Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives

Nan Sun, Ming Ding, *Senior Member, IEEE*, Jiaojiao Jiang, Weikang Xu, Xiaoxing Mo, Yonghang Tai, and Jun Zhang, *Senior Member, IEEE*

***Abstract*—Today's cyber attacks have become more severe and frequent, which calls for a new line of security defenses to protect against them. The dynamic nature of new-generation threats, which are evasive, resilient, and complex, makes traditional security systems based on heuristics and signatures struggle to match. Organizations aim to gather and share real-time cyber threat information and then turn it into threat intelligence for preventing attacks or, at the very least, responding quickly in a proactive manner. Cyber Threat Intelligence (CTI) mining, which uncovers, processes, and analyzes valuable information about cyber threats, is booming. However, most organizations today mainly focus on basic use cases, such as integrating threat data feeds with existing network and firewall systems, intrusion prevention systems, and Security Information and Event Management systems (SIEMs), without taking advantage of the insights that such new intelligence can deliver. In order to make the most of CTI so as to significantly strengthen security postures, we present a comprehensive review of recent research efforts on CTI mining from multiple data sources in this article. Specifically, we provide and devise a taxonomy to summarize the studies on CTI mining based on the intended purposes (i.e., cybersecurity-related entities and events, cyber attack tactics, techniques and procedures, profiles of hackers, indicators of compromise, vulnerability exploits and malware implementation, and threat hunting), along with a comprehensive review of the current state-of-the-art. Lastly, we discuss research challenges and possible future research directions for CTI mining.**

***Index Terms*—Cybersecurity defense, cyber threat intelligence, data mining, natural language processing, machine learning.**

Manuscript received 18 November 2022; revised 29 March 2023; accepted 24 April 2023. Date of publication 5 May 2023; date of current version 23 August 2023. This work was supported by the University of New South Wales, Artificial Intelligence Seed Funding under Project PS66804. *(Corresponding author: Nan Sun.)*

Nan Sun is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: nan.sun@adfa.edu.au).

Ming Ding is with the Data 61, Commonwealth Scientific and Industrial Research Organisation, Sydney, NSW 2015, Australia.

Jiaojiao Jiang and Weikang Xu are with the School of Computer Science and Engineering, University of New South Wales, Kensington, NSW 2052, Australia.

Xiaoxing Mo is with the Faculty of Science, Engineering and Built Environment, Deakin University, Waurn Ponds, VIC 3216, Australia.

Yonghang Tai is with the Yunnan Key Laboratory of Optoelectronic Information Technology, Yunnan Normal University, Kunming 650500, China.

Jun Zhang is with the School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC 3122, Australia.

Digital Object Identifier 10.1109/COMST.2023.3273282

I. Introduction

In THE wake of the massive disruptions that have been caused by the COVID-driven social, economic, and technological changes of the 2020s, cybersecurity adversaries have refined their tradecraft to become even more sophisticated. A series of high-profile attacks followed, such as the SolarWinds supply chain attack [1], which rocked many organizations and marked a turning point in cybersecurity. As the process of collecting, processing, and analyzing information about threat actors' motives, targets, and attack behaviors, Cyber Threat Intelligence (CTI) assists organizations, governments, and individual Internet users in making faster, more informed, data-backed security decisions and changing their behavior in order to fight threat actors from a reactive to a proactive one.

Several definitions exist for CTI. An example of what CTI is defined as is "evidence-based knowledge, including context, mechanisms, indicators, implications, and actionable advice about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard" [2]. In [3], CTI refers to "the set of data collected, assessed and applied regarding security threats, threat actors, exploits, malware, vulnerabilities and compromise indicators". Dalziel [4] describe CTI as "data that has been refined, analyzed, or processed such that it is relevant, actionable, and valuable". Generally speaking, the input of the CTI pipeline is the raw data about cybersecurity, while the output is the knowledge that can help in future decision-making for proactive cybersecurity defense, including strategies for limiting the extent and prevention of cyber attacks.

By using CTI to observe cyber risks, organizations of all shapes and sizes can better understand their attackers, respond quicker to incidents, and proactively get ahead of what threat actors will do in the near future. For small and medium-sized enterprises, CTI data is of great benefit to them because it allows them to access a level of protection they were previously unable to achieve. Meanwhile, enterprises with large security teams can reduce costs and increase the effectiveness of their analysts by leveraging external CTI.

Driven by the increasing awareness of proactively striving to achieve cyber resilience, some research efforts have been made to review related works. The existing surveys

TABLE I
LIST OF ACRONYMS USED THROUGHOUT THIS PAPER

Acronym	Definition
AARs	After Action Reports
AI	Artificial Intelligence
ANN	Artificial Neural Network
APT	Advanced Persistent Threat
BiLSTM	Bidirectional Long Short Term Memory
CRF	Conditional Random Fields
CTAs	Cyber Threat Actors
CTI	Cyber Threat Intelligence
CVE	Common Vulnerabilities and Exposures
DBIR	Data Breach Investigations Report
DDoS	Distributed Denial-of-Service
DIB	Defense Industrial Base
DL	Deep learning
ENISA	European Network and Information Security Agency
FinTech	Financial Technology
FPR	False Positive Rate
FSM	Finite State Machine
GDPR	General Data Protection Regulation
GNN	Graph Neural Network
HIN	Heterogeneous Information Network
HIs	Hazard Indicators
IOCs	Indicators of Compromise
IRC	Internet-Relay-Chat
KG	Knowledge Graph
LDA	Latent Dirichlet Allocation
LTE	Long-Term Evolution
MAEC	Malware Attribute Enumeration and Characterization
ML	Machine Learning
MOGANED	Multi-Order Graph Attention Network based method for Event Detection
NER	Named Entity Recognition
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NN	Neural Network
NTLK	Natural Language Toolkit
NVD	National Vulnerability Database
OSVDB	Open Sourced Vulnerability Database
PoS	Parts of Speech
REGEX	REGular Expression
ROC	Receiver Operating Characteristic
SIEMs	Security Information and Event Management systems
SMOBI	Smoothed Binary
SVM	Support Vector Machine
TTI	Tactical Threat Intelligence
TTPs	Tactics, Techniques, and Procedures

on CTI are summarized in Table II. Specifically, the seminar work [5] presented a study on the darknet as a practical approach to monitoring cyber activities and cybersecurity attacks. This study [5] defined darknet data components as scanning, backscatter, and misconfiguration traffic, and provided a detailed analysis of protocols, applications, and threats using a large volume of data. Case studies such as Conficker worm, Sality SIP scan botnet, and the largest DRDoS attack were used to characterize and define the darknet. The paper also reviewed the contributions of darknet measurement by analyzing data extracted from it, including cyber threats and events and identified technologies related to the darknet. Additionally, Robertson et al. [6] proposed a system consisting of a crawler, parser, and classifier to locate sites where security analysts can gather information, as well as a game theory-based framework for simulating an attacker and defender in the process of CTI mining and analyzing as a security game involving past attacks and security experts.

Further, Tounsi and Rais [7] classified the existing threat intelligence types into strategic threat intelligence, operational threat intelligence, and tactical threat intelligence. With the focus mainly on the Tactical Threat Intelligence (TTI) that was mainly generated from the Indicators of Compromise (IOCs), the work [7] provided a comprehensive study on the TTI issues, emerging research trends, and standards. With the advancements in Artificial Intelligence (AI), Ibrahim et al. provided a brief discussion on how to apply AI and Machine Learning (ML) approaches to leverage CTI to stop data breaches. Rahman et al. [11], [12] further provided a holistic discussion of various technologies in the area of ML and Natural Language Processing (NLP) for automatically extracting CTI from the textual descriptions. As the usage of CTI is one of the key steps to maximizing its effectiveness, Wagner et al. [8] reported the investigation on the state-of-the-art approaches to sharing CTI and the associated challenges of automating the sharing process with both the technical and non-technical challenges. Abu et al. [9] gave an overall survey on CTI definition, issues and challenges. Ramsdale et al. [14] summarized the current landscape of available formats and languages for sharing CTI. They also analyzed a sample of CTI feeds, including the data they contain and the challenges associated with aggregating and sharing that data.

Beyond the research works on CTI, the use and implementation of CTI is a common practice in government organizations and enterprises, reflecting the growing recognition of the critical importance of cyber security. These two parties have dedicated teams responsible for collecting, analyzing, and disseminating threat intelligence information, often through specialized CTI platforms and tools. For example, the Information Sharing and Analysis Center (ISACs) are centralized nonprofit organizations that are established to facilitate the sharing of CTI and other security-related information among their members. ISACs serve a variety of industries and sectors, including critical infrastructure, financial services, healthcare, technology, and others. They bring together organizations from within a specific industry or sector to share threat intelligence and best practices, as well as collaborate on incident response and mitigation efforts. ISACs are often supported by government agencies and other organizations, and they typically follow strict security and privacy protocols to ensure that sensitive information is protected and shared only among authorized individuals.

According to the 2022 Crowdstrike threat intelligence report, CTI is increasingly being recognized as a valuable asset, with 72 percent planning to spend more on it over the next three months in 2022 [15]. Government organizations and enterprises alike are investing significant resources into enhancing their CTI capabilities, recognizing that staying ahead of the constantly evolving threat landscape requires continuous improvement and adaptation. Such efforts include the development of in-house expertise, the establishment of partnerships with other organizations and industry leaders, and the use of cutting-edge technologies and methodologies. The efforts made by government organizations and enterprises to

TABLE II
OUR NOVEL CONTRIBUTIONS IN CYBER THREAT INTELLIGENCE MINING AND HOW THEY DIFFER FROM PREVIOUS SURVEYS. UNDER THE CATEGORY OF MAIN TOPICS, "●","◑", AND "○" REPRESENT COMPREHENSIVE REVIEW, PARTIAL REVIEW, AND NOT REVIEW, RESPECTIVELY

Related works	Main topics			Key contributions
	CTI Sharing	CTI Capabilities	CTI Mining	
[5]	○	○	◑	A survey on the *darknet* as a source of generating CTI
[6]	○	○	◑	A book on *darkweb* CTI mining and *game theory* applied to CTI
[7]	◑	◑	○	A survey on *sharing* CTI, standard CTI structures, and CTI sources
[8]	●	○	○	A survey on *sharing* CTI and discussion on technical and non-technical challenges for CTI sharing communities
[9]	○	◑	○	An overall survey on CTI definition, issues and challenges
[10]	○	○	◑	A short discussion on the challenges of implementing CTI and the future of effective CTI by applying *AI* and *ML* approaches
[11] & [12]	○	◑	◑	A holistic discussion of various essential technologies in the area of *ML* and *NLP* for CTI automatic extraction
[13]	○	◑	◑	A general overview of CTI, including its usage, value, and the requirements and processes involved in CTI teams *within organizations*
Our paper	◑	◑	●	A comprehensive survey on mining CTI for proactive security defense. Particularly, (1) A comprehensive review is conducted in line with the summarization of a six-step methodology for *CTI mining* for *proactive cybersecurity defense*. (2) An in-depth analysis of the state-of-the-art solutions is elaborated with the proposed taxonomies of *CTI knowledge acquisition*. (3) With our hope of helping other researchers expand their views on this topic, we discuss challenges and open research issues with the identification of new trends and future directions.

improve their CTI capabilities demonstrate the commitment to protecting their critical assets and safeguarding against the risks posed by cyber threats. CTI is a crucial component of a comprehensive cyber security strategy and an essential tool in the ongoing efforts to secure digital systems and networks for organizations and enterprises. Furthermore, according to the 2022 SANS CTI survey conducted by Brown and Stirparo [13], 75 percent of the participants believe that CTI improves their organization's security prediction, threat detection, and response. The survey also revealed that 52 percent of the respondents considered detailed and timely information as the most crucial characteristic for the future of CTI.

As a result of the surge in cyber attacks in recent years, a large number of attack artifacts have been reported extensively by public online sources and actively collected by different organizations [16], [17]. By mining CTI, organizations can discover evidence-based threats and improve their security posture by detecting early signs of threats and continuously improving their security controls. The source data for mining CTI can be retrieved from private channels, such as company internal network logs, as well as public channels, such as technical blogs or publicly available cybersecurity reports. In particular, cybersecurity information written in natural language comprises the majority of the CTI. Cybersecurity-related data can be gathered from a wide variety of sources, and this provides a stepping stone on the path towards mining CTI. However, mining robust, actionable, and genuine CTI while keeping pace with the rapidly increasing cybersecurity-related information is challenging. Although there is a positive trend towards higher levels of context, analysis, and relevance of CTI, 21 percent of the participants in the 2022 SANS CTI survey [13] do not perceive any improvement in their organization's overall security situation due to CTI. Currently, many organizations concentrate on fundamental usage scenarios that involve merging threat data feeds with their current network and firewall systems, intrusion prevention systems, and Security Information and Event Management systems (SIEMs). However, they do not make the most of the valuable knowledge that such new intelligence can provide. Consequently, it is important to study CTI mining consumption at fine granularities to develop effective tools. To be specific, to investigate what kind of CTI can be obtained through CTI mining, the methodology to achieve it, and how to use the acquired artifacts as proactive cybersecurity defense. Based on the above motivation, we conduct a comprehensive literature review of how CTI can be acquired from diverse data sources, especially from information written in the form of natural language texts from various data sources, to defend against cybersecurity attacks proactively. This perspective has not been explored in the existing survey works despite the fact that CTI has been extensively studied in the previous literature review.

The primary focus of this paper is to review recent studies on CTI mining. In particular, our work provides a summary of the CTI mining techniques and the CTI knowledge acquisition taxonomy. Our article presents a taxonomy that classifies CTI mining studies based on their objectives. Additionally, we offer a comprehensive analysis of the latest research on CTI mining. We also examine the challenges encountered in CTI mining research and suggest future research directions to address these issues. Below is a summary of the contributions highlighted in this paper:

- Our review summarizes a six-step methodology that transforms cybersecurity-related information into evidence-based knowledge through perception, comprehension, and projection for proactive cybersecurity defense using CTI mining.
- We collect and review the state-of-the-art solutions and provide an in-depth analysis of collected work with the proposed taxonomies based on CTI consumption, particularly seeing through the eyes of attackers for proactively defending against cyber threats.
- As part of our efforts to expand the perspectives of other researchers and CTI communities, we discuss challenges

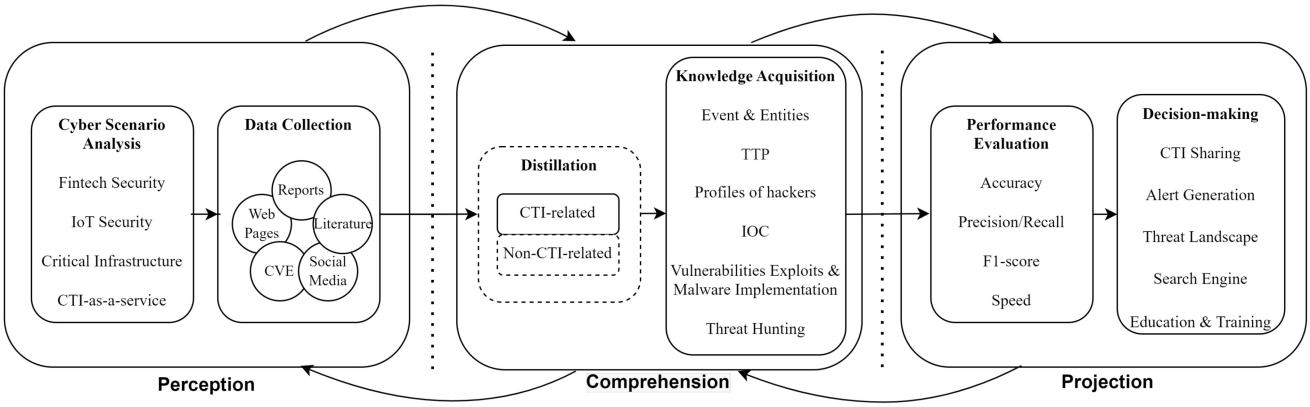


Fig. 1. Methodology of Cyber Threat Intelligence Mining for Proactive Security Defense.

and open research issues as well as identify new trends and future directions.

As follows is an overview of this survey. Firstly, Section II provides an overview of CTI mining, including its methodology of CTI mining and taxonomy. Section III presents a comprehensive review of existing work in the field of CTI mining according to our proposed taxonomy. Section IV discusses the challenges and future direction in this area. Finally, Section V concludes the paper. Table I lists and describes the acronyms used throughout this paper.

II. Cyber Threat Intelligence Mining Methodology and Taxonomy

Based on the surveyed papers, we summarize the methodology for CTI mining and the taxonomy for CTI knowledge acquisition. The process of CTI mining gradually evolved people's insights about cybersecurity from the perception of data in the environment to an understanding of the meaning of the data and finally to a projection of future decisions. Moreover, the taxonomy summarizes the most valuable information for various purposes of CTI mining and provides a new perspective on CTI mining.

A. Research Methodology

As shown in Figure 1, the methodology consists of six steps: cyber scenario analysis, data collection, CTI-related information distillation, CTI knowledge acquisition, performance evaluation, and decision-making. Cyber scenario analysis and data collection enable the perception of the specific environment across time and space. The data distillation and CTI knowledge acquisition help the comprehension of the data perceived in the previous steps by locating the targets and acquiring useful information. The last two steps, evaluation and decision-making, constitute the projection stage, where decisions are made efficiently and effectively.

