International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Analysis of the Domain Names Linked to The Important World Events Using A Longitudinal Measurement Approach

Dr. V. Nivedita[1], Siddham Singh Rao[2], Nishant Shandilya[3], V. Phanindra Varma[4]

[1]M.E, PH. D, Assistant Professor SRM Institute of Science and Technology
[2,3,4]Department of Computer Science and Engineering, SRM Institute of Science and Technology

Abstract
Since the lack of sufficient security mechanisms, domain system (DNS) has become the main operational infrastructure for cyber intruders to launch cyber-attacks. Therefore, how to discover and block the potential malicious domains and its corresponding IP addresses fast and accurately has become a hot research area as it is one of the most important method in preventing unknown cyber-attacks. In this paper, we proposed an approach to detect malicious domains by analyzing massive mobile web traffic data. We used multiple features to classify, including the textual features and the traffic statistics features of domains and presented three typical classifiers to compare the classifying effect of each. Spark framework is leveraged to speed up the calculation of a large-scale DNS traffic. The efficiency of our system makes us believe the approach can help a lot in the field of network security. The new features are harder to be tampered with and can help determine whether a domain is malicious from a more comprehensive perspective. We evaluate MalPortrait on the passive DNS traffic collected from real-world large ISP networks.

Keywords: Text determination, Investigation, Client experience, Engineering, and development (AC), Instructing and learning techniques, Characterization, AR frameworks.

1. INTRODUCTION

The rapid proliferation of internet usage globally has significantly increased the complexity and interconnectivity of computer networks worldwide. This expansion has inadvertently created new vulnerabilities and gaps within these networks that malicious actors, such as hackers and cybercriminals, actively exploit to carry out illicit activities. Their nefarious objectives range from launching cyber attacks to breach government systems, violate the privacy and data of corporations, or even manipulate individuals through deceptive phishing websites. One pressing issue that has garnered substantial attention from researchers is the detection and mitigation of malicious domain names, which serve as critical enablers for these harmful cyber activities. Moreover, the grave implications of cyber threats, information security breaches, and the associated economic damages they inflict upon organizations can no longer be disregarded or underestimated.

A recent report published in 2021 by the International Data Corporation (IDC), a prominent market intelligence firm, revealed an alarming statistic: an overwhelming 87% of organizations worldwide have fallen victim to attacks specifically targeting the Domain Name System (DNS).

The DNS plays a pivotal role in facilitating the internet revolution by functioning as a hierarchical and decentralized naming system. Its primary purpose is to enhance user experience by translating the numerical IP addresses of websites into memorable, user-friendly domain names, and vice versa. When a user attempts to access a particular website, the DNS initiates a query process, reaching out to the nearest root name server to obtain the requested information. The root server then communicates with other top-level domain (TLD) servers, such as .com, .org, and .edu, gradually piecing together the components of the domain name until it ultimately locates and returns the corresponding IP address from the authoritative DNS server hosting that website.

Although the DNS architecture appears decentralized, its operation inadvertently centralizes network traffic, creating potential vulnerabilities. Furthermore, the DNS protocol lacks inherent security mechanisms, such as authentication between nodes or encryption of data packets. This structural weakness exposes two primary security concerns: the lack of authentication and encryption mechanisms between nodes, and the tendency for DNS traffic to converge towards central points, despite the system's intended decentralized design. This centralization effect enables attackers to target multiple entities simultaneously with relative ease. Additionally, the public availability of DNS server information poses risks, particularly for companies with inadequate DNS configurations, as this exposure could compromise their internal systems and data integrity.

2. PROBLEM STATEMENT

Improve detection accuracy: A single detection model may not be able to cover all malicious behaviors and variants comprehensively. Therefore, utilizing multiple models for collaborative detection can enhance the accuracy of detection. Each model can emphasize different features or algorithms, thereby increasing the detection rate of malicious software and reducing false positives.

Counteract the evolution of malicious software: Malicious software evolves rapidly, with new variants constantly emerging. By employing multiple detection models, the sensitivity to different variants of malicious software can be increased, enabling a timely detection and response to new threats.

3. RELATABLE WORK

The authors of the current system carried out a comprehensive longitudinal measurement study with a particular focus on the Olympic Games to examine domain names linked to significant international events. The reason the Olympics were chosen as the study's topic was the amount of attention they garner from the moment the host location is revealed until the competition is over. The three-year investigation focused on the Olympic Games in Tokyo (2020–2021), Beijing (2022), and Paris (2024).

