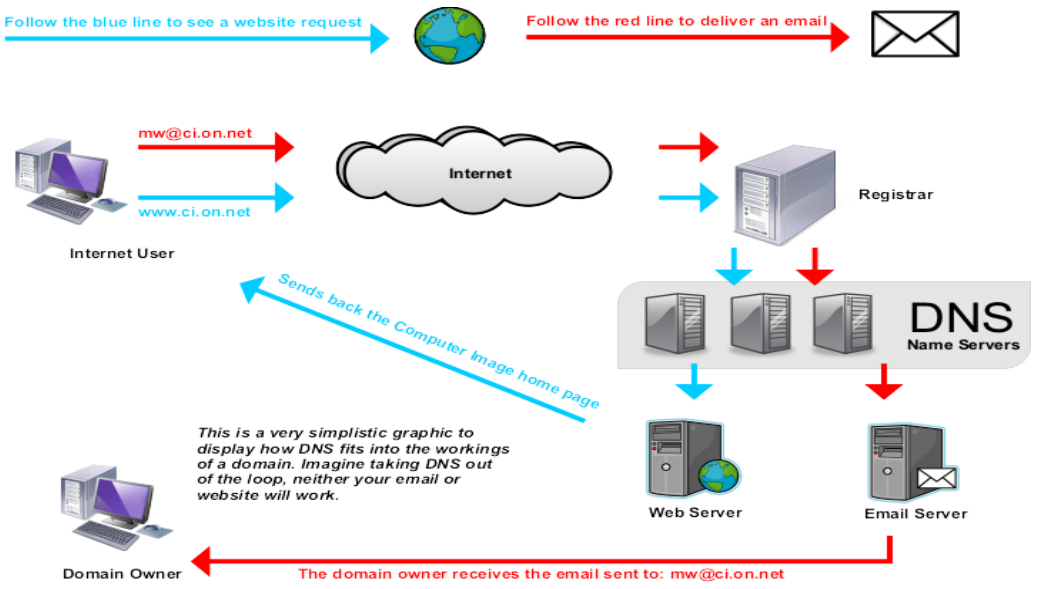
Article Info: Received: April 20, 2021. Revised: May 3, 2021.
Published online: May 10, 2021.

1. Introduction

The aim of the present paper is to propose an efficient threat hunting process by reducing DNS ambiguity in our internal network processes.
Based mainly on [1]-[5] published papers where a consistent mathematical description of cyberspace and various types of Cyber-Attacks and protection measures are presented, a mathematical approach for Cyber Threat Hunting using DNS observations. After referring to the various APT hunting techniques [5] we propose a high level, mainly, entropy-based technique for detecting the existence of various threat vectors in our networks, demystifying DNS Anomalies.

2. Malicious DNS Usage

A quite simplistic graphical display of how DNS work is shown below. We will not go to details on that since this is not the purpose of this paper and in addition, some related bits are described in the following sub chapters.
It is known and widely seen that APT's and cyber criminals utilize DNS for Command and Control (C&C), beaconing or resolution of attacker domains[5]. This is the main reason why collection and analysis of DNS events are critical for hunting, detecting, investigating and responding to threats. Defenders have challenging task to figure out what are the DNS abnormalities and distinguish them for threat hunting purposes.
All DNS related interactions for C&C communication that can be monitored and assessed in order to provide an entropy-based anomaly detection of suspicious or/and malicious activities in our network ecosystem are analysed in section 4.



[5] https://resources.infosecinstitute.com/topic/attacks-over-dns/

3. Entropy from DNS Perspective

Entropy[6] is a scientific concept, as well as a tangible physical property that is most frequently related to a state of disorder, randomness, or uncertainty. The concept itself is used in diverse fields, from classical thermodynamics, where it was first recognized, to the principles of information theory. It has found far-ranging applications in chemistry and physics, in biological systems and their relation to life, in economics, sociology, weather science, climate change, and information systems including the transmission of information in telecommunication.

In this paper, we refer to entropy from a DNS perspective and we orient it in broad DNS ecosystem. We will identify all main DNS interactions that we can measure a kind of disorder, randomness, or uncertainty and we will transfer the concept of Entropy-Based Evaluation of DNS Activity for Threat Hunting in our previous scientific research [1]-[5].

4. DNS Interactions to be Entropy-Evaluated

As already stated, there are many DNS interactions that malicious actors and more likely APTs will conduct to avoid detection. In this paper we focus on C&C related traffic. Some of them will be shortly presented below. We may not be able to keep pace with every new DNS exploitation but we may be proactive by calculating a kind of entropy-based threshold. We will show holistically that all these potentially malicious activities through DNS have high Entropy value.