1) Step 1 - Cyber Scenario Analysis: CTI mining is a process for turning raw data into actionable intelligence for decision-making and taking immediate action as needed. As the first step of the threat intelligence lifecycle, the cyber scenario analysis stage is crucial because it sets the roadmap for specific threat intelligence operations that will be conducted in the future. There are a variety of primary cyber scenarios in the reviewed studies, including Fintech security, IoT security, critical infrastructure security, and cloud-based CTI as a service. There will be a planning stage where the team will agree on the goals as well as the methodology of their intelligence program based on the requirements of the cyber scenario with various stakeholders involved in the project. Among the things the team may discover are: (1) What the attackers are and what their motivations are, as well as who they are in a specific cyber scenario? (2) Is there a surface area that is vulnerable to attacks? (3) How can their defenses be strengthened in the event of an attack in the future? Examples of primary cyber scenarios in our reviewed studies: Fintech security, IoT security, critical infrastructure, and CTI-as-a-service.

2) Step 2 - Data Collection: As a way of protecting organizations and the security community against fast-evolving cyber threats, many efforts have been made for sharing threat intelligence. There is no doubt that public sources are a significant contributor to CTI, regardless of the platform used to access it. To share unclassified CTIs, a few approaches such as AlienVault OTX [18], OpenIOC DB [19], IOC Bucket [20], and Facebook ThreatExchange [21] have been established. The information shared on these platforms can help organizations identify and mitigate security risks, prioritize their security efforts, and respond more effectively to cyber threats. As an example of a crowd-sourced platform, Facebook ThreatExchange [21] is open to any organization and allows participants to share real-time threat intelligence information, including information about malware, phishing campaigns, and other types of cyber attacks. The CTI data are usually available for Web crawling once published on online platforms. For example, we can obtain vulnerability records from the National Vulnerability Database (NVD) [22] as well as historical data breach reports in Verizon's annual Data Breach Investigations Reports (DBIR) [23]. Data generated by technical sources (i,e., security tools and systems) including log files, network traffic, and system alerts, were used as valuable sources for predicting cybersecurity incidents [24]. In addition, APIs are provided by various kinds of social media, such as Twitter, to analyze the data within these social media sites and collect threat information shared by individuals and organizations. For the restricted assessed CTI, platforms such as the Defense Industrial Base (DIB) voluntary information sharing program [25] have been created

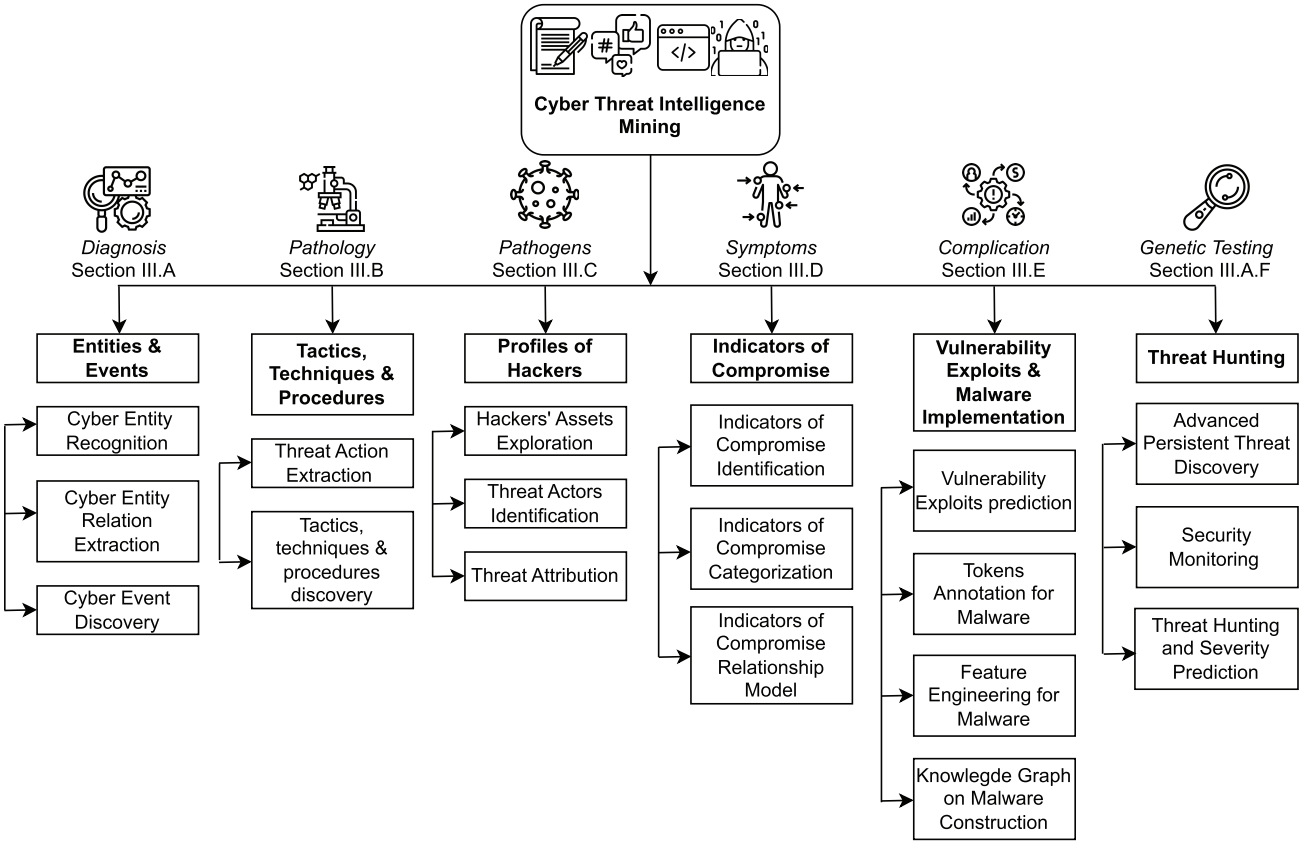


Fig. 2. Taxonomy of Cyber Threat Intelligence Mining for Proactive Security Defense.

to help organizations better protect themselves and their customers from cyber threats. These platforms provide a secure and collaborative environment for exchanging threat intelligence information between certified participants. For example, the DIB voluntary information sharing program restricted to DIB participants only is specifically designed for the Defense Industrial Base and is aimed at improving the security and resilience of the DIB against cyber threats. The program allows DIB participants to share threat intelligence information and to work together to enhance the security of the DIB against cyber threats, foreign interference, and other security risks. Last but not least, it is worth mentioning that illegal online marketplaces and forums through dark Web sources can provide information about ongoing cyber threats.

3) Step 3 - CTI-Related Information Distillation: After collecting data, it is important to distill information (i.e., articles, paragraphs, or sentences) that are related to CTI in order to prepare for the CTI knowledge acquisition. Classification is one of the widely adopted approaches for classifying the pieces of target information related or unrelated to CTI. Using examples from a variety of annotated classes (e.g., CTI-related or non-CTI-related), researchers have built machine-learning classification models to predict the classes of unseen data. Unsupervised machine learning algorithms can be considered as an alternative method of distilling information associated with CTI based on the similarity between the contents of the data clustered together.

4) Step 4 - CTI Knowledge Acquisition: Following the completion of the CTI-related information distillation, it is necessary to conduct data analysis in the form of CTI knowledge acquisition to pinpoint and locate pertinent and accurate information based on the users' requirements. The researchers and CTI community have employed NLP and ML techniques to extract CTI from textual data. Figure 2 shows a detailed taxonomy of the six specific categories of CTI knowledge acquisition based on the collected literature, respectively cybersecurity-related entities and events, cyber attack tactics, techniques and procedures, the profiles of hackers, indicators of compromise, vulnerability exploits and malware implementation, and threat hunting.

5) Step 5 - Performance Evaluation: In the fifth step, we evaluate the extracted CTI's performance against our expected objectives. It is usually measured according to various metrics in order to assess performance. Most classification or clustering tasks involve using a few standard metrics, including accuracy, recall, precision, False Positive Rate (FPR), and F1-score. In order to depict the trade-offs between benefits and costs, graphical plots are used, such as Receiver Operating Characteristic (ROC) curves with the TPR plotted on the y-axis and the FPR plotted on the x-axis. The area under the ROC curve indicates the strength of ROC curves cumulatively. Furthermore, there is a high expectation that less time will be spent on extracting requested information with the real-time CTI experience. A major challenge for cybersecurity tasks, including CTI knowledge acquisition, is often FPR because the false alarms result in excessive costs associated with manual verification, which, in many cases, is the result of the false alarms. In a way that has never been seen before, an emerging

CTI is expected to discover, for the first time, that the goal of pursuing performance is usually to maximize TPR while minimizing FPR. It is possible to determine whether a specific CTI knowledge acquisition approach produces satisfactory results by leveraging comprehensive evaluation metrics. If unsatisfactory results are achieved, it is recommended to repeat the process with the required alternations.

6) Step 6 - Decision-Making: Depending on how CTI is extracted within different categories, it can be used for a variety of purposes for decision-making. Following is a summary of key applications of acquired CTI in the process of decision-making, including CTI sharing, alert generation, threat landscape, search engine, education, and countermeasures.

CTI sharing: It is a practice in which a variety of information relating to cybersecurity is shared in order to identify risks, vulnerabilities, threats and internal security issues as well as to share good practices in this regard. The extracted CTI under various categories is expected to be shared between multiple organizations, including government agencies, IT security firms, cybersecurity researchers, etc. CTI sharing is typically driven by legal and regulatory factors (e.g., General Data Protection Regulation (GDPR) [26]), as well as economic factors (e.g., reducing the cost of resolving the consequences of data breaches).

Alert generation: According to the definition from National Institute of Standards and Technology (NIST) [27], information about a specific attack directed at an organization's information systems is called an alert in cybersecurity. An alert regarding current vulnerabilities, exploits, and other security issues that are usually human-readable can be generated directly from the extracted CTI under various categories. Several outputs can be produced, including vulnerability notes, bulletins, and recommendations.

Threat landscape: The threat landscape refers to the full spectrum of potential and recognized cybersecurity threats affecting specific industries, organizations, or user groups in a particular period. The threat landscape is constantly changing as new cyber threats emerge every day. Using the extracted CTI from the text, security experts can gain a deeper understanding of the threat landscape based on the extracted CTI.

Cybersecurity domain search Engine: The extracted CTI can serve as the basis of a cybersecurity search engine. Generally speaking, information retrieval refers to the science of finding information from text, images, and sounds, as well as information from metadata that describes the data that are being searched for [28]. Through search engines, information can be found on the Internet. Cybersecurity domain search engines are increasingly focusing on explainable cybersecurity contexts to emphasize that the amount of information users digest does not depend on the number returned, but rather on their understanding of the returned information. For example, Shodan [29] is a cybersecurity search engine for Internet-connected devices.

Education and training: There is currently a shortage of qualified cybersecurity professionals throughout the world at the moment. This shortage could reach 18,000 in Australia by 2023, according to AustCyber. By providing explainable and structured illustrations of the cybersecurity context, the extracted CTI will contribute to cybersecurity education and training. On the one hand, the education system helps address the shortage of skilled cyber professionals by building a pipeline of skilled professionals in the industry. On the other hand, cybersecurity education is also expected to help people who lack a solid understanding of cybersecurity domain knowledge increase their awareness of cybersecurity incidents and threats.

Risk management: By using CTI, organizations can enhance their risk management procedures with access to valuable intelligence on the most recent vulnerabilities, attack methods, and exploits. Keeping current with emerging risks and vulnerabilities can enable organizations to adopt preemptive measures to identify and manage risks before they are exploited, ultimately reducing the potential cost and impact of a security incident.

B. Cyber Threat Intelligence Mining Definition and Taxonomy

As far as we know, there is no formal definition of *Cyber Threat Intelligence Mining*. However, the definition of data mining has been proposed by several researchers and practitioners in the field of computer science, statistics, and data analysis. According to the definition from IBM, data mining, also known as knowledge discovery in data, is the process of uncovering patterns and other valuable information from large datasets. As one of the most widely cited definitions provided by Fayyad et al. [30], "Data mining is the application of specific algorithms for extracting patterns from data". Chakrabarti et al. [31] further explained the definition from Fayyad et al. [30] as "the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems". By limiting the scope of data in the concept of data mining, in this survey, we define *Cyber Threat Intelligence Mining* as the collection and analysis of large amounts of information from various Cyber Threat Intelligence data sources to identify information relating to cyber threats, attacks, and harmful events.

As introduced in Section II-A, the methodology of CTI mining, as shown in Figure 1, essentially turns the data broadly related to cybersecurity into the digestible CTI for final decision-making. As the bridge linking the perception and projection stages, the comprehension stage plays a role in distilling information related to CTI only and locating useful information according to various goals. As shown in Figure 2, using the stages of comprehension of CTI as a starting point, we categorize the reviewed work on CTI mining based on the aims of CTI knowledge acquisition. To shed more light on the rationale behind the identified six categories of CTI mining, in the following, we draw an analogy between CTI mining and a generic disease-treatment process.

1) Cybersecurity Related Entities and Events: The identification of cybersecurity-related entities and events in CTI mining is like a *diagnosis* step that identifies the nature of

a particular illness or disease. In cybersecurity entity and event extraction, named entities in the unstructured text are located and classified into predefined cybersecurity categories, such as impacted organizations, locations, vulnerabilities, etc, while events are classified into predefined cyber attack categories, such as phishing, Distributed Denial-of-Service (DDoS) attacks, etc.

2) Cyber Attack Tactics, Techniques, and Procedures: In this task category, the goal is to determine how cyber threat actors and hackers prepare and execute cyber attacks by analyzing their Tactics, Techniques, and Procedures (TTPs). This is analogous to *pathology* study in healthcare, which aims to understand the causes and effects of disease or injury.

3) The Profiles of Hackers: The third category in our taxonomy of CTI mining is called profiles of hackers which trace the origin of cyber attacks. The establishment of a hacker profile aims to uncover the sources and resources of a threat actor, including cyber threat attribution and hacker assets. This is similar to the identification of *pathogens* in biology, which refers to the step of finding any organism or agent (e.g., a bacterium or virus) that can produce disease.

4) Indicators of Compromise: The extraction of IoCs aims to find pieces of forensic data that provide evidence of potentially malicious activity on an organization's system, for example, the names, signatures, and hashes of malware. IOCs are similar to physical or mental *symptoms* which indicates a condition of disease.

5) Vulnerability Exploits and Malware Implementation: This category includes literature on studies analyzed documentation, such as literature and user manuals, to discover vulnerabilities under a particular product or service, predict exploits, and find information about malware implementation for predicting software characteristics. Like the *complication* of potential disease, exploiting vulnerabilities and implementing malware is highly relevant to the consequences of cyber threats.

6) Threat Hunting: The purpose of this category of task is to identify previously unknown or ongoing non-remediated threats within an organization's network. This process can be analogous to the genetic testing conducted in a generic disease-treatment process, which predicts the likelihood of a healthy individual developing a specific disease in the future [32].

III. State-of-the-Art Studies: A Proactive Defense Perspective

A. Cybersecurity Related Entities and Events

Cybersecurity attacks and incidents are widespread and have a wide range of consequences and implications, from data leaks to the potential loss of life and disruption of critical infrastructure [24]. It is crucial to develop cyber defenses based on the authoritative record of cyber events reported in the media as well as their key dimensions (e.g., exploited vulnerability, impacted system, duration of events). Cybersecurity event details recorded at fine granularity can assist various analytics efforts, including identifying cyber attacks, developing predictive indicators of attacks, tracking cyber attacks over time and space, and integrating them into cybersecurity graphs to assist automated analysis. In this section, we review the corresponding works that acquire knowledge about the cybersecurity related entities and events through CTI mining.

1) Summary of Representative Work: The entity extraction technique in NLP automatically extracts specific data from unstructured text and categorizes it based on predefined categories. Furthermore, knowledge of the entities present in a sentence can provide information that is useful for confirming the category of events and predicting event triggers. Researchers are studying cybersecurity related entities and events extraction for CTI mining, which is key to dealing with heterogeneous data sources and the huge volume of cybersecurity related information. A summary of the survey of representative studies is listed in Table III.

As a preliminary study, several approaches [33], [34] were proposed to quickly extract cybersecurity events without labeled data for the training process. A weakly supervised ML approach was proposed in [34] with no training phase requirement to extract events from Twitter stream data rapidly. The study [34] focuses on three high-impact categories of cybersecurity attacks, including data breach, DDoS and account hijacking, to demonstrate how to identify cybersecurity events based on convolution kernels and dependency parses. The highest precision in successfully detecting cybersecurity-related events can obtain 80% in this work [34]. In addition, work [33] utilized an unsupervised ML model (i.e., Latent Dirichlet Allocation (LDA)) to cluster the relevant posts in hacker forums, which demonstrates a method that can effectively extract CTI in the aspect of cybersecurity events. Although Deliu et al. [33] only evaluated the performance of the estimated cybersecurity events on the number of topics and time elapsed, the work demonstrated the approach for quickly extracting relevant cybersecurity topics and events.

