HinCTI A Cyber Threat Intelligence Modeling and Identification System Based on Heterogeneous Information Network

ABSTRACT

Cyber attacks have become increasingly complicated, persistent, organized, and weaponized. Faces with this situation, drives a rising number of organizations across the world are showing a growing willingness to leverage the open exchange of cyber threat intelligence (CTI) for obtaining a full picture of the fast-evolving cyber threat situation and protecting themselves against cyber-attacks. However, modeling CTI is challenging due to the explicit and implicit relationships among CTI and the heterogeneity of cyber-threat infrastructure nodes involved in CTI. Owing to the limited labels of cyber threat infrastructure nodes involved in CTI, automatically identifying the threat type of infrastructure nodes for early warning is also challenging. To tackle these challenges, a practical system called HinCTI is developed for modeling cyber threat intelligence and identifying threat types. We first design a threat intelligence meta-schema to depict the semantic relatedness of infrastructure nodes. We then model cyber threat intelligence on heterogeneous information network (HIN), which can integrate various types of infrastructure nodes and rich relations among them. Following, we define a meta-path and meta-graph instances-based threat Infrastructure similarity (MIIS) measure between threat infrastructure nodes and present a MIIS measure-based heterogeneous graph convolutional network (GCN) approach to identify the threat types of infrastructure nodes involved in CTI. Moreover, through the hierarchical regularization strategy, our model can alleviate the problem of overfitting and achieve good results in the threat type identification of infrastructure nodes. To the best of our knowledge,

this work is the first to model CTI on HIN for threat identification and propose a heterogeneous GCN-based approach for threat type identification of infrastructure nodes. With HinCTI, comprehensive experiments are conducted on real-world datasets, and experimental results demonstrate that our proposed approach can significantly improve the performance of threat type identification compared to the existing state-of-the-art baseline methods. Our work is beneficial to greatly relieve security analysts from heavy analysis work and efficiently protect organizations against cyber-attacks.

**EXISTING SYSTEM**

From the perspective of CTI sharing, numerous exchange formats, such as Structured Threat Information eXpression (STIX) [17], Incident Object Description and Exchange Format (IODEF) [18], and OpenIOC [19], are proposed to describe security incidents and observations related to attack campaigns. However, STIX, IODEF, and OpenIOC are not used for computational purposes. To extract and incorporate higher-level semantics of infrastructure nodes, CTI must be modeled from the perspective of computation.

The modeling of CTI based on multiple intelligence sources (e.g., IBM X-Force Exchange, and ThreatBook) can be very beneficial to discover the correlations among various cyberattack events, facilitate the analysis of cyber attacks, and obtain a complete visibility across Kill Chain phases [20]. For instance, referring to IP and DNS registration information can be useful for malware database, and referring to malware database entries is useful for IP and DNS blacklists wherever appropriate. Likewise, a vulnerability database can refer to any malware samples, which exploit that vulnerability, and vice versa. Modi et al. [4] proposed an automated CTI fusion framework called ATIS, which considers multiple threat sources and connects apparently isolated cyber events. Gascon et al. [21] proposed MANTIS, a platform for CTI that provides a unified presentation for numerous standards and correlates threat data from different sources through a novel type-agnostic similarity algorithm based on attributed graphs. However, the similarity algorithm only considers the similarity of fingerprints (hash values) of any two objects, and the available higher-level semantics (indirect relations involving other types of nodes) are totally neglected.

Boukhtouta et al. [5] presented an approach to investigate cyberthreats, in which tens of types of nodes are considered. However, the higher-level semantics among infrastructure nodes are not further analyzed.

Graph-based threat identification is an important research approach in the fields of network security and data mining, and it offers the characterization of the interaction between infrastructure nodes and the identification of influential entities and groups. By leveraging the linkage information between infrastructure nodes of interest, graph-based methods can uncover the potential relationships, which are relatively harder for attackers to evade because making a cyber attack unavoidably generate

plenty of links in the graph [22].

In recent years, a number of innovative graph-based threat identification methods have been developed for cyber security. However, existing research heavily focuses on homogeneous information networks, which can only perform simple correlation

analysis. Manadhata et al. [13] leveraged graph inference and adapted belief propagation to detect malicious domain names. However, only the host-domain graph is constructed, and ignoring IP-domain graph and other informative graphs greatly hinders the accuracy of identification. Shi et al. [23] proposed a malicious domain name identification approach based on extreme machine learning (ELM), in which construction-based, IP-based, TTL-based, and Whois-based features are extracted to characterize a domain name and fed into ELM.

**Disadvantages**

* The system is not implemented Network Representation Learning for Threat Identification.
* The system is not implemented CTI Modeling based on HIN in which CTI generally refers to cyber-attack-related evidence, involving a group of different types of threat infrastructures, such as malicious IP addresses, malicious domain names, malware hashes, and malicious email addresses.

Proposed System

1) A CTI modeling approach based on HIN is proposed from the perspective of computation (meta-path and meta-graph instances based computing). By modeling CTI based on HIN, the proposed framework can not only integrate infrastructure nodes involved in CTI in a semantically meaningful way, including domain name, IP addresses, malware hashes, email addresses, and their relations but also extract and incorporate higher-level semantics of infrastructure nodes.

2) A MIIS measure-based heterogeneous GCN approach is proposed to identify the threat types of infrastructure nodes. We define a MIIS measure between threat infrastructure nodes, and present a MIIS measure-based heterogeneous GCN approach to identify the threat type of infrastructure nodes. Through hierarchical regularization, the approach can alleviate the problem of overfitting and achieve good results in the threat type identification of infrastructure nodes. This research can also promote cyber security investigations with partial or incomplete information.

3) A practical system called HinCTI is developed for modeling cyber threat intelligence and identifying threat types. With the system, we conduct comprehensive experiments on realworld datasets, and experimental results demonstrate that

our proposed approach can significantly improve the performance of threat type identification compared with the existing state-of-the-art baseline methods.

**Advantages**

The proposed system is implemented Heterogeneous GCN-based Threat Type Identification.

The proposed system is implemented Threat Type Identifier.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - I 3 MORE

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

**Software Requirements:**

* Operating System - Windows 10 AND MORE
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL