HIN-CTIA CYBER THREAT INTELLIGENCE MODELING AND IDENTIFICATION SYSTEM BASED ON HETEROGENEOUS INFORMATION NETWORK

First Author[1], Second Author[2]

[1] Anjana, Master of Computer Application BKIT-Bhalki
[2]Prof . Poojarani, Master of Computer Application BKIT-Bhalki

Abstract -There has been an increase in the number of businesses prepared to use cyber threat intelligence (CTI) to better understand the cyber security landscape. Automatically identifying the danger type of infrastructure nodes for early warning is difficult due to the restricted labels of cyber threat infrastructure nodes included in CTI. To overcome these obstacles, we create the practical system HinCTI, which models cyber threat intelligence and classifies different kinds of threats. To illustrate the semantic connection between infrastructure nodes, we first create a threat intelligence meta-schema. We then apply the CTI model to an HIN simulation. Next, we define a threat Infrastructure similarity measure between threat infrastructure nodes based on meta-paths and meta-graph instances, and we introduce a MIIS measure-based heterogeneous graph convolutional network (GCN) approach for determining the types of infrastructure nodes that pose threats to CTI. To the best of our knowledge, this is the first effort to present a heterogeneous GCN-based method to threat type identification of infrastructure nodes and to model CTI on HIN for threat identification. Extensive tests are run on real-world datasets using HinCTI, and the findings show that our suggested methodology can greatly outperform the state-of-the-art baseline approaches in terms of threat type detection.

Key Words: Cyber threat intelligence, threat type identification, heterogeneous information network, graph convolutional network, threat infrastructure nodes.

1.INTRODUCTION

A growing number of organizations around the world are showing a growing willingness to leverage the open exchange of cyber threat intelligence (CTI) [1] to obtain the overall picture of the rapidly evolving cyber threat situation and to protect themselves from the complex, persistent, organized, and weaponized cyber-attacks. Evidence-based information about a current or emerging danger to assets is what CTI is all about, and it may be used to guide a subject's decision-making about how to respond to the threat [2]. Network infrastructures (such as domain names and Internet Protocol or IP addresses) are often exploited by cybercriminals during attacks.

The bottom three tiers of danger indicators include file hashes, IP addresses, and domain names, as shown in the Pyramid of Pain model [3]. Network security devices like intrusion detection systems (IDS), firewalls, and email server spam filters may all use these three tiers as atomic indications. Because of theusers may get massive quantities of CTI about file hashes, IP addresses, and domain names (i.e. the bottom three levels of the Pyramid of Pain model that are the subject of this research) through the application program interfaces (APIs) given by the threat intelligence sharing platforms (TISPs). Cyber-threat infrastructure nodes may be shown from various angles with the use of many information sources. For instance, not only may a domain name be characterized using data from commercial CTI sources like IBM X-Force

Exchange Platform1 and ThreatBook2, but also from passive domain name system (DNS) and domain name blacklist. Modeling CTI has various benefits [4, 5, 6, 7] in the face of increasingly complex cyber-attacks, such as gaining a comprehensive view of the rapidly shifting cyber threat scenario and revealing possible groups that are behind particular assaults. Nodes in the domain name system, for instance, may be subject to assaults like spam URLs, brute force login attempts, malware, and botnets. The fine-grained danger warning is aided by identifying the threat kinds of infrastructure nodes, which also allows for tailored defensive actions. It is important to note that in this study, we only examine CTI if it is included in a structured data format.

2. Literature survey:

In this article, we go into the world of cyberthreats and the systems that support them. To be more specific, we identify and examine cyber-threat infrastructures with the goal of revealing important participants (owners, domains, IPs, companies, malware families, etc.) and their connections. To this purpose, we offer metrics to quantify the badness of various infrastructure parts by using centrality notions and Google PageRank, both of which originate in graph theory. Furthermore, we provide quantitative analysis of the degree to which various malware samples and families share common infrastructure components in order to identify likely actors in certain assaults. In addition, we look at the historical development of cyber-threat infrastructures to extrapolate trends in cybercrime. The proposed research will allow for the generation of insights and information about cyber-threat infrastructures. Using a data set covering a full year, we find interesting things about cyber-threats and campaigns, key participants in these threats, connections between different parts of the cyber-threat infrastructure, trends of cyber-crime, etc.