The categories of automatically identified cybersecurity related entities and events have grown with the introduction of datasets with annotations and the development of NLP and deep learning techniques. Dionísio et al. [35] annotated cybersecurity related Twitter data with 5 categories of entities (as shown in Table III) that considers descriptions from the European Network and Information Security Agency (ENISA) risk management glossary [39]. In this work [35], the Bidirectional Long Short Term Memory (BiLSTM) Neural Network (NN) were implemented for name entity recognition. Pre-trained word embeddings that refer to embeddings learned in one particular task that is used for solving another similar task, including GloVE [40] and Word2Vec [41] were applied to provide a starting point for the semantic value. The BiLSTM model achieved an average F1-score of 92% in recognizing the six categories of cybersecurity related entities. The annotated data (i.e., cybersecurity related entities) built in work [35] are publicly available through their GitHub website,[1] which provides the groundtruth for name entity recognition in CTI domain. Satyapanich et al. [36]

[1] https://github.com/ndionysus/twitter-cyberthreat-detection

TABLE III
LATEST WORKS ON MINING CYBERSECURITY RELATED ENTITIES AND EVENTS

Work	Categorization of Models	Features	Pre-trained word embeddings	Performance	Data	Dataset Publicly Available	Entities or Events	# of Entities/Events
[33]	ML, LDA	Three character trigrams	No	Evaluation on the # of topics and time elapsed	Hacker forums	N	Events	5
[34]	Weakly supervised approach with no training phase requirement	Dependency parse trees, word embeddings	Word2Vec	Highest precision 80%	Twitter	N	Events	3
[35]	NER, NN, BiLSTM	Word-based embeddings, character-based embeddings	GloVE, Word2Vec	Average F-1 score 92%	Twitter	Y	Entities	5
[36]	NER, NN, BiLSTM	Word-based embeddings, linguistic features	BERT, Word2Vec	Highest F-1 score 79.94%	Cybersecurity articles	Y	Events	5
[36]	NER, NN with attention, BiLSTM	Word-based embeddings, linguistic features	BERT, Word2Vec	Highest F-1 score 74.76%	Cybersecurity articles	Y	Entities	20
[37]	NER, GNN, BiLSTM	Word-based embeddings, character-based embeddings, graph construction	Word2Vec	Average F-1 score 90.28%	Cybersecurity articles	N	Entities	4
[38]	NN, GNN, LSTM, BiLSTM, Hierarchical and Bias Tagging Networks and Gated Multi-level Attention Mechanisms	Dependency parse trees, word embeddings	BERT, Word2Vec	Highest F-1 score 68.4%	Cybersecurity articles	N	Events	4 (with 30 subcategories)

further expanded additional cybersecurity related entities and events by creating a corpus[2] of 1,000 English news articles that were labeled with rich, event-based annotations which covers cyber attacks and vulnerability related cybersecurity attacks. Along with the BiLSTM layer, the work [36] also applied attention mechanisms that have been used and proved with great advancement in NLP for learning the highlighted important parts of the text. In addition, the work [36] used Word2Vec [41] and BERT [42] embeddings in the word embedding layers, and further concatenated the embedding linguistics features to form the embedding layers, including Parts of Speech (PoS), position of the words, etc. Totally, there are 20 cybersecurity related entities (e.g., file, device, software) and 5 events (e.g., phishing) defined and can be automatically detected through the proposed approach [36].

The Graph Neural Network (GNN) that represents data as graphs aims to learn features from the graph level to classify nodes, which began to be applied in the field of information extraction [43]. The complexity of entities in the field of cybersecurity makes it difficult to capture non-local and non-sequential dependencies in name entity recognition [37]. Hence, the recent research [37], [38] proposed to use both local context and graph-level non-local dependencies extracted by GNN to conduct cybersecurity entity recognition. In the work [37], Fang et al. aimed to identify four types of entities from the cybersecurity articles, which are composed of PERSON (PER), ORGANIZATION (ORG), LOCATION (LOC) and SECURITY (SEC). During the process of graph construction, each node in the graph represented a word in each sentence and each edge constructed local context dependencies and non-local dependencies. In addition, the word level embeddings (i.e., Word2Vec [41]) and character level embeddings that capture the contextual information of the words in the sentence were applied. The CyberEyes model proposed in the work [37] can finally obtain an F1-score of 90.28% for the four types of cybersecurity entities. Trong et al. [38] annotated a large dataset that includes 30 subcategories cybersecurity events under four different stages of a cyber attack, respectively DISCOVER, PATCH, ATTACK and IMPACT. The state-of-the-art Multi-Order Graph Attention Network based method for Event Detection (MOGANED) and Attention [44] was applied with Word2Vec [41] and BERT [42] embeddings. Although the highest F1-score of cybersecurity event extraction achieved is 68.4% for their annotated dataset [38] by using a Document Embedding Enhanced Bidirectional Recurrent Neural Network

[2]https://github.com/Ebiquity/CASIE

TABLE IV
CYBERSECURITY RELATED ENTITIES IN REPRESENTATIVE WORKS

Label	Description	Papers
ORG	Company or organization	[35], [36], [37]
PRO	A product or asset	[35]
VER	A version number, possibly from the identified asset or product	[35]
VUL	May be referencing the existence of a threat or a vulnerability	[35], [36]
ID	An identifier, either from a public repository such as the National Vulnerability Database (NVD) [21], or from an update or patch	[35]
PER	Person	[36], [37]
LOC/GPE	Location	[36], [37]
SEC	The entity related to cybersecurity, including companies that provide security services and products, or conduct security research, and viruses or worms, APT organizations, security terminologies, etc.	[37]
Device	A mention of an electronic device that refers to an entity of type Device.	[36]
Data	A mention of information stolen from victims and refers to an entity of type Data.	[36]
PII	Any data that could identify a specific individual, i.e., any information that can be used to distinguish one person from another.	[36]
File	A mention of computer files.	[36]
Malware	A mention of malicious software	[36]
Patch	A mention of a new release or an update of Product, which is developed to fixed a vulnerability.	[36]
Product	A mention of a product.	[36]
Website	A mention of a specific type of system.	[36]
Number	The number of data or victims	[36]
Payment method	A mention that explained a method to pay a ransom specified in the threat from an attacker.	[36]
Money	A mention that consists of a number and a currency referred to as money.	[36]
Time	Date, time, duration that can be inferred from a day or period of time mentioned in an event	[36]
Version	A mention to specify the version of software, system, and device.	[36]
Capabilities	A phrase which explained an attack pattern, how to trick the victim, including an exploit of the vulnerability.	[36]
CVE	An identification number cassigned to a publicly known vulnerability.	[36]
Purpose	A mention of an attacker's purpose.	[36]

(RNN). When MOGANED with BERT was applied to the cybersecurity entities datasets proposed by [36], the F1-score was increased by 6.56% to 86.5%.

2) Discussion: The previous subsection reviewed seven representative studies mining cybersecurity related entities and events. A summary of the surveyed studies is presented in Table III, where we showed the critical difference in each work. Particularly, cybersecurity related entities and events defined in these studies are summarized in Table IV and Table V.

In our reviewed studies, the main techniques used in mining cybersecurity entities and events are divided into the following categories: (1) Unsupervised learning approaches, in which unsupervised algorithms are used without hand-labeled training examples; (2) Supervised learning approaches that use feature engineering in conjunction with supervised learning algorithms. The majority of the reviewed works have adopted Deep Learning (DL) based approaches that automatically discover classification representations by learning hierarchical representations of the data through multiple layers in a Neural Network. DL based approaches are particularly effective at detecting cybersecurity-related entities and events and growing rapidly. Traditional feature-based approaches require a significant amount of feature engineering skills and domain expertise, but data mining based on DL effectively learns useful representations and underlying factors from raw data. With DL, features for entity recognition can be designed in a more efficient manner. In addition, non-linear activation functions enable DL based models to learn complex and intricate

TABLE V
CYBERSECURITY RELATED EVENTS IN REPRESENTATIVE WORKS

Label	Description	Papers
Data Breach	Seed query: data leak, security breach, information stolen, password stolen, hacker stole	[34]
DDoS	Seed query: DDoS attack, slow internet, network infiltrated, malicious activity, vulnerability exploits, phishing attack	[34]
Account Hijacking	Seed query: unauthorized access, stolen identity, hacked account	[34]
DISCOVER	A vulnerability in a software or system is detected or mentioned by some entity (i.e., hackers, engineers).	[38]
PATCH	Some entity (i.e., software companies) fixes or shows how to fix a known vulnerability.	[38]
ATTACK	An attacker exploits some vulnerability to impact the systems using some means. This can be a mention of an attack or the actions involved in the attack.	[38]
IMPACT	The consequence of an attack for a system.	[38]
Attack.Databreach	An attacker compromises a system and removes data.	[36]
Attack.Phishing	An attacker imitates another entity, in an attempt to get a victim to access malicious materials, such as a website or attachments.	[36]
Attack.Ransom	An attacker breaks into a system and encrypts data, and will only decrypt the data for a ransom payment.	[36]
Discover.Vulnerability	A security expert or other entity, like a company, finds a software vulnerability. The additional roles of Discover.	[36]
Patch.Vulnerability	A software company addresses a known vulnerability by releasing or describing an appropriate update.	[36]
Leaked Credentials	Commonly used passwords, top-level domains of compromised accounts	[33]
Malicious Proxies	Malicious proxy servers	[33]
Undetected Malware	Malware evaded detection	[33]
Asset Specific CTI	Identify CTI for a specific asset or organization	[33]

features from data. Compared with linear models (e.g., linear chain Conditional Random Fields (CRF)), the non-linear mappings are generated from input to output, which benefits cybersecurity entities and events recognition.

A comparative study of the reviewed works shows that they all rely on unstructured texts such as tweets, security articles, and hacker forums. This indicates a pressing need for a structured database to store CTI data. Among the different models used, those employing Name Entity Recognition (NER) method, neural network, and BiLSTM perform better. This is because NER can identify and extract entities in sentences, ensuring that irrelevant words are not considered as CTI entities, leading to better performance. Furthermore, the two works with the highest F-1 score, namely [35] and [36], utilize character-based embedding to complement the deficiency of word-based embedding. Character-based embedding can capture morphological information such as prefixes and suffixes, which may be lost in word-based embedding, leading to more accurate and robust performance. Overall, these findings suggest that the use of NER and character-based embedding could significantly enhance the accuracy and effectiveness of CTI models in identifying and mitigating cyber threats.

In the context of natural language processing, the word embedding technique is widely regarded as the major breakthrough in deep learning. A vector can be translated into a relatively low-dimensional space known as an embedding. Machine learning is made easier using embeddings when dealing with large inputs, such as sparse vectors representing words. By placing semantically similar inputs close together in the embedding space, an embedding captures some of the semantics of the input. It is possible to learn and reuse embeddings between models. In the papers surveyed in this subsection, six out of seven work utilized pre-trained word embeddings, including Word2Vec [41], GloVE [40] and BERT [42]. Moreover, some cybersecurity entities use words in a flexible way. The word Gh0st, for example, refers to a remote access Trojan that contains both uppercase and lowercase letters. Further complicating identifications are irregular abbreviations and nesting issues within entities. To address the above challenge, character-based embeddings were applied and demonstrated in work [35] that improved entity extraction performance. The final representations of words are typically based on word-level and character-level representations, as well as additional information (e.g., linguistic features [36] and linguistic dependency [34], which are then fed into context encoding layers.

It is noted that most of the reviewed work focused exclusively on cyber-related entities and events extraction, rather than extracting relations between entities. In the process of event annotation, many challenges were encountered,

TABLE VI
REPRESENTATIVE WORKS ON MINING TACTICS, TECHNIQUES, AND PROCEDURES

Reference	Cyber scenarios	Techniques	Pre-trained word embeddings	Performance	Data	Dataset Publicly Available	Output Mapping
[48]	Comprehensive list of known tactics and techniques defined by MITRE'S ATT&CK [49] and CAPEC [50]	Threat-action ontology	BoW	Average 84% precision and 82% recall	Threats reports released by Symantec Security	N	STIX schema and other threat sharing standards
[47]	E-commerce (i.e., attacks in the e-commerce underground marketplace) TTPs	Entity extraction	Word2Vec	Identified 6,042 e-commerceTTPs with a precision of 80%	E-commerce threat corpora	Y	STIX schema
[51]	Comprehensive list of known tactics and techniques defined by MITRE'S ATT&CK [49] and CAPEC [50] and the broader CTI	Deep learning framework based on MIL ontology	FastText	Highest 64.41% precision with 82.18% precision and 52.96% recall using [48] dataset	Threats reports released by Symantec Security & other CTI dataset	N	Output technique tag
[52]	TTPs in security analysis report	Sentence level classification	TCENet	Average accuracy 94.1%	Security analysis reports with the sentence level TTPs annotation dataset	Y	STIX schemes

including annotating entities, events, and coreference relationships between events. Several distinct actions, for example, can be included in a description of a cyber attack. It is beneficial to incorporate global context across sentences or to consider non-local dependencies among phrases when performing information extraction tasks - such as name recognition, relationship extraction, event extraction, and coreference resolution [45]. Knowledge of a coreference relationship, for instance, can provide insight into the type of entity mentioned that is difficult to categorize. Furthermore, a sentence's entities can be used as inputs for event extraction, which can lead to useful information about event triggers. As a future direction, entities, events, and event coreference relationships will be combined to tap into joint CTI potentials by mining between entities in the same or adjacent sentences, while dynamic updates will model long-range cross-sentence relationships.

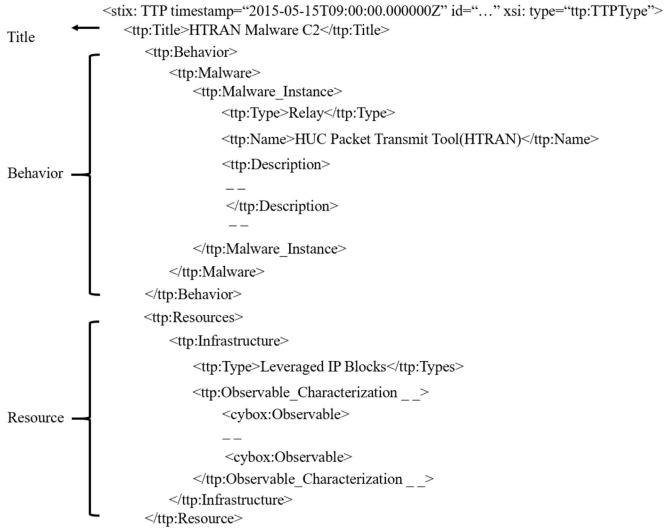


Fig. 3. Examples of TTP in STIX schema from [47].

B. Cyber Attack Tactics, Techniques and Procedures

The concept of Tactics, Techniques, and Procedures (TTPs) is crucial to CTI. The goal of identifying TTPs is to identify patterns of behavior that can be used to defend against specific threats and strategies employed by malicious actors. TTPs refer to the behaviors, including methods, tools, and strategies, that cyber threat actors and hackers utilize to prepare and execute cyber attacks. Based on the definition from the United States National Institute of Standards and Technology (NIST) [46], the tactic is the highest-level description of this behavior, techniques give a more detailed explanation in the context of a tactic, and procedures provide an even more detailed description in the context of a technique. This section reviews works on mining CTI about cyber attacks tactics, techniques, and procedures.

1) Summary of Representative Work: In Cyber Threat Intelligence, TTPs describe attack behavior associated with specific threat actors [53]. Cyber threats can be effectively identified, mitigated, and responded to when such information is collected. An example of TTPs in the Structured Threat Information eXpression (STIX) schema [54] is shown in Figure 3. The works target at mining TTPs, as summarized in Table VI, are limited but emerging due to the robustness of the roles of TTPs playing in identifying cyber threats.

The study by Husari et al. [48] described attack patterns and techniques of cyber threats using a threat-action ontology named TTPDrill. The ontology was constructed based on the MITRE's CAPEC [50] and ATT&CK [49] threat repository, which covered the procedure of pre- and post-exploit malicious actions. The threat actions and the corresponding kill-chain context in terms of tactics and techniques were

captured from micro-level (e.g., delete log file) to macro-level (e.g., defense evasion). Their work proposed an approach based on the established ontology that mapped the extracted TTPs from the unstructured data sources to the established ontology in a structured way, such as the STIX Attack Pattern schema [54] widely used in CTI. An NLP tool named Stanford typed dependency parser [55] was used to identify and extract the candidate threat actions. In addition, a set of regular expressions for common objects in the developed ontology were built to parse the special terms (e.g., strings fil_1.exe) that are used in threat reports confusing NLP tools. The candidate threat actions were applied to generate bag-of-words query and mapped to threat actions in ontology based on the calculation of similarity score.

You et al. [52] presented a novel threat context-enhanced TTP Intelligence Mining (TIM) framework for extracting TTP intelligence from unstructured threat data. The TIM framework utilizes TCENet (i.e., Threat Context Enhanced Network) to identify and categorize TTP descriptions, defined as three consecutive sentences, from textual data. You et al. [52] further enhanced the TTP classification accuracy of TCENet by utilizing the element features of TTP in the descriptions. The evaluation results demonstrate that the proposed method achieves an average classification accuracy of 94.1% across the six TTP categories. Furthermore, adding TTP element features improves classification accuracy compared to using only text features. TCENet outperforms previous document-level TTP classification works and other popular text classification methods, even in the case of few-shot training samples. The resulting TTP intelligence and rules aid defenders in deploying effective long-term threat detection and performing more realistic attack simulations to strengthen their defenses.

Ge and Wang's by proposing SeqMask as a solution for identifying and extracting TTPs for CTI using a Multi-Instance Learning (MIL) approach. SeqMask uses behavior keywords from CTI to predict TTPs labels using conditional probabilities. To ensure the validity of the extracted keywords, SeqMask employs two mechanisms, one involving expert experience verification, and the other blocking existing keywords to assess their impact on classification accuracy. The results of experiments conducted with SeqMask demonstrate a high F1 score (i.e., 86.07%) for TTPs classifications and an improved ability to extract TTPs from full-size CTI and malware.

Although the ontology based TTPs mining is able to cover a comprehensive list of tactics and techniques defined in MITRE's CAPEC [50] and ATT&CK [49] threat repository, it is difficult to adapt to diverse cyber scenarios, such as e-commerce tactics. As demonstrated in work [47], when applying TTPDrill to discover e-commerce TTPs, the recall, precision, and F1-score dropped to 50.25%, 22.38%, and 30.97% respectively. TTPDrill captured the TTPs in the traditional steps (i.e., in the phase of Cyber Kill Chain) of cyber attacks. As attacks occur before, during, and after the purchasing process, the e-commerce underground marketplace cannot be fully mapped to a conventional kill chain. To address this challenge, Wu et al. [47] built a TTP Semi-Automatic Generator (i.e., TAG) that incorporated NLP techniques, including topic term extraction and name entity recognition for identifying the e-commerce TTPs. According to the observation that topic terms in the TTPs usually share a similar semantic and lexical structure, the newly appearing topic terms were captured based on semantic and structure similarity with prevalent topic terms in [47]. In addition, the name entity recognition techniques as introduced in Section III-A combined with rule learning (i.e., a set of grammatical structure based rules for TTP entity recognition) were utilized for automatically extracting TTP entity from the unstructured data sources. After identifying TTP terms, the STIX TTP generator proposed by [47] converted the TTP terms extracted from unstructured data to the STIX schema [54]. A total of 6,042 TTPs were identified with 80% precision by TAG, which shed new light on previously unknown e-commerce CTI trends by analyzing the TTPs identified.

2) Discussion: In Table VI, the reviewed work is summarized, while the cyber attack tactics, techniques, and procedures are listed in Table VII. Since changing the attack tactic, techniques, and procedures is costly for the adversary, TTP is considered more robust and more lasting than IOC. For example, it is easy for the adversary to use IOC (e.g., different malicious domains) than to change his TTP (e.g., bulletproof hosting infrastructure) [47]. An IOC is one of the forensic artifacts that shows that a system has been infiltrated by an attack, while a TTP is one of the patterns or groups of activities associated with an individual or group of attackers. By having TTPs available, it is possible to investigate illicit activities using specific TTPs under cyber attacks in a variety of scenarios. During the recent boom in e-commerce, a number of attack patterns have emerged (such as order scalping), which have been extensively reported by public online sources. Detection, response, and containment of different types of security threats can be achieved through rapid threat analysis and deployment of TTPs to various security systems. To make TTPs tractable, a standardized and structured representation is required.

A cybersecurity corpus in contrast to an open domain corpus lacks annotation, which means more attention and effort needs to be put into it by the NLP community. Husari et al. [48] utilized the ontology based approach to sort out TTP related terms in line with the cyber kill chain. In work [47], NER was used along with human validation to guarantee the quality of critical outputs under the e-commerce TTPs domains. By using machine learning, TTP can be automatically generated from prior TTPs as the groundtruth, with the new context continuously enhancing the precision of TTPs. The TTPs extracted from [48] and [47] involve different languages, respectively English and Chinese. Dependency parsing and language processing depend heavily on language patterns. For example, a key prerequisite to language processing is the segmentation of words. In Asian languages (such as Chinese, Japanese, and Thai), words are not delimited by white space like in English. Nevertheless, TTPs can also be extracted from languages other than English. It is highly anticipated that TTPs will be extracted and converted across languages in this field.

Despite the decent performance of ML based approaches in discovering TTPs, these approaches face challenges in improving accuracy and explaining results due to their black-box

TABLE VII
CYBER ATTACKS TACTICS, TECHNIQUES, AND PROCEDURES IN REPRESENTATIVE WORKS

Category	Examples	Papers
Action	Send, encode, add, compress, extract, enumerate, query, modify, intercept, replace etc.	[48] [51]
Object	Junk data to traffic, data, credential hashes, Kerberos tickets, collected data, shortcut path etc.	[48] [51]
Intent	OS type discovery, hide communication, obtain credentials, evade defences, gather system information etc.	[48] [51]
Technique	Fingerprinting, data obfuscation, data compressed, binary padding, rootkit	[48] [51]
Kill Chain Phase	Recon, control, execute, maintain	[48] [51]
Victim targeting	E-commerce platform, e.g., Taobao	[47]
Exploit targets	Rules determining the rank of shops or products, e.g., through train	[47]
Resources	Infrastructure or tools helping make fake purchases, e.g., a website named Danghao	[47]
Attack patterns	Specific ways to make fake purchases, e.g., product cashing	[47]
Communications	Forums, mediators etc., e.g., Lotus Pond Moonlight Forum	[47]
Phishing	Dragonfly has used spearphising campaigns to gain access to victims	[52]
Scheduled task/job	Remsec schedules the execution one of its modules by creating a new scheduler task	[52]
Obfuscated files or information	Agent Tesla has had its code obfuscated in an apparent attempt to make analysis difficult	[52]
Deobfuscate/decode files or information	Carbon decrypts task and configuration files for execution	[52]
Collection	The jar file contains various classes for platform-specific implementations for capturing screenshots, capturing audio, logging keystrokes, among others	[52]
Application layer protocol	Carbon can use HTTPs in C2 communications	[52]

nature. The current extraction methods suffer from three primary limitations, namely insufficient data, incomplete verification, and a complex process. While identification methods determine classification accuracy, they do not provide reasoning behind their predictions. A simple yet comprehensive approach that combines data interpretation and high accuracy is required to obtain a complete picture of TTPs labels and evidence.

C. Profiles of Hackers

It is a never-ending game between cybersecurity attackers and defenders. By utilizing various resources, attackers are becoming more efficient and intelligent in carrying out their hacking activities. To better count hacking attempts, it is important to identify the source and resources of threat actors. This section reviews works on mining CTI for identifying the profiles of hackers, including cyber threats attribution and hacker assets.

1) Summary of Representative Work: Identifying the entity responsible for an attack is complicated and usually requires the assistance of an experienced security expert [61]. According to Hettema [62], attribution is one of the most intractable problems associated with an emerging field as a result of the technical architecture and geographies of the Internet. As the representative work shown in Table VIII, under different cyber scenarios (e.g., mobile malware, fintech security), the corresponding profiles of attackers are appropriately established with the attribution and assets.

Targeting for mobile malware threat actors as a starting point, Grisham et al. [60] used Long Short-Term Memory (LSTM) RNN architectures to identify the mobile malware attachments from CTI in online hacker forums. Furthermore, social network analysis was further utilized in this work [60] to recognize the key threat actors by understanding the threat actors' social groups and capabilities. By using networks and graph theory, social network analysis investigates social structures [63]. A networked structure is characterized by nodes (i.e., individual actors) and edges (i.e., relationships or interactions) between them. Particularly, in work [60], for a forum context, two-mode networks comprising two separate types of nodes (i.e., actor nodes affiliated with event nodes) were transferred to one-mode networks with actors linked to each other through posts in a shared thread. Hence, it is adaptable to calculate the potential centrality measures (e.g., closeness, betweenness) for a network of threat actors and further recognize the key threat actors in work [60]. It is possible, however, for the same malware to be reused by multiple actors. The actor who used malware to commit an attack might be different from the malware's author. Besides the utilized malware, a number of clues about the identity of the attacker can be gleaned from information collected during an incident. Perry et al. [58] proposed a method of identifying attack attribution named SMOBI (i.e., SMOthed BInary vector) based on CTI reports to recognize novel previously unseen threat actors and the similarities between known threat actors. The vector representation for cybersecurity related documents based on word embeddings (i.e., domain-specific word embeddings generated based on 20,630 cybersecurity articles and posts) was employed in work [58] to enhance the algorithms and reach full potential of the proposed attack attribution identification method.

For defending against data breaches, work [56] leveraged hacker source code, tutorials, and attachments directly from

TABLE VIII
REPRESENTATIVE WORKS ON MINING HACKERS' PROFILE

Work	Cyber Scenarios	Techniques	Features	Performance	Data	Dataset Publicly Available	Assests or Attribution
[56]	Data breach (e.g., Office of Personnel Management)	SVM and LDA	Term frequencies (i.e., one hundred code files for each language were used to train an SVM classifier	SVM classifier with 98.2% accuracy LDA with 0.94 Cronbach's alpha score	Underground forums	N	Assests (e.g., crypters, keyloggers, SQL Injections, and password crackers)
[56]	Data breach (e.g., Office of Personnel Management)	Social network analysis	Hackers and threads with source code assets for specific topics extracted by LDA	Evaluation on topological and node level Metrics	Underground forums	N	Attribution
[57]	Fintech cyber threats	ML, NN, and DL	High-level IOCs feature labels from ATT&CK MITRE	94% highest accuracy	CTI reports	N	Attribution
[58]	MITRE ATT&CK [49] and APT groups and operations [59]	Word embeddings, XGBoost	Domain-specific word embeddings	58.4% highest accuracy	CTI reports	Y	Attribution
[60]	Mobile malware	LSTM, RNN, social network analysis	Word embeddings, and hacker relationships	87% F1-score	Hacker forums	N	Attribution

underground hacker communities to identify malicious assets, such as crypters, keyloggers, SQL Injections, and password crackers to develop proactive CTI. In their work [56], classification models, such as Support Vector Machine (SVM), were implemented to classify the coding language. After that, LDA was used to analyze the forums' code, as well as comments, post contents, and attachments to identify malicious topics. As the last step, the metadata associated with the malicious topics was used to build social networks for identifying the attribution (i.e., key hackers) of the identified malicious topics.

The banking and financial sector is often the 'target of choice' for financially motivated Cyber Threat Actors (CTAs) [64]. Hence, it is necessary and urgent to ensure that Financial Technology (FinTech) is protected and secured against sophisticated cyber attacks from different CTAs, including state-sponsored or state-affiliated actors. Noor et al. [57] developed a machine learning based FinTech CTA framework. In their work [57], the cyber threat actors were profiled based on the high level attack patterns (e.g., Tactics, techniques and procedures taken from ATT&CK [49] MITRE [49]) extracted from CTI reports through Natural Language Processing. The accuracy of the classification model with DL achieved was 94%.

2) Discussion: It is challenging to establish a profile of hackers due to the fact that they always try to hide their identity and the assets they employed in the hacking. To profile the hackers, hybrid analyses were conducted on data sources from a variety of CTI, including code analysis, malware attachments analysis, documents (e.g., posts and comments in underground forums), and network analysis, as the representative work summarized in Table VIII.

In order to be effective, actionable CTI should incorporate not just traditional, internal approaches, but also external, open information [65]. This enables CTI to be more proactive by identifying threats before they occur, helping to understand attackers, and identifying hacker tactics. It is necessary to combine data with contextual information in order to provide relevant threats (i.e., internal incidents with external knowledge). Especially, online hacker forums are one rich-external data source that can be used to develop proactive CTI. Hackers use many venues for communicating and sharing information, including Internet-Relay-Chat (IRC), carding shops, DarkNet Marketplaces, and hacker forums [66]. Underground or hackers forums are among the ways hackers can freely share malicious tools (e.g., malicious attachments) [67], which provides practical resources for learning how threat actors operate and establishing hackers' profiles. Researchers have discovered that key hackers contribute significantly to their communities (e.g., forum moderators or senior members) [68]. Therefore, locating the key threat actors and identifying their groups through their interactions with other hackers is crucial.

D. Indicators of Compromise

Indicators of Compromise (IOCs) serve as forensic evidence of potential intrusions into a system or network. It is possible to detect intrusion attempts or other malicious activities using these artifacts by information security professionals and research community. Additionally, IOCs provide actionable threat intelligence that can be shared within the community to increase incident response and remediation efficiency. This section reviews works on mining CTI to extract IOCs and their relations.

1) Summary of Representative Work: Every year, cyber attacks are spreading widely and causing severe consequences, including data breaches, economic losses, hardware damage, etc. [76]. In view of the fast-spread speed of cyber attacks, it is imperative to proactively develop prevention methods based on recorded cyber attack event reports and log files. IOCs are pieces of forensic data identifying potentially malicious activity on an organization's system, such as system log entries or files. Examples of IOCs include attacker names, vulnerabilities,

TABLE IX
REPRESENTATIVE WORKS ON MINING INDICATORS OF COMPROMISE

work	Goal	Techniques	Pre-trained word embeddings	Data Sources	performance	Dataset Publicly Available
[69]	IOCs extraction and relationship model	NLP, dependency parser, logistic regression	N	45 technical blogs	98% precision 92% recall	Y
[70]	IOCs extraction	LSTM, REGEX	Word2Vec	687 cybersecurity articles	90.4% average precision 87.2% average recall	Y
[71]	IOCs extraction	BiLSTM, multi-head self-attention, REGEX	Token embeddings	687 English and 5427 Chinese cybersecurity articles	89.0% F1-score on English cybersecurity articles test set 81.8% F1-score on Chinese cybersecurity articles test set	Y
[72]	IOCs extraction and relationship model	Multi-granular attention, HIN, meta-path, BiLSTM + CRF	Proposed multi-granular embedding	73 international security sources	99.86% highest precision 99.81% highest recall	Y
[73]	IOCs extraction, categorization and relationship model	NLP, syntactic parser, semantic parser, REGEX	Dependency-based word embedding	10 article sources	91.9% precision 97.8% recall	N
[74]	IOCs extraction and relationship model represented in the form of actionable CTI	NLP, BiLSTM, REGEX	BERT	29,686 cybersecurity reports	86.99% accuracy 87.02% F1-score	Y
[75]	IOCs extraction	Knowledge Graph	A large general corpus	1,515 real world cybersecurity reports	88.7% F1-score for identifying IOCs	Y

IP/domain, hashes (MD5, SHA1, etc.), file names and addresses, and servers [69]. The use of IOCs aids information security and IT professionals in the detection of data breaches, malware infections, and other threats. In Table IX, we summarize the state-of-the-art work on obtaining CTI based on IOCs.

Work [69] proposed to automatically extract IOCs from unstructured texts. Liao et al. [69] proposed a method that firstly crawls blogs and removes unrelated articles. After splitting each article into multiple sentences, the method applies context terms and regular expressions to find those sentences likely have IOCs. This work [69] firstly proposed an approach that converts IOC candidates and relationships among them into a graph mining problem so that relationships can be detected according to the graph similarities. The precisions in finding IOC articles and extracting IOCs and relationships can reach up to 98% for both works.

The Bidirectional Long Short-Term Memory Neural Network (BiLSTM) and Conditional Random Fields (BiLSTM-CRF) aims to work on name entity recognition tasks, which have been shown to be applied in the field of IOC identification. Zhou et al. [70] are the first that applies the BiLSTM-CRF to IOC extraction from attack reports. The proposed approach [70] encoded the input sequence with attention-based and Word2Vec embedding. This work [70] functions well even when the number of training data is limited by using some token spelling features. The average precision in work [70] of automatically extracting and labeling IOCs is 90.4%. Based on the work of Zhou et al. [70], Long et al. [71] improved the model of Neural Network with the BiLSTM method using a multi-head self-attention module as well as more features and applied their approach to both English and Chinese datasets. The model [71] has more token features for improving the performance on a limited number of data, including spelling features, contextual features, and usage of features (i.e., the connection of spelling features and contextual features). The average precision scores of this model are 93.1% and 82.9% in the work of identifying IOCs from English and Chinese datasets, respectively. In addition, work [72] proposed a multi-granular attention Bi-LSTM-CRF model to extract IOCs with different granularities from multi-source threat texts and model the context of IOCs with a Heterogeneous Information Network (HIN). The study [72] manually defined meta-paths to present the relationships among several IOCs for better exploring contexts, which focuses on six common categories of IOCs, including the attacker, vulnerability, device, platform, malicious file, and attack type. In the work of IOC extraction, the highest precision is 99.86%, although extracting different items with different precision. The precision of threat entity recognition with the multi-granular model is 98.72% among all the experimented methods.

