Chapter 6
The Advancement of Knowledge Graphs in Cybersecurity: A Comprehensive Overview

Yuke Ma, Yonggang Chen, Yanjun Wang, Jun Yu, Yanting Li, Jinyu Lu, and Yong Wang

Abstract With the increasing complexity of artificial intelligence technology and network environments, cybersecurity is facing massive and complex data. Knowledge graphs have the potential to aggregate, represent, manage, and reason with this knowledge. Therefore, applying knowledge graphs to cybersecurity can help to characterize and present security situations, support security decision-making, and predict warnings. Over the past two decades, research on knowledge graphs for cybersecurity has received growing attention in data processing, construction, and visualization. This review provides a comprehensive comparative analysis of key technologies and application scenarios of cybersecurity knowledge graphs. Firstly, basic concepts of knowledge graphs and cybersecurity knowledge graphs are outlined, and the required datasets for their construction are compared and analyzed from both general-purpose and specialized perspectives. On this basis, a framework for building cybersecurity knowledge graphs is summarized, and key techniques for building cybersecurity knowledge graphs, including ontology construction, information extraction, and knowledge reasoning, are detailed. Finally, application scenarios of knowledge graphs in the field of cybersecurity are sorted out from the perspective of application objectives. The challenges knowledge graphs face and future development trends in this field are also pointed out.

Keywords Cybersecurity · Knowledge graph · Knowledge representation · Ontology construction · Information extraction · Knowledge reasoning

Y. Ma · Y. Wang (✉) · J. Yu · Y. Li
College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
e-mail: wyj@zzuli.edu.cn

Y. Chen
The State Information Center, Beijing 100045, China

J. Lu
Henan Province Platform Economy Development Guidance Center, Zhengzhou 450008, China

Y. Wang
Zhengzhou Aiwen Computer Technology Co., Ltd., Zhengzhou 450000, China

S. Li (ed.), *Computational and Experimental Simulations in Engineering*, Mechanisms and Machine Science 145, https://doi.org/10.1007/978-3-031-42987-3_6

6.1 Introduction

In recent years, the global cybersecurity situation is not optimistic. Cyber security is the act of protecting computer systems from attacks or illegal access, and with the increase of new technologies and devices, the causes of cyber attacks and the focus on prevention have diversified. Under the influence of the epidemic, cyber security threats such as security vulnerabilities, ransomware, ransomware, cloud services, obsolete and inefficient systems against infrastructure and important information systems are becoming increasingly serious, and the means of attack are escalating, bringing huge risks to people's lives, economic production, social stability, and national security [1]. With the rapid development of big data and artificial intelligence, new cybersecurity solutions have emerged, leveraging the vast amount of security-related data available in cyberspace. This data includes monitored cybersecurity alert data, vulnerability information repositories, and security notices. By mining the information in these data, security analysts can provide support for cybersecurity situational awareness, realize security alert predictions, and support cybersecurity decisions. However, the characteristics of network security data of massive quantification, decentralization, fragmentation, and hidden relationships make how to analyze and process these massive data in a timely and accurate manner a major problem in the field of network security. Therefore, there is a need to find effective big data analysis technologies and algorithms to achieve rapid processing and analysis of cybersecurity data to better protect cybersecurity.

Cybersecurity knowledge graph is a method that uses knowledge mapping technology to model and expresses knowledge in the field of cybersecurity, aiming to model the concepts of attackers, targets, tools, vulnerabilities, threats, risks, and other concepts involved in cyberspace and the connections between them into a unified knowledge system, to achieve a comprehensive understanding and effective control of the security posture of cyberspace to support the cyber attack and defense posture perception, threat prediction, risk assessment, and other tasks. The cybersecurity knowledge graph has the following advantages: (1) it can integrate multi-source heterogeneous data and improve data quality and credibility; (2) it can express complex semantic relationships and improve information expressiveness and readability; (3) it can support ontology-based reasoning and improve knowledge discovery and utilization; (4) it can support graph-based analysis and improve data mining and visualization.

At present, some progress has been made in cybersecurity knowledge graph research [2]. Some scholars have proposed ontology-based methods for constructing cybersecurity knowledge graphs, including knowledge extraction, ontology construction, entity identification, and other techniques, as a way to construct cybersecurity knowledge graphs. Meanwhile, some researchers are also exploring how to integrate multi-source heterogeneous data into the cybersecurity knowledge graph, to improve the completeness and accuracy of the graph. In addition, some scholars

are also studying how to use knowledge graph technology for cybersecurity situational awareness and threat intelligence analysis to achieve cybersecurity intelligence. Overall, there are still many challenges and problems in the research of cybersecurity knowledge graphs, which need further in-depth exploration. In this paper, we focus on various key technologies and application scenarios of knowledge graphs in cybersecurity, conduct a more complete and in-depth review, and give some issues that may exist or are worth exploring in the future.

This paper is organized as follows: Sect. 6.2 of this paper provides a brief overview of the development of general knowledge graph construction techniques as well as their application in the field of cybersecurity knowledge graph construction; Sect. 6.3 summarizes and analyzes the relevant datasets of cybersecurity knowledge graph; Sect. 6.4 proposes a framework for cybersecurity knowledge graph construction and sorts out the key technologies for cybersecurity knowledge graph construction; Sect. 6.5 gives the application scenarios of cybersecurity knowledge graph; Sect. 6.6 looks forward to the cybersecurity The future research direction of the knowledge graph. Finally, the whole paper is summarized.

