

Identification of URL Based Attacks from IP Data

Darshan M K, Shruthi M T

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004. Assistant professor Department of Master of Computer Applications, BIET, Davangere-577004.

ABSTRACT: The surge in online services has also resulted in an exponential rise in sophisticated and frequency-intensive cyber attacks, where malicious URLs constitute a significant factor in phishing and distribution of malwares and unwanted downloads. Existing URL detection methods primarily involve a blacklist based technique that has limited functionality in detecting new and obfuscated malicious links as it identifies URLs that have been identified previously as malicious. Thus, this project has proposed an intelligent malicious URL detection framework by combining both the features of URLs and the contextual intelligence at the IP level. The approach consists of not only extracting lexical and structural features from the URLs but also extracting contextual data from the IP resolved address: physical location, service provider information, past reputation, traffic behaviour and many more. Thus, this feature vector is much more informative which helps to attain better detection and a much smaller false positive rate. The approach is aimed to be scalable, real-time and integrable with current security infrastructures, providing a proactive threat detection capability beyond reactive signature-based approaches.

KEYWORDS: Malicious URL Detection, Cybersecurity, Phishing Attacks, Machine Learning, IP Reputation Analysis, Feature Engineering, Web Security, Threat Intelligence

I. INTRODUCTION

The project "Identification of URL-Based Attacks from IP Data" puts forth a multi-layered security framework aimed to counter malign web links. While the usage of the Internet forms a significant portion of day-to-day lives today, the invention and subsequent usage of fraudulent URLs from the simplest of credential theft phishing to ransomware injecting webpages are possible due to the underlying trust in DNS. Traditional security tools fail at this aspect by analyzing URLs in isolation; our framework rectifies

that problem by analyzing the inherent association between the web link and the network (IP network) that supports it. This method provides insight to minor malicious cues missed by traditional blacklists by analyzing both string based features and the behaviors at the network level.

A high-fidelity dataset consisting of both identified benign and malicious URLs are utilized at the core of this framework. The system uses a two layered feature extraction where lexical analysis inspects the construction of the URL (structure of the URL string) with features like character entropy, domain structure and obfuscated paths. For high accuracy of classification the framework implements a hybrid ensemble of machine learning models.

II. LITERATURE REVIEW

The number of studies on detection of malicious URLs is rapidly increasing because phishing, malware distribution and attacks on web resources grow fast. Previous methods are commonly based on Blacklist/Whitelist or Signature-Based systems, but they are not efficient in detecting novel or encoded malicious URLs. These techniques are useful only if there is knowledge about the threat, and therefore ineffective in detection of Zero-Day attacks. In order to overcome limitations of traditional techniques, the use of Machine Learning approaches has been proposed to detect malicious URLs using various characteristics of URLs like the length, usage of special characters, domain structure, and character entropy. Some common Machine Learning methods such as Support Vector Machines (SVM), Random Forest and Gradient Boosting performed well in detecting malicious URLs. Among the algorithms used to predict malicious URLs, ensemble methods provide excellent results when dealing with high dimensional data. Moreover, some works enhance accuracy using IP and network-based features such as geolocation of IP, server hosting information and reputation data. Deep learning

algorithms, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were used for detecting malicious URLs by leveraging sequential pattern learning in the URL strings. Combining of URL-centric and IP-centric data improved the prediction results and reduced false positive rates, so the idea of a hybrid, multi-dimensional security framework has been motivated.

III. EXISTING SYSTEM

Most existing methods for detecting harmful web links depend on a small group of conventional security techniques. Another commonly utilized technique for suspicious link detection is a heuristic-driven analysis approach. Here, a specific set of heuristic rules is employed to identify potentially harmful web addresses. Examples of such rules would involve checking if the given URL utilizes a numeric IP address instead of a domain name, or if it has an unnecessarily long/irregular structure. These methods will work to identify certain simple attacks effectively, however the rules are fixed, and as such malicious agents knowledgeable of the rules will simply devise and adapt URL structures in order to by-pass them, thus negating their usefulness over the long term. Reputation-based systems have more recently grown in prominence and trust levels are given to IP addresses and domains based on a variety of characteristics; the registration age of the domain, previous use in attacks, any link with known malicious IPs etc. The main disadvantage with reputation based systems is that they tend to be slow to update and thus would not detect quickly developing, or recently created malicious domain.

IV. PROPOSED SYSTEM

This developed system to detect 'Identification of URL based attacks from IP data' is an advanced and multifaceted security tool that tries to overcome limitations of already established techniques. 3.0 System Architecture and Strategic Advantage The developed system is a high speed, modular pipeline that converts network raw input data into security insights.. By shifting from static inspection to a context-aware analysis model, the system bridges the gap between traditional packet filtering and modern behavioural forensics. The Dual-Pronged Extraction Logic At the point of ingestion, the system initiates a

synchronized, two-tier feature distillation process to build a comprehensive "digital fingerprint" of the query.

ADVANTAGES OF PROPOSED SYSTYEM

Advantages of the Proposed Framework Compared to legacy security infrastructures; this system provides several critical operational improvements:

Autonomous Threat Discovery. Since the system works based on learned patterns rather than fixed signatures, it naturally prevents Zero-Day attacks that have not been listed by global blacklists.

Robustness to Infrastructure level Changes. While attackers can easily change the string of a URL, modifying a whole hosting infrastructure requires resources and considerable time. The platform has therefore a native defense against URL shortening and fast-flux based obscuring, due to the relative staticity of IP data.

"Alert Fatigue" reduction. The system provides very few false positive rate due to the addition of deep network context, allowing legitimate business traffic to proceed normally. The security system would still be maintained high.