Given the multi-stage and varied techniques utilized in cyber attacks, knowledge graphs offer a distinct advantage in comprehensively depicting the entire attack process and identifying similarities with other attacks. For example, Li et al. [75] proposed AttacKG, a new method to aggregate threat intelligence from multiple CTI reports and create an attack graph that summarizes attack workflows at the technique level. They [75] introduced the concept of a Technique Knowledge Graph (TKG) to describe the complete attack chain in CTI reports by summarizing causal techniques from attack graphs. Li et al. [75] parsed CTI reports to extract attack-relevant entities and dependencies and used technique templates built on procedure examples from the MITRE ATT&CK [49]

TABLE X
SUMMARY ON THE KEY STEPS OF MINING INDICATORS OF COMPROMISE AND THEIR RELATIONSHIPS

	Date Pre-processing	IOC Candidates Identification	Relationship Extraction
[69]	Transfer images and embedded PDF files to text, topic filtering, break articles into sentences	Use REGEX and context terms to locate sentences that contain possible IOCs	Dependency graph, Stanford dependency parser
[70]	Apply pre-trained token embedding to all cybersecurity articles	Use token spelling features to train a IOC classifier	N
[71]	Apply pre-trained token embedding to all cybersecurity articles	Concatenate spelling features and contextual features	N
[72]	Extract threat-related description from HTML source code	Apply multi-granular attention based IOC extraction method	HIN construction, meta-path design, and similarity measuring
[73]	Identify multiple word expressions	Use the output of word embedding and named entity recognition to build IOC classifier	Dependency graph
[74]	Purification, segmentation, IOC Fanging	Apply REGEX identification, Filter and Candidate IOCs Replacement to locate candidate sentences	N
[75]	NLP based report paring and graph-level processing	Identify attack technique with templates	Attack graph generation and simplification

knowledge base. A revised graph alignment algorithm was then designed to match technique templates in attack graphs, align and refine entities, and construct TKGs. The technique templates aggregate new intelligence from real-world attack scenarios in CTI reports, and attack graphs utilize this knowledge to create TKGs that introduce the report with enhanced knowledge.

It is challenging to extract a whole attack process from the CTI data, despite the fact that it is the prerequisite to understanding hacking activities and developing defense strategies. Fortunately, an attack process can be projected by identifying IOCs and their relationships. Zhu and Dumitras [73] and Liu et al. [74] split the malware delivery campaign into different stages so that the attack process can be better analyzed. Zhu and Dumitras [73] adopted Natural Language ToolKit (NTLK) and Stanford CoreNLP to represent a sentence as a directed graph to describe the actions among IOCs. Word2Vec was applied to calculate semantic similarity, and Named Entity Recognition (NER) technique was used to locate IOC candidates. Four binary neural networks were designed to classify IOCs and determine whether a candidate is an IOC. Four stages (i.e., baiting, exploitation, installation, and command & control) from STIX [54] defined the process as a set of indicators and stages in work [73]. In summary, work [73] achieved the highest precision score of 91.9% in detecting IOCs and an average precision of 78.2% in classifying campaign stages. Similarly, Liu et al. [74] designed a trigger-enhanced system to generate CTI from unstructured texts, extract IOCs, and describe the connections between IOCs and campaigns. Particularly, after crawling reports and pre-processing, the system [74] utilized regular expression and a fine-tuning BERT model to identify the IOCs. This work [74] focused on six common types of IOCs (i.e., IP address, domain name, URL, hash, email address, and CVE). With the IOCs and related sentences, a trigger vector can highly explain the campaign stages. The highest precision that this system can reach is 86.55% in the work of classifying campaign stages.

2) Discussion: As summarized in Table X, all six studies in the surveyed research adopted the methodology consisting of data pre-processing (e.g., transferring images to text, breaking text into sentences, etc.), IOC candidate identification and relationship among IOCs extraction.

In the IOC candidates identification, all of the six studies used the REGular EXpression (i.e., REGEX) as a quick and effective method to search words or patterns with specific formats as token spelling features to select IOC candidates. Designing a good set of REGEXes aids in quickly identify IOC candidate terms and improve the performance of the model.

Across the six works, the methods on relationship extraction can be categorized into the following categories: 1. Transform an IOC sentence into a dependency graph, or tree and discover the relationships among IOCs [69], [73]. 2. Treat those words that can present the characteristics of the neighbor words as contextual keywords and generate contextual features from the keywords for the IOC candidates [70], [71]. 3. Create meta-paths to describe the relationship chains among multiple IOCs [72]. A dependency tree is a directed graph that can represent the relationships among all words in a sentence. However, the dependency tree may represent every word in a sentence, including non-useful words. The contextual feature captures the context surrounding each IOC, however, it needs to locate the keywords that are hard to distinguish from IOC terms in some scenarios. Meta-path approach can easily extract the relationships among IOCs, but the meta-paths need to be defined manually, and the number of them would increase exponentially with the increase of the number of IOC types [77]. It is expected that these methods will be assembled into an efficient approach that can be generalized to a variety of types of IOCs relationship extraction.

It is worth mentioning that most of the reviewed studies mainly focused on IOC identification and a few on relationship extraction. A possible direction for future research is to predict cyber attacks that may damage our hardware or software based on the extracted IOCs and their relationships.

Extracting the detailed information and features of the attack, including but not limited to the attack type, exploiting vulnerabilities, and the target victim, is achievable to generate an attack report for cyber security experts to predict cyber attacks as well as develop a defense strategy. For example, building a series of knowledge graphs periodically with IOCs and relationships, then learning the evolutionary graphs by digging into the changes between graphs and predicting the next possible event is a feasible solution.

E. Vulnerability Exploits and Malware Implementation

It is becoming increasingly common and dangerous to be exposed to cybersecurity risks and malware threats. There are a wide range of vulnerabilities that can lead to data leaks, and threat agents can exploit them to compromise secure networks. Despite much attention paid to vulnerability and malware detection using code semantics, mining CTI sources beyond code is limited in terms of discovering practical information about vulnerability exploits and malware implementation. In this section, we comprehensively review representative works that successfully identified vulnerabilities that might be exploited and malware implementation through CTI mining.

1) Summary of Representative Work: Recently, there has been an increase in the number of software vulnerabilities exploited. Vulnerabilities are weaknesses that can be exploited by cybercriminals to gain unauthorized access to computer systems. The exploit of a vulnerability can lead to malicious code being run, malware being installed, and sensitive data being stolen by a cyberattack. It is therefore necessary to prioritize the response to new disclosures by assessing which vulnerabilities are likely to be exploited and ruling out those that are not. Furthermore, malware detection increasingly relies on machine learning techniques that focus on code semantics in order to distinguish malware from benign software. For example, human intuition and knowledge are key to the effectiveness of these techniques. In light of adversaries' efforts to evade detection, as well as the increasing amount of resources available on malware behavior online, feature engineering likely draws on a small fraction of these sources. It is therefore expected that multiple data sources will be consulted in order to obtain knowledge about vulnerability exploits and malware implementation beyond the code itself.

In work [78], Sabottke et al. studied vulnerability-related information in the wild for early exploit detection prior to the public disclosure of vulnerabilities. The study mined a large number of disseminated on Twitter that contained cybersecurity vulnerability information and constructed a machine learning model to detect which vulnerability was more likely to be exploited in the real world. In addition to mining Tweet text for word features and Twitter traffic for statistics features, information from National Vulnerability Database (NVD) [22] and Open Sourced Vulnerability Database (OSVDB) [85] are also collected and used for exploit detectors. As far as we know, this work [78] is the first technique ever used for early detection of real-world exploits using social media. Furthermore, Nunes et al. [86] developed an operational system to collect and identify vulnerability exploits and malware development information from the darknet and deepnet discussions, particularly from hacker forums and marketplaces. After extracting and structuring the information from Web pages in real-time, they [86] combined supervised and semi-supervised approaches to discover products and topics related to malicious hacking. This provided threat warnings about newly developed malware and vulnerability exploits that have not yet been deployed in a cyber attack. With limited labelled data available on the darknet and deepnet, the proposed approach reached a precision of 80% by requiring less expert knowledge and costs.

In order to detect malware, researchers propose a growing number of features derived from human knowledge and intuition that are used to characterize malware behavior. Due to adversaries' efforts to evade detection and increasing publications on malware behavior, the feature engineering process probably draws on a fraction of the available data. In order to gain greater benefit from a considerable amount of CTI regarding malware behavior, FeatureSmith [79] proposed by Zhu and Dumitraş adopted scientific papers as the source of information to discover and collect malware detection features automatically. Through the pipeline of data collection, behavior extraction from literature, behavior filtering and weighting, semantic network construction, feature generation, and explanation generation, FeatureSmith identified abstract behaviors associated with malware and then presented them as concrete features for malware detection. As a proof of concept, FeatureSmith's automatically engineered features showed no performance loss in detecting real-world Android malware, with 92.5% true positives and 1% false positives compared to a state-of-the-art feature set produced manually.

Recent literature has explored how NLP can significantly improve humans' understanding of the cybersecurity context. In the area of vulnerability exploits and malware implementation, work [80] introduced a method to annotate malware reports, which provides semantic-level information on the text and helps researchers quickly understand the capabilities of specific malware. Lim et al. annotated Advanced Persistent Threat (APT) reports with attribute labels from the Malware Attribute Enumeration and Characterization (MAEC) vocabulary as the groundtruth for the NLP tasks. They began by classifying whether a sentence is malware related or not and then predicting the tokens, relations between tokens, attribute labels, and malware signatures based on the text that describes the malware. In addition, the work of [81] leveraged diverse resources, including unlabeled text, human annotations, and specifications (i.e., MAEC vocabulary) about malware attributes to conduct malware attribution identification. WAE (Word Annotation Embedding) was applied to encode information from heterogeneous information. The results tested on SemEval SecureNLP classification task [87] showed that the model trained on features generated from the proposed annotation approach outperformed the annotation approach presented by [80], as well as the embeddings features learned by [88].

In recent studies, it has been shown that software documentation can be used to predict software vulnerabilities

TABLE XI
REPRESENTATIVE WORKS ON MINING VULNERABILITY EXPLOITS AND MALWARE IMPLEMENTATION

Work	Target	Techniques	Features	Performance	Data
[78]	Vulnerability exploits prediction	Mutual information based filter, SVM	Twitter Text, Twitter Statistics, CVSS Information and database information	87.5% precision for public PoC exploits, higher than 80% precision for private PoC exploits	Twitter, NVD, OSVDB, Exploit DB, Microsoft Security Adversaries, Symantec
[79]	Automatic feature engineering associated with malware; build malware detector using the extracted features	Semantic network construction has three types of nodes, respectively known malware families, abstract malicious behaviors, and concrete features that can be extracted from Android apps through static analysis, where links among these nodes reflect the semantic relationships	Features mining from security literature (i.e., permissions, intents, and API calls)	92.5% TPR with 1% FPR for the malware detector	Security literature (e.g., security conference papers)
[80]	Tokens annotation relevant to malware capabilities or action implied	Tokens annotation, classification	Bag-of-words, unigrams, bigrams, part-of-speech.	40.3% average F1-score for predicting token labels, 89.3% average F1-score for predicting relation labels	APT reports
[81]	Malware attribution labels prediction from cybersecurity text	Word annotation embedding algorithm, classification	Word embeddings and malware labels embeddings learned from word annotation embeddings	48.8% average F1-score for predicting attribution labels	APT reports, MAEC specification, human annotation
[82]	Predict logic vulnerabilities in payment syndication services	Dependency parsing and word embedding for recognizing entities that construct the security requirements to conduct logic flaw detection	Taking results of dependency parsing as features for entity recognition	100% accuracy for predicting potential logic flaws from documentations	Syndication documents
[83]	Discover long-term evolution vulnerabilities	Textual entailment, dependency parsing	Hazard indicators generated from LTE documentation using NLP	Reported 42 vulnerabilities from LTE NAS specification	LTE NAS specification
[84]	Construct cybersecurity knowledge graph on malware	Entities extraction, classification, entity relationship prediction	Word embeddings and relationships labels from multiple datasets	48% of entities captured correct knowledge represented in the report	AARs

without relying on the program code at all. Chen et al. [82] developed a tool that enables automatic inspection of system security specification documents instead of relying on program code analysis (e.g., model checking) to predict logic vulnerabilities in payment syndication services. They explored the use of NLP to discover logical vulnerabilities from the syndication developer's guide according to the payment models and payment service's security requirements. They extended the Finite State Machine (FSM) that was usually manually extracted for evaluating payment services by using the dependency parse tree of sentences in the developer guide to extract the parties involved in the process and the contents transmitted between them. Software documentation-specific NLP techniques were fine-tuned for the proposed approach. Furthermore, Chen et al. [83] continually applied the NLP techniques, including textual entailment and dependency parsing, to analyze Long-Term Evolution (LTE) documentation of cellar networks for Hazard Indicators (HIs). A total of 42 vulnerabilities were found in the LTE Non-Access Stratum documentation and reported to authorized parties through the proposed approach by Chen et al. [83], proving the effectiveness of this method of finding vulnerabilities.

In addition, the Knowledge Graph (KG) helps transform free-text cybersecurity into more structured formats with semantic-rich knowledge representations insights. As an example of constructing a KG from data about malware, Piplai et al. [84] proposed a cybersecurity KG from malware After Action Reports (AARs), which encloses insightful analyses of cybersecurity incidents and hereby delivers reliable information to security analysts. AARs can help deal with unidentified cybersecurity incidents by matching patterns with the predefined incidents since they provide crucial data about detection and mitigation techniques. Specifically, in work [84], the malware entity extractor based on Stanford NER [89] was created for the construction of the cybersecurity KG, and it

was trained based on data from CVEs and security blogs to identify entities required for the cybersecurity KG.

2) Discussion: In the face of enormous source code and the advancement of technology, automated vulnerability analysis and detection have emerged as a current research hotspot. Research on vulnerabilities and malware detection is anticipated to expand beyond analyzing source code to mining CTI from multiple data sources. It will significantly enhance the ability to identify, prioritize, and fix vulnerabilities if insights knowledge can be mined on vulnerabilities exploits and malware implementation.

An early identification of vulnerabilities can prevent disastrous consequences associated with their exploit. The information on vulnerabilities and malware is available in a variety of sources, including open source and classified data. There are several repositories of structured and semi-structured information on vulnerabilities and malware, including the NVD [22], IBM's XFORCE [90], US-CERT's Vulnerability Notes Database [91], and others. Informal sources, such as computer forums, hacker blogs, social media, etc, also contribute to these knowledge bases. While such unstructured sources are noisy, redundant, and often contain misinformation, they can be mined and aggregated to track the spread of new malware and vulnerabilities and alert security experts to take action. Technology in ML and NLP has enabled powerful automatic feature extraction techniques to mine features from documentation, making them more viable and timely strategies to identify relevant semantic information and understand vulnerabilities in multiple data sources, thus replacing manual detection.

F. Threat Hunting

Threat hunting is the practice of proactively searching for cyber threats that are lurking undetected in a network. Based on the definition from IBM, threat hunting is a proactive approach to identifying previously unknown, or ongoing non-remediated threats, within an organization's network [59]. During threat hunting, the suspicious activity patterns that may deemed to be resolved but isn't or have been missed are inspected. This section reviews works on mining CTI to conduct threat hunting.

1) Summary of Representative Work: The importance of threat hunting lies in the fact that sophisticated threats can get past automated cybersecurity systems [100]. A well-prepared attacker will be able to penetrate any network and avoid detection for up to 280 days on average [59]. Attackers can do less damage by reducing the time between intrusion and discovery by utilizing effective threat hunting. Knowledge about cybersecurity threats (e.g., malware employed in APT campaigns) is covered in a variety of CTI resources and presented in various formats, including natural language, structured, semi-structured, and unstructured forms. Due to the fact that the hackers usually meet online to discuss the latest hacking techniques or tools [101], work [92] applied text mining to identify the terms related to emerging cyber threats from the online chatters, such as Twitter and dark Web forums. Furthermore, [93] proposed a diachronic graph embedding framework that helps in dynamically capturing the evolution of hacker terms over time.

There are, however, fragmented views of cyber threats that can be extracted by approaches focusing on extracting terms related to emerging threats, such as signatures (e.g., hashes of artifacts), file names, IP addresses and timestamps. Using predefined rules, such as correlating suspicious threats using heuristics, we could discover emerging threats. It is hard and lacks the precision to show the complete picture of how the threat evolved, especially over long periods. Hence, recent research efforts are dedicated to correlating the relationships between threat terms (i.e., IOC artifacts) and representing the attackers' steps in the form of graphs, which includes clues on the behavior of the attacks. In this case, even if the hackers update their strategies (e.g., signatures) to conduct attacks, threat hunting is still effective compared to concentrating on the threat terms only. Satvat et al. [94] extracted the full picture of the attack behavior from the CTI reports and represented it as a group to identify the APT. Through the proposed approach by work [94], the complicated descriptions from the CTI report are processed to be as a provenance graph, where nodes signify the entities (e.g., domain names, username and file), and the edges point to system calls (e.g., write, send, decode and log). Furthermore, Milajerdi et al. [96] bridged the gap between the low level system-call view and the high level APT kill chain view by building an intermediate layer between them. The intermediate layer is established based on MITRE's ATT&CK [49] threat repository that describes hundreds of behavioral patterns defined as TTPs, which summarizes the observations from the nodes and edges in the provenance graph.