There has been a recent uptick in interest in threat intelligence sharing platforms from businesses. The increasing need to defend against sophisticated cyber assaults has led to an open sharing of information and expertise among businesses on risks, vulnerabilities, incidents, and mitigation techniques. We held a focus group with 10 expert stakeholders from security operations centers of different internationally operating enterprises to learn more about potential difficulties with data quality in threat intelligence sharing. The quality of collected, processed, shared, and stored threat intelligence data is investigated, along with a number of other aspects. The study confirms what was already suspected: the limitations and complexities of integrating and consolidating shared threat intelligence from different sources while guaranteeing the data's usefulness for an inhomogeneous group of participants are the primary factors that affect shared threat intelligence data.When it comes to pooled threat intelligence, the quality of the data is crucial. Our research shows that sharing threat information does not introduce any new data quality challenges. Although several threat intelligence sharing systems are being hastily brought to market, some data quality concerns, most notably those concerning scalability and data source integration, need special attention because of the novelty of the threat intelligence sharing field.

Indicators of Compromise (IOC) (such as malware signatures, botnet IPs) are being actively exchanged by security experts through public sources (such as blogs, forums, tweets, etc.) in order to keep up with the ever-changing nature of cyber threats. In order to automate analysis and speed up deployment to different security mechanisms like an intrusion detection system, the material offered in articles, blogs, white papers, etc. may be transformed into the OpenIOC format. There are hundreds of thousands of sources out there generating IOC data at breakneck speeds, making it more difficult for humans to keep up with. However, current Natural Language Processing (NLP) techniques fall short of the high standard (in terms of accuracy and coverage) expected from the IOCs that could serve as direct input to a defense system, stymying efforts to automatically gather such information from unstructured text. In this study, we introduce iACE, a novel approach to AI-powered IOC extraction. We take use of the fact that IOCs in technical publications are often specified in a consistent fashion, with each word being linked to a small group of related terms (such as "download") through well-established grammatical connections.

The growing number of serious security events has sparked fresh interest in teaming together to counteract online dangers. Structured and consistent incident description formats are an absolute prerequisite for facilitating collaboration in the detection and prevention of assaults. Due of their complexity and size, corresponding formats are typically developed for automated processing and interchange. These flaws make it difficult to interpret and comprehend the recorded episode by humans. This is a serious issue since security professionals are crucial to the success and efficacy of any security strategy.

To address these gaps, we present a visual analytics approach for analyzing and enriching semi-structured cyber threat intelligence data by security professionals. Our method integrates a novel storage mechanism for this data with a dynamic display feature for examining and modifying the threat intelligence. We show that our idea can be implemented by utilizing the current standard for reporting cyber security threats called the Structured Threat Information eXpression.

3. Motivation**:**

Without a doubt, the most essential need for any cyber threat protection and warning system should be the modeling of CTI and the identification of danger types of infrastructure nodes.

Many state-of-the-art research, such as [7], [10], and [11], have been conducted on this issue in recent years, attracting the attention of academic and industrial groups in the disciplines of cybersecurity and data mining. Some of them are rather unique and complex, yet they all have at least one of the following two major problems.

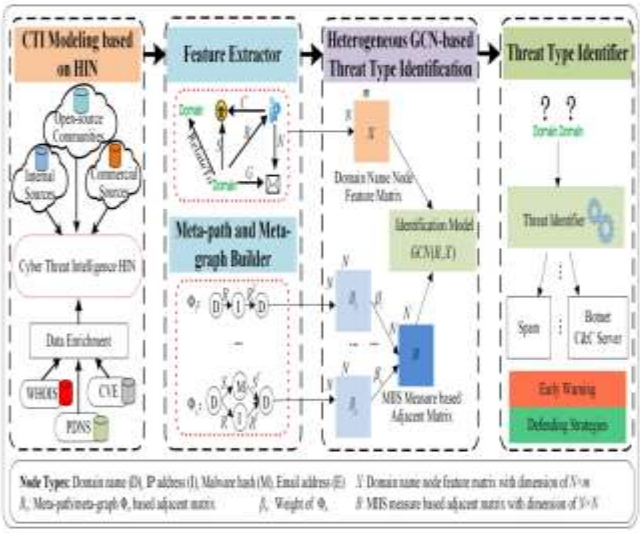
To begin, the issue of incomplete threat category labeling for CTI infrastructure nodes has received very little attention from researchers. Intelligence providers and security analysts mark the threat labels of cyberthreat infrastructure nodes with threat kinds due to the high expense of manual labeling [11]. Thus, most security analysts and operators are concerned with and tasked with figuring out how to reliably and effectively learn from the limited labeled infrastructure nodes and a vast number of interactions among them to anticipate the danger categories of unlabeled nodes