Elastic Performance. Since it is built to match the demands of modern enterprises, the system is able to processes massive number of requests in parallel and provide sub millisecond inference rate at high traffic network gateways.

DESIGN AND IMPLEMENTATION

The developed system, namely Identification of URL-Based Attacks from IP Data, implements a framework based on machine learning that detects malicious links by combining both URL-related structural characteristics and IP-based information. A labeled dataset of benign and malicious URLs is gathered from existing cyber security sources. We extract two sets of features per URL: one is lexical, describing aspects of the URL (e.g. Length of the URL, presence of special characters, number of sub-domains, use of HTTPS). The second set consists of IP-related information of the website related to the URL (e.g. Domain age, whether the IP is listed in blacklists, its DNS records, and geolocation)

During implementation, the dataset undergoes preprocessing steps such as handling missing values, encoding categorical variables, and splitting into training and testing sets (80:20 ratio). A Random Forest classifier is employed due to its robustness and high

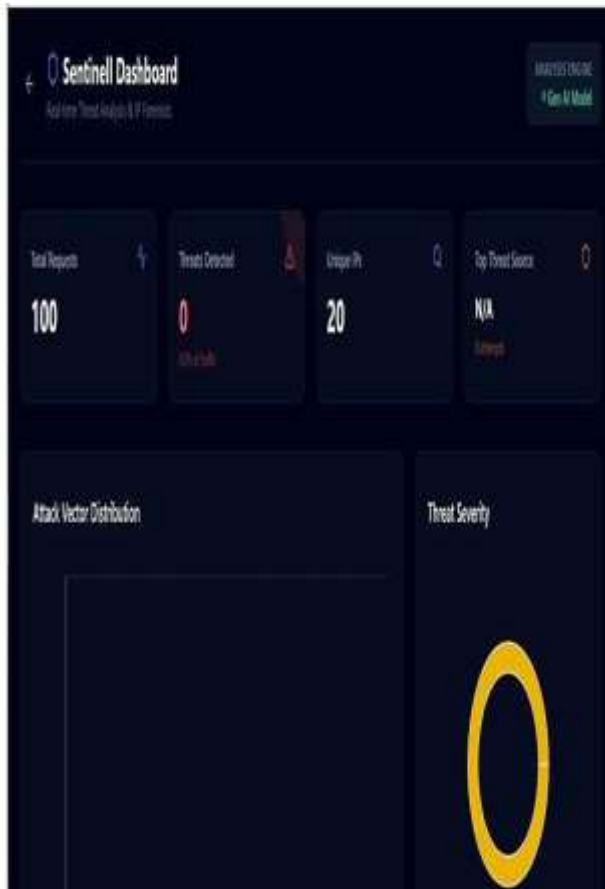
(about 5–7%) under moderate noise conditions, whereas the baseline system showed higher sensitivity to accent and noise variations. User evaluation studies involving bilingual speakers reported higher satisfaction scores in terms of pronunciation handling, contextual accuracy, and naturalness of translated speech. Overall, the experimental results confirm that accent adaptation improves both speech recognition reliability and multilingual translation accuracy, making the system more suitable for real-world, diverse linguistic environments.



(a)

Timestamp	IP Address	Method	URL Path	Status	Event
14:33 PM	192.168.1.3	GET	/assets/Caps.png	Success	Log
15:04 PM	192.168.1.5	GET	/assets/js/jquery.min.js	Success	Log
15:09 PM	192.168.1.8	GET	/api/users	Success	Log
15:09 PM	192.168.1.10	GET	/api/users	Success	Log
15:16 PM	192.168.1.8	GET	/api/players/matrix?W=191	Success	Log
15:45 PM	192.168.1.3	GET	/.../assets/	Success	Log
15:46 PM	192.168.1.3	GET	/api/config	Success	Log
15:47 PM	192.168.1.34	GET	/api/config	Success	Log

(b)



(c)

VI. CONCLUSION

The “Identification of URL-Based Attacks from IP Data” project represents a significant advancement in cybersecurity by shifting from conventional, reactive defenses to a proactive, data-driven framework. By integrating IP address intelligence, the system adds a vital contextual layer—such as geolocation, ASN reputation, and hosting provider data than traditional URL-only analysis. Leveraging advanced machine learning and deep learning, the system continuously adapts to evolving attack strategies, ensuring high resilience against zero-day attacks that often bypass static, rule-based filters. Its real-time processing and scalable cloud architecture make it a practical solution for any environment, from small businesses to large enterprises. The study clearly shows how AI and data science can effectively solve real-world security challenges, assisting organizations in minimizing economic impact and safeguarding reputation, while contributing to a safer online ecosystem.

VII. REFERENCES

- [1] McGrath et al., “Behind Phishing: An Examination of Phisher Modi Operandi,” In UseNIX workshop on Large Scale Exploits and Emergent Threats (2008).
- [2] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious Proceedings URLs,” of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
- [3] A. Blum, B. Wardman, T. Solorio, and G. Warner, “Lexical Feature Based Phishing URL Detection Using Online Learning,” Proceedings of the ACM Workshop on Artificial Intelligence and Security, 2010
- [4] S. Marchal, J. François, R. State, and T. Engel, “Proactive Discovery of Phishing Related Domain Names,” Research in Attacks, Intrusions and Defenses (RAID), 2014.
- [5] A. Le, A. Markopoulou, and M. Faloutsos, “PhishDef: URL Names Say It All,”
- [6] A. O. Almomani et al., “Phishing Website Detection with Semantic Features,” Applied Soft Computing, vol. 23, pp. 554-569,