It's expected that threat intelligence will gather information from multiple sources to provide more insights. Gao et al. [95] proposed an approach that described the CTI instances involving different types of threat infrastructure nodes (i.e., domain name, IP address, malware hash, and email address) and edges (i.e., relation matrices between nodes). By utilizing the open source CTI, such as Common Vulnerabilities and Exposures (CVE) [102] to discover the relationships of exploiting the same vulnerability, it can be possible to discover more information between two malware hashes. Using heterogeneous graph convolutional networks, a threat infrastructure similarity measure-based approach for modeling and identifying threats (e.g., malicious code, Botnet, and unauthorized access) involved in CTI has been proposed [95]. Meta-path and meta-graph were defined in work [95] to capture the high level relationships over nodes from various semantic meanings. Another example of combining CTI from multiple sources is that Milajerdi et al. [97] adopted a novel similarity metric to assess the alignment between attack behavior graph extracted from IOC open standards and system behavior graph from kernel audit logs. Furthermore, THREATRAPTOR, a system created by Gao et al. [99], enables the process of threat hunting with the use of Open Source Cyber Threat Intelligence (OSCTI). The system accomplishes this by developing an unsupervised NLP pipeline that extracts organized actions from unstructured open source CTI. These organized actions can be effortlessly searched using the proposed domain

TABLE XII
REPRESENTATIVE WORKS ON THREAT HUNTING

Work	Techniques	Performance	Data	Dataset Publicly Available	Threat hunting	Warning
[92]	Text mining for threat hunting; Rules setting for warning generation.	Identify terms related to emerging cyber threats with precision above 80%	Online chatter (e.g., Twitter, Blogs, dark web forums)	N	Y (e.g., data breach, malware, ransomware, botnet, exploit, IoT, DDoS)	Y
[93]	Diachronic Graph Embedding Framework that maps the evolution of hacker terms over time by operating on a Graph-of-Words representation of text	Multiple intrinsic and extrinsic experiments, case studies	Online hacker forums	Y	Y (e.g., web application threats that target PHP technologies, DoS threats that target the Windows operating system)	N
[94]	Concisely extract the full picture of the attack behavior from CTI reports and represent as a graph	Evaluate using real-world incident reports as well as reports of DARPA adversarial engagements	CTI reports	Y	Y (e.g., APT)	N
[95]	Model CTI on heterogeneous information network integrating various types of infrastructure nodes and relations among the nodes	Evaluate on comprehensive experiments on real-world datasets and demonstrate the results for threat type identification is state-of-the-art (the highest Macro-F1 can reach 80%)	CTI reports and cybersecurity databases	N	Y (e.g., spam URLs, brute force login attack)	Y
[96]	Generate high level graph to summarize attackers' steps in real-time	Evaluate on real-world APTs and demonstrate the APT can be detected with high precision and low alarm rate	Audit logs	Y	Y (e.g., APT)	Y
[97]	Calculate the similarity to assess an alignment between query graph (i.e., based on CTI correlations) and a provenance graph (i.e., based on kernel audit log records)	Evaluate on real-world cyber attacks and attack scenarios and the attacks can be detected with high confidence, no false signals, and in a matter of minutes.	CTI reports and IOC descriptions (i.e., Kernel audit records)	Y	Y (e.g., the likelihood of success of an attack campaign)	Y
[98]	The temporal and probabilistic dependence between alerts are explored based on the suffix-based probabilistic deterministic finite automation model	Evaluate on comprehensive experiments and use cases to show the utility and how the proposed approach present a clear picture of the attack as well as the strategies of the involved teams	Intrusion alters	Y	Y (e.g., severity of attackers' actions)	Y
[99]	Extract threat behavior using the unsupervised NLP pipeline and hunt malicious activities by designing a query language and building query synthesis mechanism and query execution engine	Evaluate on a broad set of attack cases that show the accuracy and efficiency of the developed threat hunting system	Audit logs	Y	Y (e.g., severity of attackers' actions)	N

specific query language, query synthesis mechanism, and query execution engine.

2) Discussion: Keeping up with cyber threats and responding to potential attacks rapidly is becoming increasingly important as enterprises strive to stay ahead of the latest threats [103]. An effective threat hunting strategy is one that proactively searches for cyber threats lurking in a network that go undetected. Threat hunting digs deep into the target environment to find malicious actors that have slipped past its endpoint security measures. Upon sneaking into a network, an attacker can gain access to data, confidential information, or login credentials that will allow later movement. Organizations often lack the advanced detection capabilities to detect advanced persistent threats once adversaries evade detection and penetrate their defenses. Hence, threat hunting is an essential part of any defense strategy. Hence, threat hunting is an essential part of any defense strategy.

There are several challenges involved in threat hunting inside an enterprise: (1) Attackers often perform their attack steps over long periods of time, for example, lurking over several months before discovery [59]. In this manner, a significant data breach can be launched by siphoning off data and exposing enough confidential information to enable further access. A method of linking related IOCs together is therefore necessary due to the attack activities occurring over a long period of time [104]. (2) Effective threat hunting must be able to identify whether an attack campaign will affect system, even if the attacker has modified artifacts like file hashes and IP addresses to avoid detection. Hence, a robust approach should uncover the entire threat scenario, instead of looking for matching IOCs

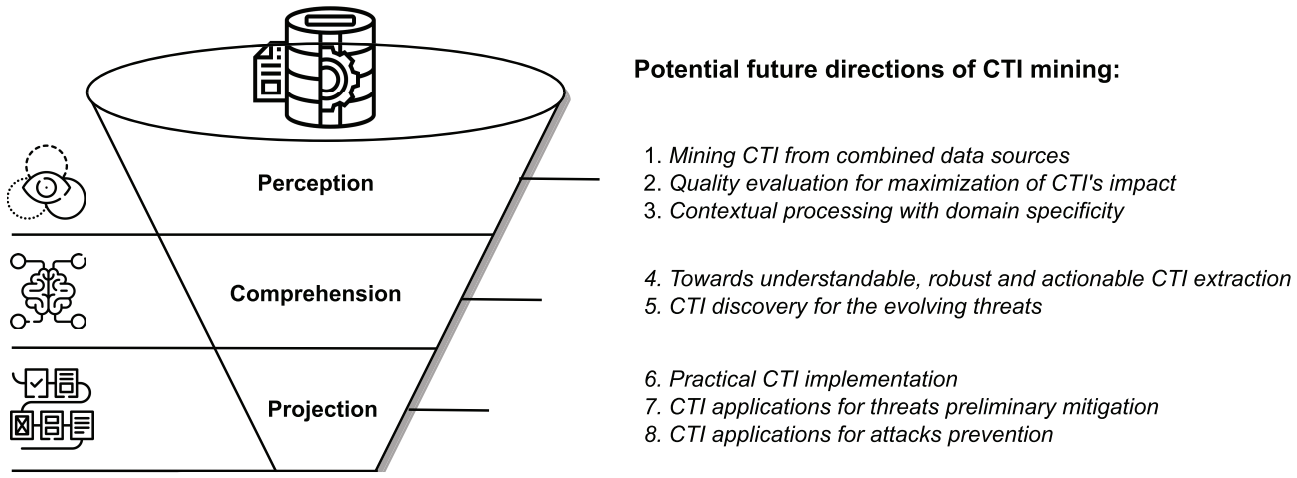


Fig. 4. Future Directions of Cyber Threat Intelligence Mining for Proactive Security Defense.

in isolation [24]. (3) In order for a cyber analyst to analyze and respond to a threat incident in a timely manner, the approach must be efficient and not produce many false positives so those appropriate cyber-response operations can be initiated [97].

To overcome the above mentioned limitations and build a robust detection system for threat hunting, it is important to consider the correlation between indicators of compromise. CTI reports present information about cybersecurity threats in a variety of forms, such as natural language, structured, and semi-structured. The security community has adopted open standards such as STIX [54] and OpenIOC [19], in order to facilitate the exchange of CTI in the form of IOCs and enable the characterization of TTPs. A standard's description of indicators or observables often illustrates how they are related to each other to provide a better perception of attacks [7]. The relationships between IOC artifacts provide essential clues about attacks inside a compromised system, which are tied to attacker goals, and are therefore difficult to change [97].

IV. Challenges and Future Directions

Despite numerous investigations advocating the use of CTI mining to achieve proactive cybersecurity defense, as discussed in Section III, there remain a multitude of challenges that must be addressed. This section will delve into the difficulties encountered in this field. To combat these challenges, potential future directions will be outlined in accordance with the perception, comprehension, and projection process pipeline, which was introduced in Section II and is depicted in Figure 4.

A. Perception

1) Future Direction 1 (Mining CTI From Combined Data Sources): We have seen a paradigm shift in understanding and defending against evolving cyber threats, from primarily reactive detection to proactive prediction, driven by the increasing scale and high profile cybersecurity incidents related to public data in recent years [24]. The amount of information about cybersecurity is rapidly increasing from multiple sources, including open source cyber threat intelligence and restricted-access classified information.

While the vast amount of information sources makes it possible to mine more valuable CTI than ever, it is common for threat reports to contain a significant amount of irrelevant text [105]. In other words, only a small portion of the report is dedicated to the description of attack behavior. For instance, describing the geographical origin of the attacker is of interest. However, it does not contribute to clarifying the attack behavior in an attacking activity if that information is not provided. In addition, in previous research, most work only used one source of data, even though different studies employed different sources. For instance, Table III summarizes recent work on mining cybersecurity-related entities and events, where only data from a single source was used in most works.

It is envisioned that CTI will be extracted from multiple data sources by aggregating information from these different resources in the future. Furthermore, it is expected that the relationships between these data sources will be investigated in order to provide a holistic picture of the attack activity by using multi level information about CTI, such as with the aid of heterogeneous knowledge graph. In addition, it is important to check for issues related to quality, such as false alarms and consistency, when it comes to extracted CTI.

2) Future Direction (Quality Evaluation for Maximization of CTI's Impact): CTI can be obtained from a variety of sources, including but not limited to government agencies, security vendors, research organizations, and open-source information. The challenge lies in identifying credible and reliable sources of CTI, as the quality of the information can vary greatly. In addition, the dynamic nature of CTI means that the information is constantly changing and evolving, making it crucial to carefully evaluate the quality of the information and its sources when trying to understand and predict potential cyber threats. Collecting high-quality CTI is a challenge that requires a thorough understanding of the sources and a systematic approach to evaluating the credibility and reliability of the information, which ultimately decides the impact of CTI.

There have been a few studies on accessing the quality of CTI and its sources in recent years [106], [107], [108]. For example, Schaberreiter et al. [106] and Griffioen et al. [107] proposed the quantitative assessment of parameters to evaluate

the quality of CTI, such as extensiveness, maintenance, compliance, timeliness, completeness, etc. Schlette et al. [108] proposed a series of quality dimensions and showcased how to make quality assessment transparent. The field of cybersecurity is constantly evolving, and the exploration of CTI and its quality is an ongoing pursuit. As more is understood about the dynamics of CTI and the factors that influence its quality, organizations can better assess the CTI they receive and make more informed decisions about their security posture. The continued development of methodologies and frameworks for evaluating the quality of CTI will help to ensure that organizations can effectively use CTI to improve their security posture.

Furthermore, it is crucial to consider the impact of CTI on evaluating its quality and the quality of its sources. The assessment of CTI's quality should be based on solid evidence instead of subjective opinions. For example, in a study by Liao et al. [69], the authors utilized IOCs to track emerging cyber threats and determined high-quality intelligence sources by evaluating the comprehensiveness, timeliness, and dependability of their IOCs. This integrated approach of considering both the quality of the information and its impact provides a more comprehensive evaluation of CTI. Developing a systematic and evidence-based method for assessing the quality of CTI and its sources is essential for ensuring that the information is accurate and reliable and can be effectively used to protect against cyber attacks.

3) Future Direction 3 (Contextual Processing With Domain Specificity): Furthermore, among the assumptions made by the reviewed studies is that the text structure of the CTI reports follows a relatively simple structure [109]. For example, grammatically follows a specific pattern, assuming the cybersecurity related terms can be captured by regular expression, taking into account stable grammatical relations in the form of subject, verb, and object in the sentence. The fact is that CTI reports, in general, contain a great deal more complex domain-specific context than most other reports [110]. As a result of the complex syntactic and semantic structure of CTI reports, the prevalence of technical terms, as well as a lack of proper punctuation in these reports, these factors can easily influence how the report is interpreted and how the attack behaviors are extracted.

A few research efforts worked on creating cybersecurity domain groundtruth datasets. Satyapanich et al. [36] created and published a corpus containing 1000 annotations for five types of cybersecurity attacks, thus providing a foundation for simplifying the process of extracting cybersecurity related information from the raw data and facilitating the development of domain-specific groundtruth. Behzadan et al. [111] manually labeled 21,000 cybersecurity related tweets for future usage. In addition, in contrast to general pre-trained models (e.g., word2vec [88], glove [40]), cybersecurity specific NER models and word embeddings (e.g., sec2vec [112] modified by EmTaggeR [113]) are shown to improve performance in processing complex domain-specific contexts [36], [114].

B. Comprehension

1) Future Direction 4 (Towards Understandable, Robust and Actionable CTI Extraction): In recent years, researchers have made significant contributions to the automation of the extraction of CTIs from multiple data sources [12]. However, there are still some challenges to overcome: (1) Due to the severe shortage of experienced professionals, many organisations cannot handle the flood of CTI feeds, causing them to be burdened. (2) As a result of fake CTI generated by adversaries, false alarms might occur. In addition, adversaries can make use of fake CTI to corrupt cyber defence systems. (3) The extracted CTI can be difficult to utilise for actionable advice, for example, prioritizing the following actions for cybersecurity defence. It is essential that the next generation of CTI is understandable, robust, and actionable in order to overcome these challenges. Firstly, understandable CTI facilitates people without strong cybersecurity domain knowledge with the interpretation of key security elements. For example, in work [115], 15 categories of entities related to cybersecurity events were extracted and indexed from text through supervised approaches based on neural networks. Cybersecurity related information, such as the impacted date, time and organisation of a security event, is extracted and used to explain a specific cybersecurity event. With the interpretation of the annotated entities, the CTI becomes more accessible and understandable for further analysis. The explainability of CTI can be improved by including more entities and variety that will facilitate the explanation of CTI by expanding entities through enlarging the groundtruth data and embedding supplementary semantic features to concatenate with word embedding. In addition, because cybersecurity events are language independent, the study on turning unstructured text from sources across different languages into a structured format is expected.

Secondly, robust CTI ensures the extracted data is genuine instead of fake by adversaries. Fake CTI examples are used as input to corrupt cyber defence systems, which serve for attackers to achieve malicious needs through training models on incorrect inputs [116]. Recent work [116] demonstrated that the majority of fake CTI samples generated by GPT-2 transformers are labelled as true even by cybersecurity professionals and threat hunters. Linguistic errors and disfluencies that generative transformers commonly produce but humans rarely are expected to be explored and utilised as the key features to distill genuine CTI. To detect fake CTI samples, aspects such as aesthetic, readability, source credibility, novelty, and propagation identified through the analysis of users' propagation and perceptions of real and fake cyber news [117] are worth investigating.

Last but not least, actionable CTI delivers complete and accurate information that is relevant and trustworthy to the consuming organisation. The CTI can be called actionable if the CTI is relevant and trustworthy to the operations of organisations, provide complete and accurate information, and can be ingested into CTI sharing platforms [12]. The output of CTI mining aims to provide actionable suggestions, including risk mitigation, security practice recommendation, and relationship establishment between the extracted CTI. For example, users are expected to be provided with actionable CTI outputs with the help of publicly available security datasets, recommendations, and knowledge graphs that represent the relationships among various CTI.

2) Future Direction 5 (CTI Discovery for the Evolving Threats): Cyber defence tools are constantly updating and becoming more and more sophisticated [118]. Yet, we are still facing a slow response to the ever-evolving of cyber threats, such as phishing to steal our information, ransomware to encrypt our data and demand a ransom in exchange, and malware to compromise our critical infrastructures. Ensuring the timely and automated intelligence discovery of evolving threats from publicly available sources, such as hacker forums and threat reports, is paramount in helping organizations keep pace with ever-changing threat landscapes. However, existing threat intelligence extraction techniques ignore the ever-evolving nature of cyber threats. Recent development in AI compounds the problem by taking advantage of adversaries that can adapt to attacks, generate variants, and evade detection: "This new era of offensive AI leverages various forms of machine learning to supercharge cyberattacks, resulting in unpredictable, contextualised, speedier, and stealthier assaults that can cripple unprotected organizations", Forrester Consulting [119].

Current approaches to extracting open source CTI, use various NLP and machine learning ML techniques, for example, text memorization, information extraction, named entity recognition, decision tree and neural networks, to understand the means and the consequence of different cyber attacks. However, current CTI work has three major limitations: (1) static and isolated CTI hardly depicts the dynamics of threat attacks and the vast landscape of threat events; (2) fragmented views of CTI, such as suspicious domain names and hashes of artifacts, can hardly help security analysts to hunt down the target of an advanced persistent threat in an enterprise; (3) the inter-dependency among CTI, which can help us to reveal a big picture of how the threat behaviors, are unexplored. Furthermore, AI-powered adaptive cyber attacks bring more challenges in those different variants of the attack can develop and multiple cyber attacks can even cooperate to cause large-scale organized crime. In general, CTI extraction is a significant and challenging task for enterprises and individuals and current work cannot address this growing issue of national intelligence and security. Hence, to develop focused theory and techniques for the automatic extraction of interconnected and evolving CTI from heterogeneous open sources, constructing a dynamic CTI knowledge graph to uncover how cyber attacks evolve and how multiple cyber attacks coordinate in infiltrating a system is expected to realise timely and responsive cyber threat hunting in a complex system.