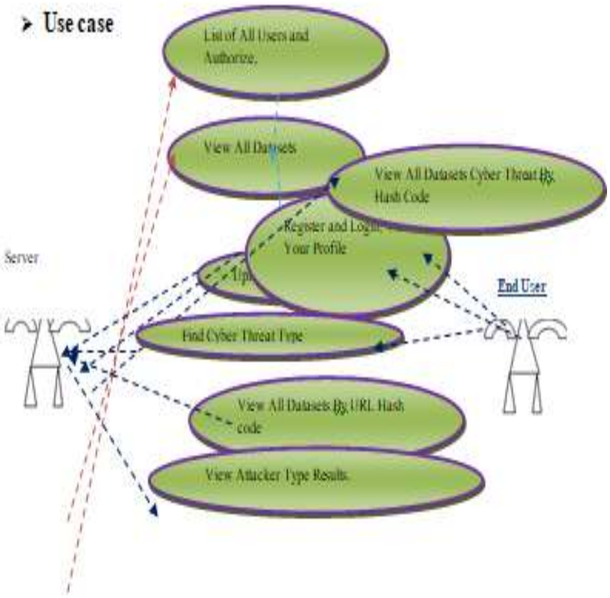
4. SYSTEM ANALYSIS:

Proposed System

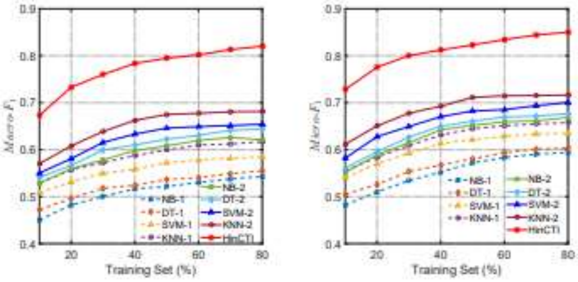
Here, we lay the groundwork for future work by discussing feature extraction and then developing meta-paths and meta-graphs. We next detail a method to threat type detection using heterogeneous GCNs, and lastly illustrate how a hierarchical regularization technique helps to address the issue of overfitting. We zero in on the detection of threats to domain name infrastructure nodes because CTI pertaining to domain names is more stable and efficient than CTI pertaining to other kinds of infrastructure nodes in cybersecurity [3].

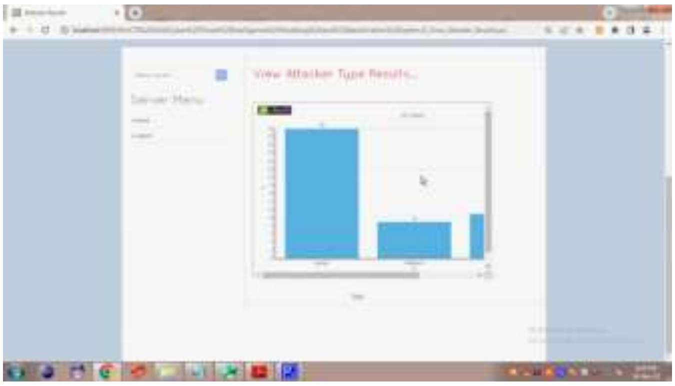
4. ARCHITECTURE


CTI Modeling based on HIN
Feature Extractor
Heterogeneous GCN-based Threat Type Identification
Threat Type Identifier
Open-source Communities
Internal Sources
Commercial Sources
Cyber Threat Intelligence HIN
Data Enrichment
WHOIS
CVE
PDNS
Domain
Domain
Meta-path and Meta-graph Builder
Domain Name Node Feature Matrix
Identification Model GCN(E,X)
MIIS Measure based Adjacent Matrix
Domain Domain
Threat Identifier
Spam
Botnet C&C Server
Early Warning
Defending Strategies
Node Types: Domain name (D), IP address (I), Malware hash (M), Email address (E)
X: Domain name node feature matrix with dimension of N×m
A: Meta-path/meta-graph Φ based adjacent matrix
β: Weight of Φ
E: MIIS measure based adjacent matrix with dimension of N×N