After conducting a thorough analysis, the authors of the current system have identified a number of significant findings. First of all, they noted a marked rise in the registration of Olympic-related domain names (ODNs) in tandem with the 2020 Tokyo Olympics being postponed and the 2022 Beijing Olympics being diplomatically boycotted. There are a number of reasons for this spike in registrations, such as speculative activity, potential malice, or opportunistic registrations.

Moreover, the analysis revealed a significant increase in the quantity of ODNs used for malevolent websites shortly before the Olympic Games started. This research emphasizes the possible dangers of

domain squatting and cybercrime that takes advantage of the increased interest and focus around big events like the Olympics.

The authors of the current system also mentioned how many Open Data Networks (ODNs) were obtained to represent the regional aspect of every Olympic Games, possibly focusing on particular regions or linguistic groups. From a security standpoint, these domain names needed to be closely monitored in order to reduce potential threats and guarantee the integrity of official Olympic-related websites and online presence.

The authors of the current system suggested, in light of their research, the establishment of a generic top-level domain (gTLD) with an Olympic theme, such as.olympic, which would only be accessible by reputable businesses, official Olympic websites, and organizations connected to the Olympic Games hosted in each nation. By taking this precaution, the likelihood of phishing attempts, domain squatting, and other malicious activity aimed at the Olympic events and brand may be decreased.

Additionally, the existing system authors emphasized the importance of organizations planning future events paying close attention to regional keywords and domain registration patterns specific to the locality. By proactively monitoring and addressing potential threats related to domain registrations, event organizers can enhance their security preparations and safeguard against potential cyber attacks or brand infringement.

With a focus on the Olympic Games, the authors of the current system carried out a thorough analysis of domain names linked to significant international events over the course of a three-year longitudinal measurement study. The investigation focused on the Olympics because of the ongoing interest and attention they garner around the world, from the announcement of the host venue until the end of the competition. The research focused on the next three Olympic Games, which will be held in Paris in 2024, Beijing in 2022, and Tokyo in 2020–2021.

This thorough investigation's conclusions showed a number of important patterns and trends. Notably, the creators of the current system noticed a sharp rise in the registration of Olympic-related domain names (ODNs) in tandem with two significant occurrences: the COVID-19 pandemic-related postponement of the 2020 Tokyo Olympics and the diplomatic boycott of the 2022 Beijing Olympics by a number of countries. There are several possible explanations for this spike in ODN registrations, including speculative registrations by people or organizations hoping to profit from the increased interest, potentially malicious intent involving cybercrime, or even opportunistic registrations by individuals or entities hoping to profit from the increased interest.

4. PROPOSED WORK AND EXPERIMENT

Identifying malicious domain names has become an effective approach to protecting internet users from potential threats and malicious activities online. While previous works have achieved notable success in this domain, they heavily rely on historical Domain Name System (DNS) responses and external intelligence sources. Consequently, these existing methods may fail to identify previously unknown or undocumented domain names for which no prior knowledge or intelligence exists.

In this research, we propose a novel feature ensemble-based approach to identifying malicious domain names directly from valid DNS responses, without relying on historical data or external sources. Our proposed approach, termed Glacier, addresses the aforementioned limitation by leveraging two distinct types of features extracted from domain name strings: linguistic features and statistical features.

Linguistic features are vector representations generated from the character sequences of domain names

using a bidirectional long short-term memory (LSTM) neural network. It is noteworthy that we have modified the final LSTM layer to enhance the expressiveness and representational capacity of these linguistic features. By capturing the sequential patterns and contextual information within domain name strings, these linguistic features can effectively encode the inherent linguistic characteristics and potential indicators of malicious intent.

Complementing the linguistic features are statistical features, which consist of six manually designed statistics that represent the structural information and characteristics of a domain name. These statistical features encode aspects of the domain name structure that may be challenging for an LSTM neural network to learn directly from the character sequence alone. By combining these two feature types, our approach leverages both the linguistic patterns and the structural properties of domain names to enhance the identification of malicious domains.

Furthermore, our approach recognizes that domain names resolved to the same IP addresses might share similarities in reputation and potentially malicious intent. To capture this association information, we construct a domain association graph, which quantifies the relationship between different domain names based on their shared IP resolutions. Intuitively, if a domain name is resolved to the same IP addresses as known malicious domains, there is a higher likelihood that the domain name itself is also malicious.

By incorporating linguistic features, statistical features, and domain association information into a unified ensemble model, our proposed Glacier approach aims to provide a comprehensive and effective solution for identifying malicious domain names directly from valid DNS responses, without relying on historical data or external intelligence sources. This approach has the potential to enhance the proactive detection and prevention of emerging cyber threats and malicious online activities, thereby contributing to a more secure and trustworthy internet ecosystem.