4.1 DNS Queries to Unusual Destinations

Disorder, randomness, or uncertainty increases when DNS queries to unusual destinations increases. Let DNS_1 be the queries to abnormal destinations. This abnormality may be identified internally using statistical analysis.

4.2 DNS Traffic Bypassing DNS Forwarding Servers

Traffic is expected to be forwarded to Internet Service Provider (ISP) or DNS Service Provider, from internal hosts, via a DNS forwarding server. Any traffic bypassing DNS forwarding Server/Entity indicates high DNS entropy. Let DNS_2 be the bypassing traffic. We should not expect the outbound DNS to be other than the IP of forwarding server.

[6] https://en.wikipedia.org/wiki/Entropy

4.3 DNS Queries Using Multiple Subdomains

Given that DNS queries do not normally use subdomains (some exceptions exist though), DNS entropy increases proportionally to the number of dots (subdomains) of the queries. For example, the following query indicates high disorder, randomness, or uncertainty:
SV1DVV.2.8572Y8.YY.19283OLPDL20PPLPOLK.CSEC.apple.net
Let DNS_3 be the multi subdomains queries.

4.4 DNS Queries Using Lengthy Domains

Combination of the existence of multiple domains in DNS queries with lengthy ones is indicative of high DNS entropy and the risk of malicious activity is greater. Let DNS_4 be the lengthy DNS queries. Therefore, by combining both the character length and sub domain in a query is critical.

4.5 DNS Queries Including a Mix of Upper/Lowercases

Queries that include a mix of upper and lowercases indicate a higher DNS entropy, according to the approach of this paper. The more frequent interchange of upper and lowercase in a DNS query the higher potential DNS entropy. Let DNS_5 be the queries that mix Upper/Lowercases.

4.6 DNS Queries Including Alphanumeric Characters

Regular DNS queries are expected to contain only numbers, roman alphabetical letters, dots and hyphens. Not unlike the example of mixed case lettering, non-alpha numeric strings may indicate base 64 encoded strings. Which would be an unusual event. Let DNS_6 be the DNS queries that include Alphanumeric characters.

4.7 Allowed Traffic on Port 53 Inbound Transmission Control Protocol (TCP)

This port is used for zone transfer and should only be allowed between primary and secondary DNS servers. In a perfect scenario, searching for outbound traffic over Port 53 will, at best, yield misconfigured systems. Alternatively, at worse, potentially malicious traffic. Exceptions to those IP's should be used as a trigger point for investigation. Let DNS_7 be the above DNS traffic exceptions.

4.8 Outbound Traffic, Other than 53 and 123, Should not be User Datagram Protocol (UDP)

Genuine services often rely on TCP. So, if any outbound communication is made to any port on the UDP, this should be a trigger point for investigation. Let DNS_8 be the above DNS exceptions.

4.9 Fast Flux DNS

DNS fluxing is a technique used by attackers to hide an actual phishing or malware domain behind constantly changing compromised hosts (IP) which are acting like proxies. To accomplish this, the Time to Live (TTL) for DNS is set very low (close to 5 min) so that the changes made in DNS will reflect quickly over the internet. Let DNS_9 be the DNS queries for domains, having a TTL less than 5-10 mins.

4.10 Domain Generation Algorithm (DGA)

Domain generation algorithms are used by many malware families to create a large number of domain names that can be used as a rendezvous point with C2. It can be detected based on a myriad of non-responsive DNS queries from the source towards Randomised Domains. Let DNS_{10} be the non-responsive DNS queries.

5. Entropy-Based Evaluation of DNS Activity

Assume a node V and the $[\mathcal{C}_{available}(V)](t_i)]$ that is the set of ordered columns of all available constituents $\left(dev_1^{(V)}, \dots, dev_{m_V}^{(V)}, res_1^{(V)}, \dots, res_{\ell_V}^{(V)}\right)^T$ of V, over the time $t_i \in [0,1]$ and $[DNS(V)](t_i)]$ that is the set of ordered columns of all DNS activity (mainly queries) of all constituents of V $\left(DNS(dev_1^{(V)}), \dots, DNS(dev_{m_V}^{(V)}), DNS(res_1^{(V)}), \dots, DNS(res_{\ell_V}^{(V)})\right)^T$ over the time $t_i \in [0,1]$.

Let assume also the set of DNS characteristics of DNS activity ($[DNS(V)](t_i)]$) that have to be evaluated $\{DNS_1,\ DNS_2,\ DNS_3 \dots DNS_{10}\}$ as we have described in Section 4.