Use case
List of All Users and Authorize.
View All Datasets
View All Datasets Cyber Threat By Hash Code
Register and Login Your Profile
Up
Server
End User
Find Cyber Threat Type
View All Datasets By URL Hash code
View Attacker Type Results

6. Results and Analysis:





Conclusion:

In this study, we offer HinCTI, an HIN-based system for CTI modeling and threat type recognition. To model CTI on HIN, we provide a meta-schema, as well as a collection of meta-paths and meta-graphs, that can extract and include higher-level semantics of cyber-threat infrastructure nodes. We address the problem of insufficient labels for cyber-threat infrastructure nodes by proposing a MIIS measure-based heterogeneous GCN-based method to threat type identification. Our identification method may help reduce overfitting thanks to the use of hierarchical regularization.

Based on real-world dataset experiments, we show that our created system HinCTI that incorporates our suggested technique can considerably outperform the current state-of-the-art baseline approaches in threat type detection.

In order to further enhance the performance of our method, we want to investigate alternative data that may be used to augment the node properties and relations of the cyber threat intelligence HIN. Using topic modeling and NLP methods, future research might focus on extracting fine-grained structured data (such as nodes and their connections) from intelligence reports written in plain language. By doing so, the heterogeneous information network will be substantially enriched, and threat detection performance will be greatly improved.

ACKNOWLEDGEMENT

We would like to extend our deepest appreciation to the manuscript's anonymous reviewers and editors for their insightful remarks and recommendations. The funding for this study came from the NSFC-General Technology FundamentalResearch Joint Fund (grant number U1836215) and the BUPT ExcellentPh.D. Students Foundation (grant number CX2018216).

Grants III-1763325, III-1909323, and CNS-1930941 from the US National Science Foundation helped fund Philip S. Yu's research..

REFERENCES

[1] S. Samtani, M. Abate, V. Benjamin, and W. Li, Cybersecurity as an Industry: A Cyber Threat Intelligence Perspective, pp. 1–20. Cham: Springer International Publishing, 2019.
[2] R. McMillan, “Definition: threat intelligence.” https://www.gartner. com/doc/2487216/definition-threat-intelligence, 2013. Retrieved January, 2019.
[3] D. Bianco, “The Pyramid of Pain.” http://detectrespond.blogspot.com/2013/03/the-pyramid-of-pain.html, 2013.
[4] A. Modi, Z. Sun, A. Panwar, T. Khairnar, Z. Zhao, A. Doupé, G.-J. Ahn, and P. Black, “Towards automated threat intelligence fusion,” in IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), pp. 408–416, IEEE, 2016.
[5] A. Boukhtouta, D. Mouheb, M. Debbabi, O. Alfandi, F. Iqbal, and M. El Barachi, “Graph-theoretic characterization of cyber-threat infrastructures,” Digital Investigation, vol. 14, pp. S3–S15, 2015.
[6] C. Sillaber, C. Sauerwein, A. Mussmann, and R. Breu, “Data quality challenges and future research directions in threat intelligence sharing practice,” in Workshop on Information Sharing and Collaborative Security, pp. 65–70, ACM, 2016.
[7] S. Lee, H. Cho, N. Kim, B. Kim, and J. Park, “Managing cyber threat intelligence in a graph database: Methods of analyzing intrusion sets, threat actors, and campaigns,” in International Conference on Platform Technology and Service (PlatCon), pp. 1–6, IEEE, 2018.
[8] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, “Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence,” in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 755–766, ACM, 2016.
[9] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, “TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources,” in Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 103–115, ACM, 2017.
[10] F. Böhm, F. Menges, and G. Pernul, “Graph-based visual analytics for cyber threat intelligence,” Cybersecurity, vol. 1, no. 1, p. 16, 2018.
[11] U. Noor, Z. Anwar, A. W. Malik, S. Khan, and S. Saleem, “A machine learning framework for investigating data breaches based on semantic analysis of adversary’s attack patterns in threat intelligence repositories,” Future Generation Computer Systems, 2019.