6.2 Background Knowledge

6.2.1 Knowledge Graph

The early idea of the Knowledge Graph originated from the vision of Tim Berners-Lee, the father of the World Wide Web, on Semantic Web [3]. The core idea of the Semantic Web is to add machine-understandable semantic information to web data, thus improving the comprehension ability of machines. In 2012, Google introduced the concept of Knowledge Graph to enhance the search quality and user experience of search engines [4]. Subsequently, Knowledge Graph has been extensively utilized in diverse domains, including but not limited to finance, education, and medicine.

Knowledge graph is a structured data model for representing and storing knowledge, which organizes knowledge elements such as entities, concepts, relationships, and attributes in graph form to achieve description, query, reasoning and application of knowledge [5]. In essence, a knowledge graph is a semantic network that shows entities and relationships between entities and is a formal description of things and relationships in the real world. Knowledge graphs are generally represented by a triad, i.e. $K = (E, R, S)$ where K denotes the knowledge base; $E = \{e_1, e_2, \ldots, e_{|E|}\}$ denotes the set of entities in K. There are entities in the set of entities $|E|$ kinds. $R = \{r_1, r_2, \ldots, r_{|R|}\}$ denotes the set of relations in K, and There are $|R|$ different kinds of relations in the set of relations. The basic forms of the triples are <concept, attribute, attribute value> and <entity1, relationship, entity2>, etc. The most fundamental components of knowledge graph K are entities, which are interconnected through different types of relationships. Concepts are used to describe the category or type of things, objects, or collections, including place names and individuals.

Attributes, such as birthplace and birth year, refer to the inherent characteristics and features of entities. Attribute values represent the specific values of the attributes assigned to entities or relationships, such as "Beijing." A unique identifier can be assigned to each entity, while attribute and value pairs are used to describe the entity's intrinsic characteristics. Relationships connect entities and indicate the associations between them.

The logical structure of the knowledge graph can be divided into a schema layer and a data layer. The schema layer is the basis of knowledge graph construction, which defines the meta-information and meta-structure of data and is usually designed through an ontology library. Ontology is a structured knowledge representation, which describes the relationships and attributes between different concepts. Data layers are concrete knowledge instances, including entities, relationships, and attributes, generated according to the ontology specification of the schema layer. They describe specific knowledge facts of a class or a concept. The schema layer and the data layer can be compared to the relationship between the skeleton and the flesh and blood.

Figure 6.1 illustrates the general process of constructing a knowledge graph, which is continually updated and refined through cognitive ability. Generally, there are two methods for building knowledge graphs: top-down and bottom-up. The top-down approach involves defining ontology and the knowledge graph data model first, and then adding knowledge to the database. On the other hand, the bottom-up approach involves extracting knowledge from open unstructured data and selecting those with higher confidence levels to add to the knowledge graph. This approach also involves building the top-level ontology model.

There are two main categories of knowledge graphs based on their knowledge scope and application scenarios, namely general knowledge graphs and domain-specific knowledge graphs. General knowledge graphs are extensive knowledge

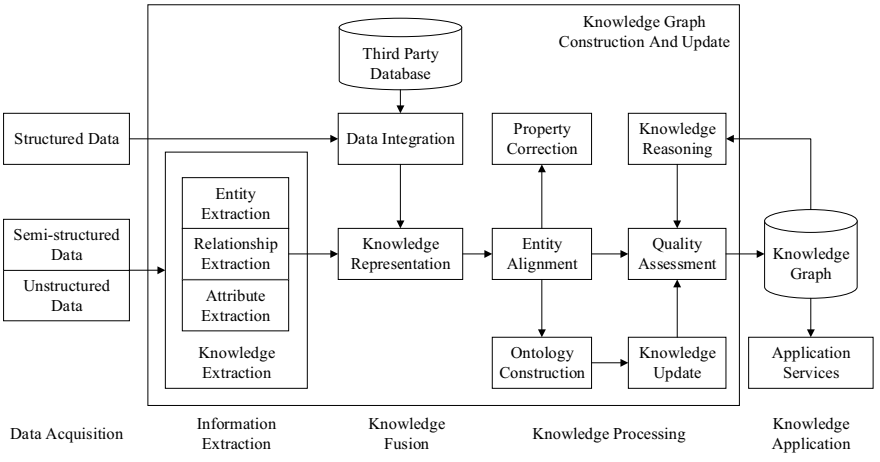

Fig. 6.1 The general construction process of knowledge graph

bases, such as Freebase [6], Yago [7], DBpedia [8], etc., which are mainly used for pervasive intelligent search and recommendation scenarios and provide broad basic knowledge associations. Domain-specific knowledge graphs, on the other hand, build a knowledge space with depth based on a certain knowledge sub-domain to serve specific query and analysis requirements in the domain. The cybersecurity knowledge graph can be positioned as a domain-specific knowledge graph in this classification system, providing deep knowledge space and specific query and analysis demand services in the cybersecurity domain, and we will introduce the cybersecurity knowledge graph in Sect. 6.2.2.