5. IMPLEMENTATION AND DISCUSSIONS

A distributed computing architecture was used to implement the suggested system in order to meet the demands of large-scale data processing and analysis. The data collection module collected information from government agencies, news outlets, and domain registries by using web scraping techniques and APIs. Using Apache Spark and Hadoop for parallel processing, the gathered data underwent preprocessing, which included timeline synchronization, standardization, and cleaning. The event-domain mapping component identified pertinent domain names based on semantic relevance to real-world events by using natural language processing (NLP) techniques like named entity recognition and topic modeling. As supplementary mapping criteria, temporal proximity and geolocation were also applied. Using time-series analysis and anomaly detection algorithms, the longitudinal analysis module found patterns and trends in domain registrations over time and connected them to the identified world event.

A distributed web crawler was used to collect historical information on malware signatures, web content, and IP resolutions for domain reputation analysis. In order to evaluate the reliability and possible risks connected with the identified domains, this data was subjected to machine learning models and integrated with pre-existing domain reputation databases.

The visualization and reporting module made use of D3.js and React to create interactive dashboards and reports that let stakeholders examine the results using a variety of visualizations, including network graphs, timelines, and maps. For real-time data ingestion and processing, the continuous monitoring module used Apache Kafka and Apache Flink. This allowed for the instantaneous detection of new world events and domain registrations, triggering alerts and enabling quick response.

The system that was put into place effectively illustrated the viability and importance of using a longitudinal measurement approach to examine the connection between significant global events and domain names. Through the integration of various data sources and the utilization of sophisticated methods for event-domain mapping, longitudinal analysis, and domain reputation evaluation, the system offered valuable insights into the dynamics of domain registrations related to noteworthy worldwide events.

The acquisition and integration of data from various sources, each with its own format and structure, posed a significant implementation challenge. A significant amount of work was put into preprocessing the data to make sure that it was compatible and consistent across datasets.

6. CONCLUSION

In conclusion, this study has demonstrated the value of employing a longitudinal measurement approach to analyze the relationship between domain name registrations and important world events. By combining data from diverse sources, preprocessing techniques, and robust methodologies for event-domain mapping, longitudinal analysis, and domain reputation assessment, we have uncovered valuable insights into the dynamics of domain registrations surrounding significant global occurrences. The findings highlight the potential risks associated with domain registrations during and around major world events, including the possibility of domain abuse, cybersquatting, and other malicious activities. The visualizations and reports generated by this analysis provide a comprehensive overview of the patterns, trends, and anomalies observed, enabling stakeholders to take proactive measures in monitoring and mitigating potential threats.

Furthermore, the continuous monitoring and updating mechanisms established as part of this architecture ensure that the analysis remains relevant and up-to-date, allowing for timely detection and response to emerging threats or changes in the domain registration landscape. While this study has made significant contributions, there are opportunities for further research and refinement. Incorporating additional data sources, refining the event-domain mapping criteria, and exploring advanced machine learning techniques for pattern recognition and anomaly detection could enhance the accuracy and effectiveness of the analysis.

Overall, this longitudinal approach to analyzing domain names linked to important world events has demonstrated its value in enhancing cybersecurity preparedness, raising awareness, and supporting proactive measures to combat potential domain abuse and protect valuable digital assets.

REFERENCES

1. Sultan H. Almotiri Integrated Fuzzy Based Computational Mechanism for the Selection of Effective Malicious Traffic Detection Approach IEEE Access, 2021
2. Inwoo Ro,Boojoong Kang,Choonghyun Seo,Eul Gyu Im Detection Method for Randomly Generated User IDs: Lift the Curse of Dimensionality IEEE Access, 2022
3. Haojun Wang,Haixia Long,Ailan Wang,Tianyue Liu,Haiyan Fu Deep Learning and Regularization Algorithms for Malicious Code Classification IEEE Access, 2021
4. P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-GÃ3mez, N. Pitropakis, et al., Hiding in plain sight: A longitudinal study of combosquatting abuse, Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 569-586, Oct. 2017.
5. H. Suzuki, D. Chiba, Y. Yoneya, T. Mori and S. Goto, ShamFinder: An automated framework for