C. Projection

1) Future Direction 6 (Practical CTI Implementation): CTI mining studies have the challenge of transforming the research studies into practical implementations and applications of CTI and demonstrating their practical significance to the maximum extent possible. Many CTI tools are available on the market that facilitate the collection, analysis, and sharing of CTI data. In our review of the existing CTI tools, we summarized them into four categories: (1) Open source and enterprise tools that can access threat intelligence and offer advanced management options (e,g., functions including filtering, analysis, finding correlations, search). (2) The CTI protocol set is a set of languages for describing and sharing CTI information. (3) The sharing platforms for CTI. (4) Incident response systems given the collected CTI.

Though many organizations wish to share their CTIs, a universally accepted format for CTI exchange is expected. For example, in order to facilitate CTI exchange, MITRE developed the STIX scheme [54] that is widely adopted by research studies and CTI applications. It is important that data formats are compatible with the different systems of stakeholders. In order to exchange CTI in a timely manner, unnecessary data transformations must be avoided.

It is the core idea behind CTI sharing that by sharing information about the most recent threats and vulnerabilities among stakeholders, as well as implementing the remedies as quickly as possible, stakeholders will become aware of the situation [8]. CTI sharing offers a new way to create situation awareness among sharing stakeholders. In addition, it is seen as a necessity to prepare for future attacks in order to preempt them rather than react to them as in the current practice. CTI sharing is expected to become an integral part of proactive cybersecurity for organizations in the future to share their information. Implementing the way of CTI sharing in a way that consumes and disseminates information in a timely manner will be of great benefit to the industry, whose future depends on how well the CTI is comprehended and implemented its remedies.

2) Future Direction 7 (CTI Applications for Threats Preliminary Mitigation): By taking a more proactive, forward-thinking approach from the start, companies can address and mitigate future disruptions and cyber threats [120]. Working actively to prevent threats promotes complete control over the cybersecurity strategy. This helps to prioritize risks and address them accordingly. By identifying vulnerabilities early on, and preparing for the worst-case scenarios ahead of time, we will be able to take action rapidly and decisively during a cyber incident. While proactive measures help to prevent breaches, reactive measures strike if and when a breach occurs. The proactive security market was worth USD 20.81 million in 2020, and it is expected to grow to USD 45.67 million by 2026 [121].

Threat mitigation is the process of reducing the severity of threats from physical, software, hardware, etc., of IT systems. From the perspective of CTI mining applications, we illustrate how threats can be mitigated in a proactive manner. First, the acquired CTI can assist in organisational strategies that refer to physical security measures, training, and education. Secondly, in terms of networking strategies that use technical implementations for threats mitigation, monitoring network activities from the CTI and anticipating cyber attacks are potential future directions. For example, by using security events data from commercial intrusion prevention systems, Shen et al. [122] predict the specific steps that will be taken by the adversary to perform cyberattacks. The demand for special security solutions that are customized to the organization is also on the rise. It is expected that organizations have access to specialized security expertise that can easily analyze a system

and transform its security from zero to a significant level within a short timeframe. For example, an innovative method for integrating heterogeneous data into customized and understandable cybersecurity information was proposed in recent research work [123], which can be applied for cybersecurity consultation and specialized security solutions.

3) Future Direction 8 (CTI Applications for Attacks Prevention): Recently, the number of cyber threats is constantly increasing. There are ten times more malwares now than ten years ago. More and more security organizations start collecting threat details and applying measures to prevent them. Thus, threat prediction is essential to detect and prevent potential attacks and loss.

By collecting massive CTI reports and forums from external sources and extracting useful information, including attack name, characteristics, vulnerabilities the attack may explore, objects, etc., it is possible to predict whether a threat may attack specific devices [72]. For example, if there is an attack report that illustrates that an attack damaged a device by exploring a vulnerability and the same vulnerability exists in a device of an organization, the attack may also damage the organization device. As a result, a security expert is able to apply defenses prior to the possible unhappened attacks.

However, this method can only predict happened attacks, which means that only attacks and threats that appear in the collected texts can be predicted. How to predict unhappened attacks keeps being a problem and challenges.

V. Concluding Remarks

A. Lessons Learned

Cyber Threat Intelligence (CTI) mining is a powerful tool that can provide valuable insights into potential cyber threats and attacks, enabling proactive defense measures to be taken. To generate robust and actionable intelligence, we need to conduct CTI mining with diverse data sources, including open-source and classified information. This involves a variety of techniques, such as data collection, pre-processing, feature extraction, and machine learning algorithms, which must be carefully selected and optimized to achieve accurate and reliable results. However, CTI mining has its challenges. The high volume and complexity of data, the need for real-time analysis, and the difficulty of distinguishing between genuine threats and false positives can all pose significant obstacles. Quality control is essential in CTI mining to ensure accuracy and consistency in the extracted intelligence, avoiding the risk of making decisions based on incomplete or inaccurate information. CTI mining is an ongoing process that requires constant monitoring and adaptation to keep pace with the rapidly evolving threat landscape. Nonetheless, it can have significant benefits for both academia and industry. These include improved threat detection and response, enhanced cybersecurity posture, and increased awareness of emerging threats and trends. Overall, our review of the state-of-the-art works on CTI mining revealed that this field is complex and challenging, but ultimately valuable, capable of enhancing our ability to defend against cyberattacks.

B. Conclusion

In this survey, we provided a detailed review of the most significant works on CTI mining that have been published so far. In our paper, we proposed a classification scheme for organizing and categorizing existing research works on the basis of the purposes of CTI knowledge acquisition, and we highlighted the methodology adopted by the existing studies. In accordance with the proposed classification scheme, we thoroughly review and discuss current works, including cybersecurity related entities and events, cyber attack tactics, techniques and procedures, profiles of hackers, indicators of compromise, vulnerability exploits and malware implementation, and threat hunting. Furthermore, we discussed current challenges and promising future research directions. Over the past several decades, there has been tremendous interest in CTI mining, specifically for proactive cybersecurity defense. Many people have come to the attention that an enormous number of new techniques and models are developed every year. Hopefully, this survey helps readers understand the critical aspects of this field, clarifies the most notable advances, and sheds light on future research.

Acknowledgment

The authors wish to acknowledge the anonymous reviewers for their valuable comments and extend special thanks to Lihua Wang (University of New South Wales) for assisting in the preparation of this manuscript.

References

[1] "SolarWinds hackers linked to known Russian spying tools, investigators say." 2022. Accessed: Oct. 10, 2022. [Online]. Available: https://cybernews.com/news/solarwinds-hackers-linked-to-known-russian-spying-tools-investigators-say/

[2] R. McMillan. "Definition: Threat intelligence." Accessed: Nov. 10, 2022. [Online]. Available: https://gartner.com/

[3] D. Shackleford, *Who's Using Cyberthreat Intelligence and How*, SANS Inst., North Bethesda, MD, USA, 2015.

[4] H. Dalziel, *How to Define and Build an Effective Cyber Threat Intelligence Capability*, Syngress, Waltham, MA, USA, 2014.

[5] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1197–1227, 2nd Quart., 2015.

[6] J. Robertson et al., *Darkweb Cyber Threat Intelligence Mining*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[7] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber attacks," *Comput. Security*, vol. 72, pp. 212–233, Jan. 2018.

[8] T. D. Wagner, K. Mahbub, E. Palomar, and A. E. Abdallah, "Cyber threat intelligence sharing: Survey and research directions," *Comput. Security*, vol. 87, Nov. 2019, Art. no. 101589.

[9] M. S. Abu, S. R. Selamat, A. Ariffin, and R. Yusof, "Cyber threat intelligence—Issue and challenges," *Ind. J. Elect. Eng. Comput. Sci.*, vol. 10, no. 1, pp. 371–379, 2018.

[10] A. Ibrahim, D. Thiruvady, J.-G. Schneider, and M. Abdelrazek, "The challenges of leveraging threat intelligence to stop data breaches," *Front. Comput. Sci.*, vol. 2, p. 36, Aug. 2020.

[11] M. R. Rahman, R. Mahdavi-Hezaveh, and L. Williams, "What are the attackers doing now? Automating cyber threat intelligence extraction from text on pace with the changing threat landscape: A survey," 2021, *arXiv:2109.06808*.

[12] M. R. Rahman, R. Mahdavi-Hezaveh, and L. Williams, "A literature review on mining cyberthreat intelligence from unstructured texts," in *Proc. Int. Conf. Data Min. Workshops (ICDMW)*, 2020, pp. 516–525.

[13] R. Brown and P. Stirparo, *SANS 2022 Cyber Threat Intelligence Survey*, SANS Inst., North Bethesda, MD, USA, 2022.

[14] A. Ramsdale, S. Shiaeles, and N. Kolokotronis, "A comparative analysis of cyber-threat intelligence sources, formats and languages," *Electronics*, vol. 9, no. 5, p. 824, 2020.

[15] "What is cyber threat intelligence? 2022 threat intelligence report." 2022. Accessed: Feb. 13, 2023. [Online]. Available: https://www.crowdstrike.com/cybersecurity-101/threat-intelligence/

[16] N. Sun, C.-T. Li, H. Chan, M. Z. Islam, M. R. Islam, and W. Armstrong, "How do organizations seek cyber assurance? Investigations on the adoption of the common criteria and beyond," *IEEE Access*, vol. 10, pp. 71749–71763, 2022.

[17] N. Sun, J. Zhang, S. Gao, L. Y. Zhang, S. Camtepe, and Y. Xiang, "Data analytics of crowdsourced resources for cybersecurity intelligence," in *Proc. 14th Int. Conf. Netw. Syst. Security (NSS)*, Melbourne, VIC, Austraila, Nov. 2020, pp. 3–21.

[18] "AlienVault open threat intelligence." 2022. Accessed: Oct. 10, 2022. [Online]. Available: https://otx.alienvault.com/

[19] "A community OpenIOC resource." Accessed: Oct. 10, 2022. [Online]. Available: https://openiocdb.com/

[20] "IOCbucket." Accessed: Oct. 10, 2022. [Online]. Available: https://www.iocbucket.com/

[21] "Facebook ThreatExchange." 2022. Accessed: Oct. 10, 2022. [Online]. Available: https://developers.facebook.com/products/threat-exchange

[22] "National vulnerability database." Accessed: Oct. 10, 2022. [Online]. Available: https://nvd.nist.gov/vuln

[23] "2018 verizon annual data breach investigations report." Accessed: Nov. 10, 2022. [Online]. Available: https://www.verizonenterprise.com/verizon-insights-lab/dbir/

[24] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xiang, "Data-driven cybersecurity incident prediction: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1744–1772, 2nd Quart., 2018.

[25] "Defense industrial base cybersecurity information sharing program." 2022. Accessed: Oct. 10, 2022. [Online]. Available: https://dibnet.dod.mil/portal/intranet/

[26] R. Borden, J. Mooney, M. Taylor, and M. Sharkey, "Threat information sharing under GDPR," *Scitech Lawyer*, vol. 15, no. 3, pp. 30–35, 2019.

[27] NIST. "Alert." Accessed: Nov. 10, 2022. [Online]. Available: https://csrc.nist.gov/glossary/term/alert

[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[29] "Shodan." 2019. Accessed: Apr. 2, 2022. [Online]. Available: https://www.shodan.io/

[30] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[31] S. Chakrabarti et al., "Data mining curriculum: A proposal (version 1.0)," in *Proc. Intensive Workshop ACM SIGKDD Curriculum Committee*, vol. 140, 2006, pp. 1–10.

[32] Y. Liu et al., "Cloudy with a chance of breach: Forecasting cyber security incidents," in *Proc. 24th USENIX Security Symp. (USENIX Security)*, 2015, pp. 1009–1024.

[33] I. Deliu, C. Leichter, and K. Franke, "Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2018, pp. 5008–5013.

[34] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in *Proc. ACM Conf. Inf. Knowl. Manag.*, 2017, pp. 1049–1057.

[35] N. Dionísio, F. Alves, P. M. Ferreira, and A. Bessani, "Cyberthreat detection from Twitter using deep neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8.

[36] T. Satyapanich, F. Ferraro, and T. Finin, "CASIE: Extracting cybersecurity event information from text," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 8749–8757.

[37] Y. Fang, Y. Zhang, and C. Huang, "CyberEyes: Cybersecurity entity recognition model based on graph convolutional network," *Comput. J.*, vol. 64, no. 8, pp. 1215–1225, 2021.

[38] H. M. D. Trong, D.-T. Le, A. P. B. Veyseh, T. Nguyen, and T. H. Nguyen, "Introducing a new dataset for event detection in cybersecurity texts," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2020, pp. 5381–5390.

[39] "ENISA risk management—Glossary." [Online]. Available: https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/risk-management-inventory/glossary

[40] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[43] A. Daigavane, B. Ravindran, and G. Aggarwal, "Understanding convolutions on graphs," *Distill*, vol. 6, no. 9, p. e32, 2021.

[44] H. Yan, X. Jin, X. Meng, J. Guo, and X. Cheng, "Event detection with multi-order graph convolution and aggregated attention," in *Proc. Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5766–5770.

[45] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," 2019, *arXiv:1909.03546*.

[46] NIST. "Tactics, techniques, and procedures (TTP)." Accessed: Nov. 10, 2022. [Online]. Available: https://csrc.nist.gov/glossary/term/tactics_techniques_and_procedures

[47] Y. Wu et al., "Price TAG: Towards semi-automatically discovery tactics, techniques and procedures of E-commerce cyber threat intelligence," *IEEE Trans. Depend. Secure Comput.*, early access, Oct. 15, 2021, doi: 10.1109/TDSC.2021.3120415.

[48] 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 *Proc. 33rd Annu. Comput. Security Appl. Conf. (ACSAC)*, 2017, pp. 103–115.

[49] "Adversarial tactics, techniques & common knowledge (ATT&CK)." Accessed: Nov. 10, 2022. [Online]. Available: https://attack.mitre.org/5

[50] "Common attack pattern enumerations and classifications (CAPEC)." Accessed: Nov. 10, 2022. [Online]. Available: https://capec.mitre.org/

[51] W. Ge and J. Wang, "SeqMask: Behavior extraction over cyber threat intelligence via multi-instance learning," *Comput. J.*, to be published.

[52] Y. You et al., "TIM: Threat context-enhanced TTP intelligence mining on unstructured threat data," *Cybersecurity*, vol. 5, no. 1, p. 3, 2022.

[53] "Definitive guide to cyber threat intelligence." 2015. Accessed: Nov. 10, 2022. [Online]. Available: https://cryptome.org/2015/09/cti-guide.pdf

[54] "A structured language for cyber threat intelligence: Structured threat information expression (STIX)." Accessed: Nov. 10, 2022. [Online]. Available: https://oasis-open.github.io/cti-documentation/

[55] M.-C. De Marneffe and C. D. Manning, "The stanford typed dependencies representation," in *Proc. Workshop Cross Framework Cross Domain Parser Eval. (Coling)*, 2008, pp. 1–8.

[56] S. Samtani, R. Chinn, H. Chen, and J. F. Nunamaker, "Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence," *J. Manag. Inf. Syst.*, vol. 34, no. 4, pp. 1023–1053, 2017.

[57] U. Noor, Z. Anwar, T. Amjad, and K.-K. R. Choo, "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise," *Future Gener. Comput. Syst.*, vol. 96, pp. 227–242, Jul. 2019.

[58] L. Perry, B. Shapira, and R. Puzis, "No-doubt: Attack attribution based on threat intelligence reports," in *Proc. IEEE Int. Conf. Intell. Security Inf. (ISI)*, 2019, pp. 80–85.

[59] "APT groups and operations." Accessed: Nov. 10, 2022. https://www.ibm.com/au-en/topics/threat-hunting#::text=Thre%20hunti%%20al%20kno%20as, threa%%20with%20%20organizati%2%20network

[60] J. Grisham, S. Samtani, M. Patton, and H. Chen, "Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence," in *Proc. IEEE Int. Conf. Intell. Security Inf. (ISI)*, 2017, pp. 13–18.

[61] D. Sahoo, "Cyber threat attribution with multi-view heuristic analysis," in *Handbook of Big Data Analytics and Forensics*. Cham, Switzerland: Springer, 2022, pp. 53–73.

[62] H. Hettema, "Rationality constraints in cyber defense: Incident handling, attribution and cyber threat intelligence," *Comput. Security*, vol. 109, Oct. 2021, Art. no. 102396.

[63] S. Tabassum, F. S. Pereira, S. Fernandes, and J. Gama, "Social network analysis: An overview," *Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 8, no. 5, 2018, Art. no. e1256.