6.2.2 Cybersecurity Knowledge Graph

The technology of knowledge graphs has broad applications in the field of cybersecurity [9]. As the network environment becomes increasingly complex, the amount of data involved in cybersecurity is also rapidly expanding. Consequently, there is a pressing need to identify correlations and attack patterns from the vast, fragmented, and heterogeneous data related to cyberspace that come from multiple sources. Network security knowledge mapping is to organize and manage a massive amount of network security knowledge with the technology of knowledge mapping. The present challenge in conducting cybersecurity posture analysis is not the absence of available information, but rather, the difficulty in integrating heterogeneous information from multiple sources into a single model. This integration is crucial in obtaining a comprehensive understanding of the cybersecurity posture and providing decision support. Knowledge graph technology can effectively integrate heterogeneous cybersecurity data from multiple sources, build a security knowledge ontology architecture, and support tasks such as cybersecurity situational awareness and early warning prediction.

The role of knowledge graph technology in the cybersecurity field is mainly in the following aspects: integrating multi-source heterogeneous cybersecurity data to build a unified and structured cyberspace situational portrait; revealing the intrinsic connections and influencing factors among different entities to provide rich cyberspace correlation analysis capabilities; supporting intelligent reasoning capabilities to realize dynamic perception and prediction of cyber threat behaviors and risk states; and providing other cybersecurity technologies with reliable data sources and decision bases. When introducing knowledge graph into the cybersecurity field, core concepts and relationships can be defined based on ontology and threat modeling methods, entity and relationship information can be extracted from multi-source heterogeneous cybersecurity data, and data in the security knowledge graph can be stored and managed using graph databases and graph computing frameworks, and complex reasoning tasks can be performed based on methods such as logical reasoning or machine learning, such as attack path analysis, risk assessment, and anomaly detection, etc.

The construction of cybersecurity knowledge graph generally starts with ontology construction and the design of the conceptual structure of cybersecurity knowledge graph, including the definition of entity types, attribute types and relationship types, as well as the hierarchical relationships and constraints among entities. Reference can be made to existing standards and frameworks in the field of cybersecurity, such as STIX, ATT&CK, CAPEC, etc. Next, knowledge extraction is performed to identify named entities and relationships from unstructured or semi-structured cybersecurity data and map them to ontologies. Natural language processing, machine learning, rule matching, and other methods are generally used. Then, knowledge storage is performed, and the extracted entities and relations are stored in the knowledge graph database to form a cybersecurity knowledge graph. Suitable data models and query languages, such as RDF, SPARQL, etc., can be selected. Finally, knowledge inference is performed using logical reasoning, path search, and machine learning to generate new knowledge to support prediction and inference tasks by using the existing cybersecurity knowledge graph. In the latest study, Ren et al. [10] designed a complete knowledge graph framework including knowledge extraction, knowledge storage, knowledge inference, and knowledge visualization, and used deep learning and expert knowledge to complement and update the knowledge graph.

Relevant cybersecurity knowledge bases have been constructed in the field of cybersecurity [11]. Building a cybersecurity knowledge graph can integrate cybersecurity-related information from different sources, analyze and mine them to understand the cybersecurity posture, and provide a basis for decision-making. This is very important for detecting intrusions and monitoring the cybersecurity posture. Given the increasingly systematic knowledge available in the field of network security, the construction of network security knowledge graphs typically employs a top-down approach. Compared to foreign countries, domestic research on knowledge graphs in the field of cybersecurity has been relatively limited and started later. Iannacone et al. proposes the cyber threat intelligence platform STUCCO [12], which is dedicated to cyber attack detection and contextual understanding; MITRE, the National Security Engineering Center, develops a Neo4j-based cyber attack knowledge graph tool CyGraph [13], which is mainly oriented to cyber warfare task analysis, visualization analysis and knowledge management.

Previous review articles on cybersecurity knowledge graphs are mainly in the areas of building key technologies and cybersecurity assessment. Zhang et al. [14] reviewed the application and development of knowledge graphs in the field of cybersecurity assessment, analyzed the advantages and challenges of knowledge graphs, proposed a framework for cybersecurity assessment based on knowledge graphs, and gave a case study, but only limited to the field of cybersecurity assessment. Li et al. [15] summarize the current research progress of knowledge graph-related technologies at home and abroad and their application status in the field of network security, but only confine to the key technical aspects and do not cover the application scenarios of network security knowledge graph. Ding et al. [16] briefly elaborate on several application directions of network security knowledge graph on the basis of introducing the construction technology of network security knowledge graph. Liu et al. [9] provide an overview of A comparative review of different works describing recent advances

in cybersecurity knowledge application scenarios is presented. Other articles mainly review the research on cybersecurity knowledge graphs in the dimensions of graph-based approaches [17], knowledge inference approaches [18], etc. However, there is still a lack of literature that comprehensively reviews cybersecurity knowledge graph construction techniques and application scenarios. The primary objective of this paper is to offer a comprehensive and in-depth review of the datasets, key technologies, and application scenarios of knowledge graphs in the field of cybersecurity, incorporating the latest advancements.

6.3 Cybersecurity Knowledge Graph Dataset

The internet comprises numerous elements related to cybersecurity, and several security companies have made significant strides in collecting and integrating cyberspace resources [19]. For instance, KnownSec has developed ZoomEye, a cyber radar system; FOFA, a web search engine launched by the 100 Club; and Shodan, a web device search engine. Shodan is a well-known open web search engine that scans internet devices and identifies information about them. Censys is a free search engine that scans IPV4 addresses, domain names and certificates. ZoomEye is a cyberspace search engine with two detection engines that can identify Internet devices and websites.