detecting IDN homographs, Proc. Internet Meas. Conf., pp. 449-462, Oct. 2019.
6. A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot and H. Pfister, UpSet: Visualization of intersecting sets, IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 1983-1992, Dec. 2014.
7. S. Bird, E. Loper and E. Klein, Natural Language Processing With Python, Sebastopol, CA, USA:Oâ€TMReilly Media, 2009.
8. Y. Sakurai, T. Watanabe, T. Okuda, M. Akiyama and T. Mori, Discovering HTTPSified phishing websites using the TLS certificates footprints, Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS&PW), pp. 522-531, Sep. 2020.
9. S. Torabi, A. Boukhtouta, C. Assi and M. Debbabi, Detecting Internet abuse by analyzing passive DNS traffic: A survey of implemented systems, IEEE Commun. Surveys Tuts., vol. 20, pp. 3389-3415, 4th Quart. 2018.
10. S. SchÃ1⁄4ppen, D. Teubert, P. Herrmann and U. Meyer, FANCI: Feature-based automated NXDomain classification and intelligence, Proc. USENIX Secur. Symp. (USENIX Secur), pp. 1165-1181, Aug. 2018.
11. Y. Zhauniarovich, I. Khalil, T. Yu and M. Dacier, A survey on malicious domains detection through DNS data analysis, ACM Comput. Surv., vol. 51, no. 4, Sep. 2018.
12. H. Gao, V. Yegneswaran, J. Jiang, Y. Chen, P. Porras and S. Ghosh, Reexamining DNS from a global recursive resolver perspective, IEEE/ACM Trans. Netw., vol. 24, no. 1, pp. 43-57, Feb. 2016.
13. R. Kozik, M. Pawlicki and M. Choras, Cost-sensitive distributed machine learning for netflow-based botnet activity detection, Secur. Commun. Netw., vol. 2018, Dec. 2018.
14. Y. Zhauniarovich, I. Khalil, T. Yu and M. Dacier, A survey on malicious domains detection through DNS data analysis, ACM Comput. Surv., vol. 51, no. 4, pp. 1-36, Sep. 2018.
15. S. Vosoughi, P. Vijayaraghavan and D. Roy, Tweet2Vec: Learning tweet embeddings using character-level CNN- LSTM encoder-decoder, Proc. 39th Int. ACM SIGIR Conf., pp. 1041-1044, 2016.
16. D. S. Berman, DGA CapsNet: 1D application of capsule networks to DGA detection, Information, vol. 10, no. 5, pp. 157, Apr. 2019.
17. Y. Qiao, B. Zhang, W. Zhang, A. K. Sangaiah and H. Wu, DGA domain name classification method based on long short-term memory with attention mechanism, Appl. Sci., vol. 9, no. 20, pp. 4205, Oct. 2019.
18. L. Yang, G. Liu, Y. Dai, J. Wang and J. Zhai, Detecting stealthy domain generation algorithms using heterogeneous deep neural network framework, IEEE Access, vol. 8, pp. 82876-82889, 2020.
19. Y. Fu et al., Stealthy domain generation algorithms, IEEE Trans. Inf. Forensics Security, vol. 12, no. 6, pp. 1430- 1443, Jul. 2017.
20. Z. Chen, M. Roussopoulos, Z. Liang, Y. Zhang, Z. Chen and A. Delis, Malware characteristics and threats on the Internet ecosystem, J. Syst. Softw., vol. 85, no. 7, pp. 1650-1672, 2012.
21. T. Nelms, R. Perdisci and M. Ahamad, ExecScent: Mining for new C&C domains in live networks with adaptive control protocol templates, Proc. USENIX Conf. Security Symp., pp. 589-604, 2013.
22. R. Vinayakumar, K. P. Soman, P. Poornachandran and S. S. Kumar, Evaluating deep learning approaches to characterize and classify the DGAs at scale, J. Intell. Fuzzy Syst., vol. 34, no. 3, pp. 1265-1276, 2018.
23. G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen and F. Petitjean, Characterizing concept drift, Data Min. Knowl. Discovery, vol. 30, no. 4, pp. 964-994, 2016.
24. J.-Y. Bisiaux, DNS threats and mitigation strategies, Netw. Secur., vol. 2014, no. 7, pp. 5-9, Jul. 2014.

25. K. L. Chiew, K. S. C. Yong and C. L. Tan, A survey of phishing attacks: Their types vectors and technical approaches, Expert Syst. Appl., vol. 106, pp. 1-20, Sep. 2018.
26. A. Aleroud and L. Zhou, Phishing environments techniques and countermeasures: A survey, Comput. Secur., vol. 68, pp. 160-196, Jul. 2017.
27. M. Khonji, Y. Iraqi and A. Jones, Phishing detection: A literature survey, IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091-2121, Apr. 2013.
28. M. Janbeglou, M. Zamani and S. Ibrahim, Redirecting outgoing DNS requests toward a fake DNS server in a LAN, Proc. IEEE Int. Conf. Softw. Eng. Service Sci., pp. 29-32, Jul. 2010.