We will presume the following notations [2] for the sets of relative valuations of parts (fractions) of possible constituents:

$$
\begin{aligned}
&{}_{S_W}\mathfrak{C}^{(fraction)}(V) \\
&= \left\{\left(S_W[x_1, x_2, x_3, t\,]\left(fr\left(DNS(dev_1^{(V)})\right)\right), \dots, S_W[x_1, x_2, x_3, t\,]\left(fr\left(DNS(dev_{M_V}^{(V)}),\right)\right),\right.\right. \\
&\left.\left. S_W[x_1, x_2, x_3, t,\ id_t\,]\left(fr\left(DNS(res_1^{(V)})\right)\right), \dots, S_W[x_1, x_2, x_3, t\,]\left(fr\left(DNS(res_{L_V}^{(V)})\right)\right)\right)\right\}^T \\
&S_W[x_1, x_2, x_3, t\,]\left(fr\left(DNS(dev_K^{(V)})\right)\right)\ is\ valuation\ of\ part \\
&of\ possible\ DNS\ activity\ in\ V\ subject\ to\ W,\ \ k \le \mathcal{M}_V\ \ with\ \mathcal{M}_V \in \mathbb{N} \\
&S_W[x_1, x_2, x_3, t\,]\left(fr\left(DNS\left(res_\xi^{(V)}\right)\right)\right)\ is\ valuation\ of\ possible\ DNS\ activity \\
&in\ V\ subject\ to\ W, \xi \le \mathcal{L}_V\ with\ \ \mathcal{L}_V \in \mathbb{N}\,, \\
&at\ the\ spatiotemporal\ point\ (x_1, x_2, x_3, t)\ \in \mathbb{R}^3 \times [0,1]\}:
\end{aligned}
$$

the set of all ordered columns of relative valuations of parts (fractions) of possible DNS activity of V, from the viewpoint of the (user(s) of) node W, over the space time $\mathbb{R}^3 \times [0,1]$;
We employed Shannon's function in order to calculate the entropy (evaluation) of $\mathcal{S}_W\mathfrak{C}^{(fraction)}(V)$ as $H(X)$:

$$H(X) = -\sum_{i \in X} P(i) log_2 P(i)$$

where X is the data set $\{DNS_1,\ DNS_2,\ DNS_3 \ldots DNS_{10}\}$ of the DNS characteristics (mainly for C&C).
We calculate periodically for each element of $[DNS(V)](t_i)$ set the $H(X)$. Afterwards, we get the mean value of $H(X)$ over the time t_i. Having in advance identify internally a "safe" threshold for DNS transactions in our ecosystem, we compare this value to the mean.
If:

- $H(X) \ll mean,$
  $then\ the\ propability\ of\ a\ threat\ actor\ in\ our\ ecosystem\ is\ almost\ 0$
- $H(X) < mean,$
  $then\ the\ propability\ of\ a\ threat\ actor\ in\ our\ ecosystem\ is\ relatively\ low$
- $H(X) \approx mean,$
  $then\ the\ propability\ of\ a\ threat\ actor\ in\ our\ ecosystem\ is\ medium$
- $H(X) > mean,$
  $then\ the\ propability\ of\ a\ threat\ actor\ in\ our\ ecosystem\ is\ relatively\ high$
- $H(X) \gg mean,$
  $then\ the\ propability\ of\ a\ threat\ actor\ in\ our\ ecosystem\ is\ very\ high$

References

[1] Daras, N.J. (2018). On the mathematical definition of cyberspace, Theoretical Mathematics & Applications, Vol.8, no.1, pp. 9-45. Scienpress Ltd, 2018.

[2] Daras, N.J and Alexopoulos, A. (2017). Mathematical description of cyber-attacks and proactive defenses, Journal of Applied Mathematics & Bioinformatics, Vol.7, no.1, pp. 71-142. Scienpress Ltd, 2017.

[3] Daras, N.J and Alexopoulos, A. (2017). Modeling Cyber-Security, Journal of Applied Mathematics & Bioinformatics, Vol. 7, no.1, 2017, pp. 71-142. Scienpress Ltd, 2017.

[4] Alexopoulos, A. and Daras, N. (2018). Mathematical Study of Various Types of Cyber-Attacks and Protection, Journal of Computations & Modelling, Vol.8, no.2, 2018 pp. 1-61. Scienpress Ltd, 2018.

[5] Alexopoulos, A. and Daras, N. (2020). Mathematical Study of Advanced Persistent Threat (APT) Hunting Techniques, Journal of Computations & Modelling, Vol. 10, no. 2, pp. 1-24. Scientific Press Ltd, 2020.