Cybersecurity knowledge mapping data must usually be obtained from many different sources, the first source is structured data, such as structured intelligence databases and intelligence from STIX. The second source is semi-structured data, such as the knowledge base under MITRE, including CVE, CWE, CAPEC, CPE, ATT&CK and CTI [20]. Such information is typically collected and stored in semi-structured vulnerability databases, including NVD, CNVD, and CNNVD [21]. Public disclosure of important security information also appears in the databases of well-known companies, such as Kaspersky (Kaspersky Anti-Virus [22] is one of the world's most technologically advanced antivirus software), IBM, VERIS Community [23], and other open-source intelligence community sites. Third, security engineers can also find some key information from cybersecurity blogs (such as Talos blog [24]), cybersecurity reports (such as GitHub APT report [25]), Internet chat rooms, and any publicly available cybersecurity texts. These are better resources that can be mined for concepts, abstractions, entities, attributes, and relationships.

In this paper, based on the purpose of constructing knowledge graphs, the cybersecurity knowledge graph datasets are divided into two categories: general-purpose and professional-purpose. The general-purpose knowledge graph aims to cover all aspects of the cybersecurity domain, including threat intelligence, vulnerability information, security events, etc.; while the specialized knowledge graph focuses on a specific domain, such as industrial control system security, cloud security, etc.

6.3.1 Generalized Dataset of Cybersecurity Knowledge Graph

To improve the analysis and decision-making capability of cybersecurity, the literature [26] introduces the SEPSES knowledge graph. The knowledge graph integrates a variety of cybersecurity data and knowledge resources, including data on threat intelligence, vulnerabilities, attack patterns, and security events, and shows the relationships and semantics among them. The SEPSES knowledge graph is represented by RDF/OWL specification and supports cybersecurity analysis and decision-making by providing SPARQL endpoints and web interfaces to query and visualize the data. Among other things, SEPSES Knowledge Graph can help users discover attackers' strategies, assess system vulnerabilities, predict future threats, etc., and improve the visualization and operability of cybersecurity.

To build a comprehensive cybersecurity knowledge graph (CSKG), knowledge mapping techniques and multi-source heterogeneous security knowledge bases are utilized, as discussed in the literature [27]. Publicly available vulnerability and threat bases on the network are analyzed to extract knowledge and form a multi-source heterogeneous information security knowledge graph for threat analysis. The article presents a description of the security knowledge ontology model during the construction process, as well as methods such as threat modeling, which enable the processing and integration of multi-source heterogeneous cyber security domain information into a structured intelligent security domain knowledge base. Among the datasets are NVD, CAPEC, CWE, etc. The article also explores applications of this knowledge graph, such as threat intelligence analysis, malicious activity detection, and advanced persistent threat (APT) organization attribution.

Literature [28] introduces a unique dataset containing manually labeled cybersecurity-related terms in six categories: attackers, targets, vulnerabilities, attack methods, consequences, and solutions. The dataset helps organizations and government agencies to automatically extract cybersecurity terms, quickly understand and discover vulnerabilities in their systems, and take appropriate measures to strengthen security, while also tracking unofficial data sources to discover potential threats. Data sources include open blogs and official company security bulletins, among others.

The data sources used in the study [29] were structured data from the cyber security domain, including the National Vulnerability Database (NVD), the Open Source Vulnerability Database (OSVDB), and the Exploit Database (Exploit-DB). These data sources provide a number of entity tags associated with text descriptions, such as vulnerability name, software name, attack type, etc. The article uses these tags to automatically tag text descriptions with entities, thus generating a corpus containing cybersecurity entities, and exposing the corpus. The application scenario of the article is to use this corpus to train a supervised learning algorithm based on a maximum entropy model to extract cybersecurity entities from other unlabeled texts, such as blogs, news articles, and tweets.

6.3.2 *Cybersecurity Knowledge Graph Professional Type Dataset*

The literature [30] discusses the construction of a new annotated malware text database. The article presents an annotation framework for defining malware features based on MAEC vocabularies and a database containing 39 annotated APT reports with 6819 sentences. The authors also use this database to construct models that can help cybersecurity researchers in data collection and analysis. The advantage of this approach is that it can help cybersecurity researchers to better collect and analyze data, but the disadvantage is that it requires a large amount of annotated data.

Sun et al. [31] use CWE (Common Weakness Enumeration) based knowledge graphs to analyze Twitter data to discover and predict cybersecurity-related events and trends. The article describes the method of constructing CWE knowledge graphs and how to use knowledge graphs and machine learning techniques to classify, cluster, correlate, and visualize Twitter data. The article also shows some experimental results to demonstrate the effectiveness and application value of the CWE knowledge graph-based Twitter data analysis method in the field of cybersecurity.

In [32], a novel approach for the automatic extraction of core information from CTI reports is presented, using a Named Entity Recognition (NER) system. The study also includes the publication of a dataset containing 498,000 tagged examples. With countless CTI reports being used by companies worldwide for security purposes, it is essential to extract useful information from large volumes of textual data to secure critical cybersecurity information. The advantage of this approach is the ability to quickly extract useful information from large amounts of text data, but the disadvantage is the large amount of labeled data required.

Open-CyKG [33] is an open-source knowledge graph framework that aims to extract valuable cyber threat intelligence (CTI) from unstructured advanced persistent threat (APT) reports using an attention mechanism-based neural open information extraction (OIE) model. The framework consists of three modules: data preprocessing, entity recognition and relation extraction. The NER model is designed to identify security-related entities, such as malware names, IP addresses, and file names, while the RE model identifies the relationships between these entities. The attention mechanism is used to improve the accuracy of the extraction by giving more attention to important parts of the text. The extracted information is then represented as a knowledge graph, which can be used for various cybersecurity applications, such as threat analysis and attack prediction. The article also shows some application cases, such as APT organization attribution, attack vector analysis, and vulnerability exploitation analysis using knowledge graphs.