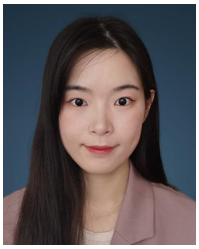
[64] K.-K. R. Choo, "Cyber threat landscape faced by financial and insurance industry," in *Trends Issues Crime Criminal Justice*. Sydney, NSW, Australia: Aust. Inst. Criminol., 2011.

[65] M. Bromiley, *Threat Intelligence: What It Is, and How to Use It Effectively*, SANS Inst., North Bethesda, MD, USA, 2016.

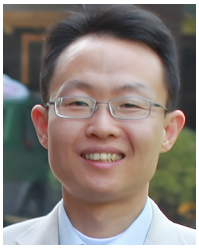
[66] V. Benjamin, W. Li, T. Holt, and H. Chen, "Exploring threats and vulnerabilities in hacker Web: Forums, IRC and carding shops," in *Proc. IEEE Int. Conf. Intell. Security Inf. (ISI)*, 2015, pp. 85–90.

[67] S. Samtani, K. Chinn, C. Larson, and H. Chen, "Azsecure hacker assets portal: Cyber threat intelligence and malware analysis," in *Proc. IEEE Conf. Intell. Security Inf. (ISI)*, 2016, pp. 19–24.
[68] S. Samtani and H. Chen, "Using social network analysis to identify key hackers for keylogging tools in hacker forums," in *Proc. IEEE Conf. Intell. Security Inf. (ISI)*, 2016, pp. 319–321.
[69] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. A. Beyah, "Acing the IoC game: Toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2016, pp. 755–766.
[70] S. Zhou, Z. Long, L. Tan, and H. Guo, "Automatic identification of indicators of compromise using neural-based sequence labelling," 2018, *arXiv:1810.10156*.
[71] Z. Long, L. Tan, S. Zhou, C. He, and X. Liu, "Collecting indicators of compromise from unstructured text of cybersecurity articles using neural-based sequence labelling," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8.
[72] J. Zhao, Q. Yan, X. Liu, B. Li, and G. Zuo, "Cyber threat intelligence modeling based on heterogeneous graph convolutional network," in *Proc. 23rd Int. Symp. Res. Attacks Intrusions Defenses (RAID)*, 2020, pp. 241–256.
[73] Z. Zhu and T. Dumitras, "Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports," in *Proc. IEEE Eur. Symp. Security Privacy (Euro S&P)*, 2018, pp. 458–472.
[74] J. Liu et al., "TriCTI: An actionable cyber threat intelligence discovery system via trigger-enhanced neural network," *Cybersecurity*, vol. 5, no. 1, p. 8, 2022.
[75] Z. Li, J. Zeng, Y. Chen, and Z. Liang, "AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports," in *Proc. 27th Eur. Symp. Res. Comput. Security (ESORICS)*, Copenhagen, Denmark, Sep. 2022, pp. 589–609.
[76] H. Shin, W. Shim, S. Kim, S. Lee, Y. G. Kang, and Y. H. Hwang, "# twiti: Social listening for threat intelligence," in *Proc. Web Conf.*, 2021, pp. 92–104.
[77] L. Luo, Y. Fang, X. Cao, X. Zhang, and W. Zhang, "Detecting communities from heterogeneous graphs: A context path-based graph neural network model," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.*, 2021, pp. 1170–1180.
[78] C. Sabottke, O. Suciu, and T. Dumitras, "Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits," in *Proc. 24th USENIX Security Symp. (USENIX Security)*, 2015, pp. 1041–1056.
[79] Z. Zhu and T. Dumitraș, "FeatureSmith: Automatically engineering features for malware detection by mining the security literature," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2016, pp. 767–778.
[80] S. K. Lim, A. O. Muis, W. Lu, and C. H. Ong, "MalwareTextDB: A database for annotated malware articles," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguist. Long Papers*, vol. 1, 2017, pp. 1557–1567.
[81] A. Roy, Y. Park, and S. Pan, "Predicting malware attributes from cybersecurity texts," in *Proc. Conf. North Amer. Assoc. Comput. Linguist. Human Lang. Technol.*, vol. 1, 2019, pp. 2857–2861.
[82] Y. Chen et al., "Devils in the guidance: Predicting logic vulnerabilities in payment syndication services through automated documentation analysis," in *Proc. 28th USENIX Security Symp. (USENIX Security)*, 2019, pp. 747–764.
[83] Y. Chen et al., "Bookworm game: Automatic discovery of LTE vulnerabilities through documentation analysis," in *Proc. IEEE Symp. Security Privacy (S&P)*, 2021, pp. 1197–1214.
[84] A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, and R. Zak, "Creating cybersecurity knowledge graphs from malware after action reports," *IEEE Access*, vol. 8, pp. 211691–211703, 2020.
[85] "Open sourced vulnerability database." Accessed: Oct. 10, 2022. [Online]. Available: http://www.osvdb.org/
[86] E. Nunes et al., "Darknet and DeepNet mining for proactive cybersecurity threat intelligence," in *Proc. IEEE Conf. Intell. Security Inf. (ISI)*, 2016, pp. 7–12.
[87] "SemEval." Accessed: Oct. 10, 2022. [Online]. Available: https://semeval.github.io//
[88] "Word2vec—TensorFlow core." Accessed: Nov. 10, 2022. [Online]. Available: https://www.tensorflow.org/tutorials/text/word2vec
[89] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2005, pp. 363–370.
[90] "Internet security systems X-force security threats." Accessed: Nov. 10, 2022. [Online]. Available: http://xforce.iss.net
[91] "US-CERT, vulnerability notes database." Accessed: Nov. 10, 2022. [Online]. Available: http://www. kb.cert.org/vuls/
[92] A. Sapienza, S. K. Ernala, A. Bessi, K. Lerman, and E. Ferrara, "DISCOVER: Mining online chatter for emerging cyber threats," in *Proc. Companion Web Conf.*, 2018, pp. 983–990.
[93] S. Samtani, H. Zhu, and H. Chen, "Proactively identifying emerging hacker threats from the dark Web: A diachronic graph embedding framework (D-GEF)," *ACM Trans. Privacy Security*, vol. 23, no. 4, pp. 1–33, 2020.
[94] K. Satvat, R. Gjomemo, and V. Venkatakrishnan, "EXTRACTOR: Extracting attack behavior from threat reports," in *Proc. IEEE Eur. Symp. Security Privacy (Euro S&P)*, 2021, pp. 598–615.
[95] Y. Gao, X. Li, H. Peng, B. Fang, and P. S. Yu, "HinCTI: A cyber threat intelligence modeling and identification system based on heterogeneous information network," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 708–722, Feb. 2022.
[96] S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. Venkatakrishnan, "HOLMES: Real-time APT detection through correlation of suspicious information flows," in *Proc. IEEE Symp. Security Privacy (S&P)*, 2019, pp. 1137–1152.
[97] S. M. Milajerdi, B. Eshete, R. Gjomemo, and V. Venkatakrishnan, "POIROT: Aligning attack behavior with kernel audit records for cyber threat hunting," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2019, pp. 1795–1812.
[98] A. Nadeem, S. Verwer, S. Moskal, and S. J. Yang, "Alert-driven attack graph generation using S-PDFA," *IEEE Trans. Depend. Secure Comput.*, vol. 19, no. 2, pp. 731–746, Mar./Apr. 2022.
[99] P. Gao et al., "Enabling efficient cyber threat hunting with cyber threat intelligence," in *Proc. IEEE 37th Int. Conf. Data Eng. (ICDE)*, 2021, pp. 193–204.
[100] W. Yang and K.-Y. Lam, "Automated cyber threat intelligence reports classification for early warning of cyber attacks in next generation SOC," in *Proc. Int. Conf. Inf. Commun. Security*, 2019, pp. 145–164.
[101] B. Biswas, A. Mukhopadhyay, S. Bhattacharjee, A. Kumar, and D. Delen, "A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums," *Decis. Support Syst.*, vol. 152, Jan. 2022, Art. no. 113651.
[102] "Common vulnerabilities and exposures." Accessed: Mar. 11, 2022. [Online]. Available: http://cve.mitre.org/
[103] B. Bhattarai and H. H. Huang, "SteinerLog: Prize collecting the audit logs for threat hunting on enterprise network," in *Proc. ACM Asia Conf. Comput. Commun. Security (Asia CCS)*, 2022, pp. 97–108.
[104] W. U. Hassan et al., "This is why we can't cache nice things: Lightning-fast threat hunting using suspicion-based hierarchical storage," in *Proc. Annu. Comput. Security Appl. Conf. (ACSAC)*, 2020, pp. 165–178.
[105] S. Purohit et al., "Cyber threat intelligence sharing for cooperative defense in multi-domain entities," *IEEE Trans. Depend. Secure Comput.*, early access, Oct. 13, 2022, doi: 10.1109/TDSC.2022.3214423.
[106] T. Schaberreiter et al., "A quantitative evaluation of trust in the quality of cyber threat intelligence sources," in *Proc. 14th Int. Conf. Availability Rel. Security*, 2019, pp. 1–10.
[107] H. Griffioen, T. Booij, and C. Doerr, "Quality evaluation of cyber threat intelligence feeds," in *Proc. 18th Int. Conf. Appl. Cryptography Netw. Security (ACNS)*, Rome, Italy, Oct. 2020, pp. 277–296.
[108] D. Schlette, F. Böhm, M. Caselli, and G. Pernul, "Measuring and visualizing cyber threat intelligence quality," *Int. J. Inf. Security*, vol. 20, pp. 21–38, Feb. 2021.
[109] P. Rajesh, M. Alam, M. Tahernezhadi, A. Monika, and G. Chanakya, "Analysis of cyber threat detection and emulation using mitre attack framework," in *Proc. IEEE Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, 2022, pp. 4–12.
[110] K. Liu, F. Wang, Z. Ding, S. Liang, Z. Yu, and Y. Zhou, "Recent progress of using knowledge graph for cybersecurity," *Electronics*, vol. 11, no. 15, p. 2287, 2022.
[111] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in Twitter stream," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2018, pp. 5002–5007.
[112] "Sec2vec." Accessed: Oct. 10, 2022. [Online]. Available: https://github.com/0xyd/sec2vec
[113] K. Dey, R. Shrivastava, S. Kaushik, and L. V. Subramaniam, "EmTaggeR: A word embedding based novel method for hashtag recommendation on Twitter," in *Proc. IEEE Int. Conf. Data Min. Workshops (ICDMW)*, 2017, pp. 1025–1032.

[114] Y. Guo, J. Liu, W. Tang, and C. Huang, "Exsense: Extract sensitive information from unstructured data," *Comput. Security*, vol. 102, Mar. 2021, Art. no. 102156.
[115] N. Sun, J. Zhang, S. Gao, L. Y. Zhang, S. Camtepe, and Y. Xiang, "Cyber information retrieval through pragmatics understanding and visualization," *IEEE Trans. Depend. Secure Comput.*, vol. 20, no. 2, pp. 1186–1199, Mar./Apr. 2023.
[116] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin, "Generating fake cyber threat intelligence using transformer-based models," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2021, pp. 1–9.
[117] M. Maasberg, E. Ayaburi, C. Liu, and Y. Au, "Exploring the propagation of fake cyber news: An experimental approach," in *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, 2018, pp. 1–10.
[118] D. Liebowitz et al., "Deception for cyber defence: Challenges and opportunities," in *Proc. IEEE 3rd Int. Conf. Trust Privacy Security Intell. Syst. Appl. (TPS-ISA)*, 2021, pp. 173–182.
[119] "The forrester threat report: The emergence of offensive AI." Accessed: Nov. 10, 2022. [Online]. Available: https://knowledgehubmedia.com/the-forrester-threat-report-the-emergence-of-offensive-ai-3/
[120] N. Sun et al., "Defining security requirements with the common criteria: Applications, adoptions, and challenges," *IEEE Access*, vol. 10, pp. 44756–44777, 2022.
[121] "Why human error is #1 cyber security threat to businesses in 2021." 2022. Accessed: Nov. 9, 2022. [Online]. Available: https://thehackernews.com/2021/02/why-human-error-is-1-cyber-security.html#::text='Hum%20err%20w%2%20major,%20%%20%20a%20breaches.&text=Mitigati%20%20hum%20err%20must,cyb%20busine%20securi%20%202021
[122] Y. Shen, E. Mariconti, P.-A. Vervier, and G. Stringhini, "Tiresias: Predicting security events through deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2018, pp. 592–605.
[123] N. Sun, J. Zhang, S. Gao, L. Y. Zhang, S. Camtepe, and Y. Xiang, "My security: An interactive search engine for cybersecurity," in *Proc. 54th Hawaii Int. Conf. Syst. Sci. (HICSS-54)*, 2021, pp. 6206–6215.

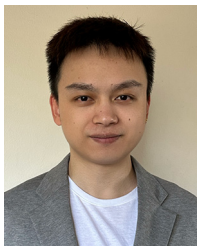


Nan Sun received the B.S. degree (Hons.) in software engineering and the Ph.D. degree in cybersecurity from Deakin University. She is currently a Lecturer of Cybersecurity with the School of Engineering and Information Technology, University of New South Wales (UNSW), Canberra, Australia. Before joining UNSW, she was a Postdoctoral Research Fellow with Deakin University. Her current research focuses on cybersecurity.



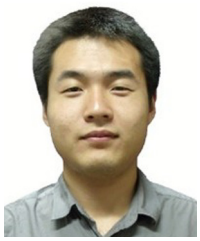
Ming Ding (Senior Member, IEEE) received the B.S. and M.S. degrees (with First-Class Hons.) in electronics engineering and the Ph.D. degree in signal and information processing from Shanghai Jiao Tong University, Shanghai, China, in 2004, 2007, and 2011, respectively. From April 2007 to September 2014, he worked with the Sharp Laboratories of China, Shanghai, as a Researcher/Senior Researcher/Principal Researcher. He is currently a Principal Research Scientist with Data61, CSIRO, Sydney, NSW, Australia. He has authored more than 200 papers in IEEE journals and conferences, all in recognized venues, and around 20 3GPP standardization contributions, as well as two books, i.e., *Multi-Point Cooperative Communication Systems: Theory and Applications* (Springer, 2013) and *Fundamentals of Ultra-Dense Wireless Networks* (Cambridge University Press, 2022). Also, he holds 21 U.S. patents and has co-invented another 100+ patents on 4G/5G technologies. His research interests include information technology, data privacy and security, and machine learning and AI. He is currently an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. Besides, he has served as a guest editor/co-chair/co-tutor/TPC member for multiple IEEE top-tier journals/conferences and received several awards for his research work and professional services, including the prestigious IEEE Signal Processing Society Best Paper Award in 2022.

Jiaojiao Jiang received the Ph.D. degree from Deakin University, Melbourne, VIC, Australia. She is currently a Lecturer with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia. She has authored or coauthored more than 30 articles in high-quality journals and conferences, including IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE Trustcom, and IEEE Globecom. Her research interests include cybersecurity, complex networks, and service virtualization.

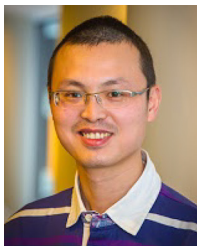


Weikang Xu received the master's degree in information technology from the University of New South Wales in 2020, where he is currently pursuing the Ph.D. degree. His current research interests include cyber security and cyber threat intelligence.

Xiaoxing Mo is currently pursuing the Ph.D. degree in cyber security and AI with Deakin University, Australia. His research focuses on integrating AI into cyber security to develop more robust and effective security strategies. He is also dedicated to contributing to the wider academic community, sharing his expertise and learning from other experts in the field.



Yonghang Tai received the Ph.D. degree in computer science from Deakin University, Melbourne, Australia, in 2019. He is a Professor with the Yunnan Key Laboratory of Opto-Electronic Information Technology, Yunnan Normal University. He is also an Associate Researcher with the Cybersecurity Lab and the Digital Research and Innovation Capability Platform, Swinburne University of Technology, Melbourne. He has published more than 80 research papers in many international journals and conferences, e.g., IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON HAPTICS, IEEE INTERNET OF THINGS JOURNAL, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS and conferences, e.g., IEEE INFOCOM, IEEE IECON, and IEEE SECON. His research interests include physics-based simulation, AI-based medical implementations, virtual reality/augmented reality, Internet of Medical Things, and big data. He gave nine keynotes at international conferences. He is widely regarded as one of the most active and influential young scientists and experts in VR/AR/IoT/data science/AI, as he has the experience to develop ten different services for multiple disciplines.



Jun Zhang (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Wollongong, Wollongong, NSW, Australia, in 2011. He is currently a Full Professor and the Director of the Cybersecurity Lab, Swinburne University of Technology, Australia. He was recognized in The Australian's top researchers special edition publication as the Leading Researcher in the field of computer security and cryptography in 2020. He leaded Swinburne cybersecurity research and produced excellent outcome, including many high-impact research papers and multimillion-dollar research projects. Swinburne was named in The Australian's 2021 Research Magazine, the top research institution in the field of computer security and cryptography. He has been serving as a Steering Committee Member of the P-TECH Program at Melbourne since 2019, in which the Australian Government invested in, promoting STEM education. He devotes himself to communication and community engagement, boosting the awareness of cybersecurity.