The literature [34] presents MalKG: a framework for generating and predicting knowledge graphs (KGs) related to malware threat intelligence. The article describes methods for collecting and processing malware-related data from different open data sources, and how to use relational extraction (RE) techniques to generate malware knowledge graphs. The article also describes how to use graph neural network

(GNN)-based models to predict missing entities and relationships in the knowledge graph and how to use the knowledge graph for threat intelligence analysis.

Kurniawan et al. [35] proposed a knowledge graph-based dataset, ATT&CK-KG, that includes 665 attack techniques and 14 attack tactics from the MITRE ATT&CK framework. Each technique and tactic has unique attributes such as ID, name, description, platform, data source, and impact. The dataset is stored and represented in RDF format and can be integrated with other cybersecurity-related knowledge graphs to form a larger and more comprehensive cybersecurity knowledge graph. The literature [10] proposes a cybersecurity knowledge graph dataset CSKG4APT for APT organization attribution. The dataset is based on ontology, and by collecting and analyzing APT organization information in open source threat intelligence and combining threat intelligence data standards such as STIX and CYBOX, a security knowledge graph model containing APT organizations, attack activities, attack techniques, vulnerabilities, malware and other entities and their relationships is constructed. A security knowledge graph model containing APT organizations, attack activities, attack techniques, vulnerabilities, malware and other entities and their relationships. The data sources of this dataset are mainly MITRE's ATT&CK framework and Threat Group Cards, as well as other publicly available threat intelligence platforms and reports, providing cybersecurity analysts with an intelligent knowledge graph-based auxiliary platform that can help analysts quickly identify the characteristics of APT organizations through query, inference and visualization, etc. and behavior patterns, thereby improving the efficiency and accuracy of cyber attack attribution.

Li et al. [36] developed AttacKG, a method for constructing technical knowledge graphs (TKGs) from cyber threat intelligence (CTI) reports to analyze cyber-attacks. They used two datasets: the publicly available MISP CTI dataset and a private CTI Corpus collected and labeled by the authors. The datasets were preprocessed and labeled to extract attack techniques and relationships, which were then aligned with the MITRE ATT&CK framework. Hanks et al. [37] proposed a new cybersecurity entity annotation dataset for identifying and extracting cybersecurity-related entities, such as attackers, attack techniques, vulnerabilities, etc., from cyber threat intelligence (CTI) texts. This dataset is an unstructured CTI corpus collected from multiple open sources and contains texts of different types and styles (Table 6.1).

The Cybersecurity Knowledge Graph dataset is a very valuable resource that provides a common, scalable, and reusable cybersecurity knowledge base for academia and industry. By modeling and constructing the knowledge graph, we can better understand and master the knowledge and technologies in the field of cybersecurity, thus improving our ability to identify and respond to cybersecurity issues. In this paper, we introduce two types of general-purpose and professional cybersecurity knowledge graphs, and their establishment will provide valuable support and help for researchers and practitioners in different fields. It is believed that the release and use of these datasets will greatly contribute to the development and progress of the cybersecurity field.

Table 6.1 Cybersecurity knowledge graph dataset

Year	Knowledge graph	References	Data sources	Purpose
2019	SEPSES CKB	[26]	CVE, CWE, CAPEC, CPE, CVSS	Security event prediction
2020	CSKG	[27]	CVE, CWE, CAPEC	Cybersecurity knowledge graph
2019	CWE-KG	[31]	CWE, CAPEC, Twitter data	Twitter data analysis
2021	Open-CyKG	[33]	APT reports, CTI reports	Open cyber threat intelligence knowledge graph
2021	MalKG	[34]	CVE, Malware reports	Malware threat intelligence
2021	ATT&CK-KG	[35]	ATT&CK	Network security event detection and analysis
2022	AttacKG	[36]	AlienVault OTX, Emerging threats, and CTI reports et al	Cyber attack
2022	CSKG4APT	[10]	STIX, CYBOX	Cyber attack attribution

6.4 Cybersecurity Knowledge Graph Construction Techniques

6.4.1 Technical Architecture of CyberSecurity Knowledge Graph

The construction process of a cybersecurity knowledge graph follows a common framework similar to that of a general knowledge graph. A top-down construction model is typically adopted due to the relative maturity and completeness of the cybersecurity domain's knowledge system [38]. In this model, existing cybersecurity knowledge maps are combined to link fragmented knowledge. Information extraction and fusion techniques are used to separate entities and relationships from original data, which are then connected into the knowledge graph representation under the guidance of the ontology framework. Knowledge inference techniques are applied to generate new knowledge based on the existing knowledge graphs to support prediction and inference tasks. The resulting cybersecurity knowledge graph can be used in several application scenarios, including cyberspace situational awareness, attack path analysis, risk assessment, and anomaly detection. Figure 6.2 shows the construction framework of a cybersecurity knowledge graph.

The focus of this chapter is to provide a comparative analysis of the current research status of key technologies in the field of cybersecurity knowledge graphs, including ontology modeling, knowledge extraction, and knowledge inference.

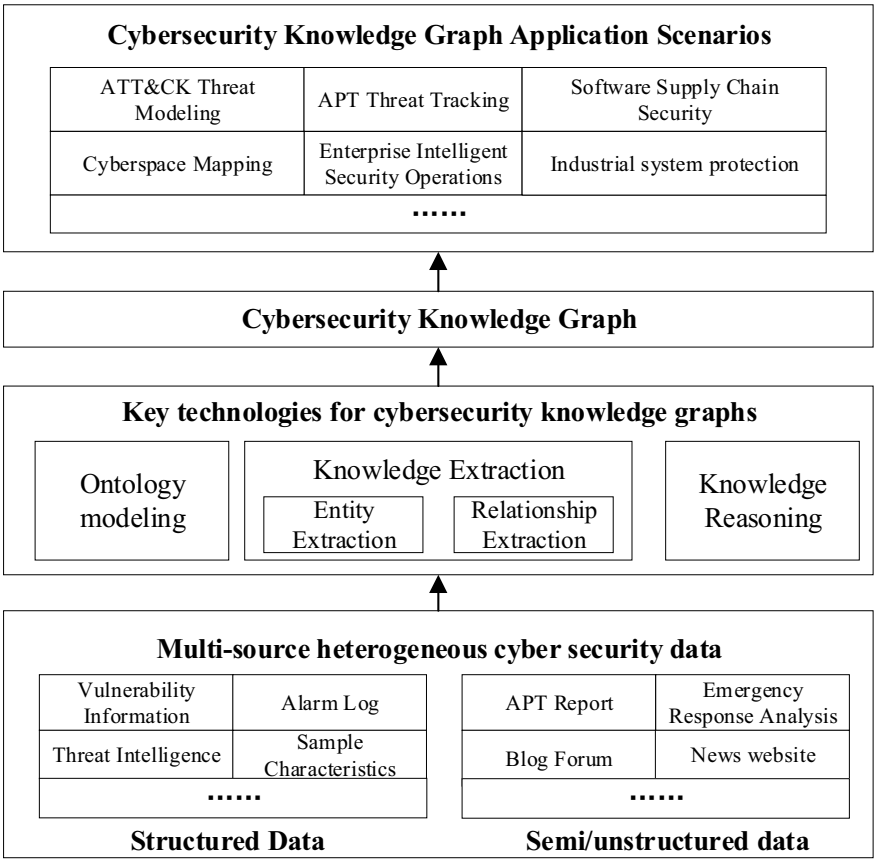

Fig. 6.2 Framework for constructing cybersecurity knowledge graph

6.4.2 Network Security Domain Ontology Modeling

Ontology Definition

Ontologies are the schema layer of knowledge graphs, derived from the philosophical domain [39], which philosophically refers to the discipline of inquiring into the nature of things in the world, and in the fields of computer science and information, science refers to the representation of categories, properties and relationships among concepts, data and entities. Ontologies are commonly used to organize data from different domains into information and knowledge, in order to reduce complexity, improve knowledge sharing, and promote reuse. Ontologies can be categorized into three types based on their scope of application: top-level ontology, domain ontology, and hybrid ontology. A top-level ontology represents generic domain concepts and relationships that are suitable for all domains. Domain ontology represents domain-specific concepts and relationships that are restricted to a particular domain. Hybrid

ontology lies between top-level ontology and domain ontology. This chapter focuses on cyber security domain ontology construction. The modeling meta-language of ontology includes five elements: class (or concept), instance, relationship, function and axiom, which are the basic elements to compose the ontology model.

The ontology model is usually represented by a five-tuple (C, I, R, F, A), where C denotes the set of classes or concepts, I denotes the set of class instances, R denotes the set of relationships between classes and classes, F denotes the set of special relationships of function classes, and A denotes the set of axioms that constrain classes and relationships. The set of relations R contains four basic types: whole and local relations (part-of), parent and subclass relations (kind-of), class and instance relations (instance-of), and attribute relations (attribute-of), which can be customized according to the specific domain relations.

Ontology construction methods can be classified into manual construction, semi-automatic construction, and automatic construction [40]. Manual construction of domain ontology relies on the knowledge of domain experts but lacks standardization and evaluation criteria for results. The semi-automatic approach reuses existing ontology libraries for extension, which can reduce the cost of ontology construction, but can also result in ontology conflict problems. The automatic construction approach is challenging because of the difficulty in handling noisy data, which makes it challenging to ensure the quality of the ontology. There are six more mature ontology construction methods, including IDEF5, skeleton method, TOVE Methodology, Methodology method, seven-step method and cyclic acquisition method. Based on the W3C standard specification, the ontology description language can be divided into RDF [41], RDFS [42] and OWL [43]. When building the domain ontology, the domain characteristics and the corresponding description language should be fully considered.

Cybersecurity Domain Ontology

In the field of cyber security, ontologies are widely researched, but there is no unified security ontology for reference [44]. Current research mainly focuses on specific areas of security, such as malware classification [45], threat intelligence analysis [46], etc. These research results provide the basis for the construction of network security knowledge graphs. At present, ontologies are mainly constructed by manual editing, because the data types involved are small and manual editing is more efficient [47].

According to the level and granularity of ontologies, this paper divides cybersecurity ontologies into three levels: (1) The highest level, unified security ontology, describes the most fundamental and core concepts and relationships in the cybersecurity domain, such as security policies, security events, security threats, etc. Ontologies at this level can provide an overall security knowledge framework and provide the basis for higher-level ontologies. (2) Intermediate level, including intrusion detection ontology, malware classification and behavior modeling ontology, vulnerability analysis ontology, etc. These ontologies describe instances or events with high generality and importance, such as different types of malware, common intrusion detection methods, and vulnerability analysis tools. Ontologies at this level can provide support

for specific security application scenarios. (3) The lowest level of the cybersecurity ontology includes several sub-ontologies, such as cyber threat intelligence analysis, cyber attack analysis, threat and security assessment, and threat actor analysis, etc. These ontologies describe situations or behaviors in network security with high detail and specificity, such as specific attack methods, attacker behavior patterns, threat intelligence analysis methods, etc. Ontologies at this level can provide more detailed knowledge support for cybersecurity professionals.

Each of these three levels is described below to provide content experience for the construction of security ontologies.

The highest level ontology in cyber security

The highest-level ontology in the network security domain includes the most fundamental and core concepts and relationships in the network security domain, such as security policy, authentication, access control, encryption and decryption, etc. The construction of the unified security ontology requires deep excavation of the nature and purpose of network security, analysis and abstraction of various security mechanisms and security protocols, as well as modeling and prediction of network attacks and threats.

Iannacone et al. proposed STUCCO [12], an ontology for cybersecurity knowledge graph databases. STUCCO is designed to integrate multiple structured and unstructured data sources and contains concepts, relationships, and rules related to cybersecurity. It mainly includes the following: (1) defines the core concepts in the cyber security domain, such as attacker, target, event, behavior, tool, etc.; (2) defines the relationships among these concepts, such as belong, cause, use, etc.; (3) defines the attributes of these concepts and relationships, such as name, time, type, etc.; (4) uses OWL language to represent the ontology and uses SPARQL query language for data retrieval. Such ontologies can help to extract meaningful cybersecurity information from data of different sources and formats, and to integrate and analyze them. Meanwhile, the article introduces the design method and implementation process of the ontology, and how to use the ontology for knowledge graph construction, querying and reasoning. In a practical case, the application effect of the ontology is demonstrated and how to use the knowledge graph to support cybersecurity analysis and decision-making is shown.

Syed et al. proposed the Unified Cybersecurity Ontology (UCO) to facilitate information integration and cyber situational awareness in cybersecurity systems [48]. The UCO combines and integrates heterogeneous data and knowledge models from different cybersecurity systems and standards, including CAPEC, CVE, CWE, STIX, TAXII [49], and Att&ck [50]. This integration makes UCO a comprehensive and practical cybersecurity ontology for information sharing and exchange. Although STIX was designed with the integration of other framework standards in mind, STIX is mainly oriented to threat intelligence information and does not cover some data representations with low information content, while its XML format representation of information is not conducive to automatic information inference. The UCO ontology, through the study of existing threat intelligence standards and ontologies, merges multiple standards into a unified standard by merging similar categories and

parent class abstraction, and covers the data ontologies represented by the current mainstream standards. operations. While many of the ontologies discussed in the literature are not publicly available, the Unified Cybersecurity Ontology (UCO) provides downloads at [51], including some example instances from industry standard repositories. However, the instance data in the dumps are not complete or updated, and there are no available public endpoints. Another example of an RDF ontology available for download is the Cyber Intelligence Ontology [52], which provides classes, properties, and many industry-standard constraints, but no instance data.

Intermediate level ontology in cyber security

Intermediate-level ontologies in the network security domain, i.e., instantiated ontologies, describe instances or events with high generality and importance in network security, such as intrusion detection, malware classification, vulnerability analysis, etc. The construction of instantiated ontology needs to focus on common security problems in practical application scenarios, extract the features and relationships through analysis and modeling of existing instances, and build the corresponding ontology model.

The role of an intrusion detection ontology is to describe and represent the concepts, properties, relationships and rules of the intrusion detection domain, as well as the semantic connections between them. The literature [53] proposes a target-centric ontology for intrusion detection to describe the composition, state and behavior of computer systems, as well as the goals, strategies and means of attackers. The article first analyzes the needs and challenges in the field of intrusion detection, and then introduces the design principles and methods of the ontology, as well as the concepts, properties, and relationships contained in the ontology. Then, the article shows how to use ontologies for knowledge representation, reasoning, and querying for intrusion detection, and how to integrate ontologies with other information sources. Finally, the article discusses the strengths and limitations of ontologies and points out future research directions. The shortcomings of this article are that the ontology may not be complete and general enough, and needs to be extended and updated for different computer systems and attack types.

The role of malware ontologies is to represent malware types, behaviors and prevention methods with semantic knowledge to improve malware identification, analysis and defense. The literature [45] focuses on an ontology-based knowledge representation approach for processing and storing complex behavioral knowledge of a large number of malware families and individuals. Using the ontology-based malware knowledge base can support a variety of research tasks, such as malware sample analysis, infection process understanding, and potential damage level assessment. In addition, the paper proposes an ontology-based reasoning approach for malware classification, which uses malware behavioral features and similarity calculations to achieve automatic categorization of unknown malware individuals. However, the specific steps and tools for extracting behavioral features from malware samples and building an ontology knowledge base are not detailed. The literature [54] mainly introduces an ontology model based on malware behavior,

which can describe the details of the malware infection process, attack target, execution mode and impact range, as well as the association between malware and threat actors and attack techniques. This ontology model can support tasks such as malware detection, analysis, and classification to improve the understanding and identification of malware behavior. However, the paper does not explain how to extract relevant information from malware samples and populate the ontology knowledge base.

Qin et al. proposed an Automated Analysis and Reasoning Model (AARV) based on a vulnerability knowledge graph [55] for vulnerability analysis. AARV consists of three components: knowledge graph construction, vulnerability description analysis, and security domain knowledge graph inference. The model can extract and store vulnerability knowledge from several widely used vulnerability databases, process and analyze the latest vulnerability descriptions using natural language processing techniques, and perform security domain knowledge graph inference using graph database query language. Based on the model, security personnel can quickly locate vulnerabilities, assess their impact and make remediation recommendations to improve cybersecurity defenses.

Lowest level ontology in cyber security

The lowest-level ontology in cybersecurity describes situations or behaviors in cybersecurity with high detail and specificity, such as cyber threat intelligence analysis, cyber attack analysis, threat and security assessment, and threatener analysis. The construction of event ontology requires an in-depth study of the details of various events in cyber security, and analysis and modeling of the source, purpose, means, and characteristics of events to better understand and predict the occurrence and evolution of cyber security events.

To provide a unified representation of multi-source heterogeneous cyber threat intelligence, Philpot et al. [56] introduce a Cyber Intelligence Ontology (CIO) to support the collection, analysis, and sharing of cyber threat intelligence. The ontology employs the OWL language for formal representation and the SPARQL language for querying and reasoning. The ontology provides an open-source knowledge model for different application scenarios and requirements. The advantage of the article is that it provides a generic web intelligence ontology. The literature [57] proposes an ontology-based information security assessment model that can describe knowledge about the structure, functionality, vulnerabilities and threats of information systems. The authors construct an information security knowledge graph based on the ontology model to store and manage information security-related data and rules. And based on the knowledge graph, a rule-based inference engine is designed to automate and automate the information security assessment. Recently, Gao et al. [46] proposed an ontology-based technique for cyber threat intelligence analysis, which uses a general ontology modeling approach to abstract the elements involved in cyber threat intelligence, such as entities, attributes, relationships, and rules, into ontology concepts and use the OWL language for formal representation. The technique can structure, semantically and logically process cyber threat intelligence and improve cyber security situational awareness. Its advantage is that it realizes the unified description and storage of cyber threat intelligence from different sources

and formats. The article also designs an ontology-based network threat intelligence analysis system for the collection, storage, query and visualization of network threat intelligence, and verifies the effectiveness and feasibility of the technique through experiments.

For cyber attacks, the literature [58] presents an ontology of cybersecurity attacks intended to represent and organize knowledge of relevant concepts, services, threats, vulnerabilities, and failure modes in the cybersecurity domain to support cybersecurity analysis and assessment. The strength of the paper is that it provides a framework for cybersecurity attacks based on standard literature and accepted methods that classify attacks into five dimensions: attack target, attack source, attack method, attack outcome, and attack defense. The literature [59] introduces an ontology-based security framework for detecting and defending zero-day attacks and complex attacks against web applications. The article proposes two ontology models that store information about application layer attacks and HTTP communication protocols, respectively, and uses this information for content filtering and attack identification. The article also demonstrates the effectiveness of the framework in a real-world environment, proving its outperformance over traditional security solutions. More recently, the literature [60] proposes a cybersecurity ontology to support the collection and representation of risk-related information in cyber-physical systems. This cybersecurity ontology includes concepts such as components, attributes, vulnerabilities, threats, attacks, and impacts of cyber-physical systems, and the relationships among them. It can be used to construct risk models of cyber-physical systems, analyze possible attack paths, and assess risk levels and impact levels. It also demonstrates the effectiveness and scalability of this cybersecurity ontology in practical applications through a case study of a smart grid.

The literature [61] presents a framework for creating a knowledge graph of threat actors, including the construction of a threat actor ontology and a named entity identification system. The threat actor ontology is used to describe the attributes, relationships, and categories of threat actors, as well as their associations with other cybersecurity-related entities. The named entity recognition system is used to extract cybersecurity-related entities from articles and generate knowledge graphs based on the threat actor ontology. And with a case study, it is demonstrated that the framework can help understand cybersecurity threat posture, especially information about threat actors. Similarly, the literature [62] proposes an ontology-based knowledge graph model that can accurately describe cybersecurity threats, vulnerabilities, attacks, and defenses, and can support complex reasoning and analysis.

In the domain of social engineering in cybersecurity, Wang et al. [63] define 11 core entity concepts and the relationships among them, as well as attributes describing social engineering attack scenarios and events; and demonstrates the usefulness of ontologies and knowledge graphs for understanding and analyzing social engineering attacks through knowledge graph applications to evaluate the effectiveness of ontologies. The strengths of this paper are: a systematic, standardized, and formalized ontology for the social engineering domain is proposed, filling the gap of lacking a unified conceptual framework in the field.

Summary and Discussion

The research on security ontology provides the reference for the construction of cybersecurity knowledge graphs in terms of content and methods. Although the above ontologies have proposed ontologies covering various security elements from various perspectives, they are relatively independent and do not consider the interconnection and integration with other ontology standards, which is still slightly insufficient for building a comprehensive knowledge ontology. This chapter proposes three levels of cybersecurity ontologies and points out their respective roles and meanings, as shown in Fig. 6.3. Establishing a network security ontology model can improve the description and understanding of knowledge and practices in the field of network security, leading to enhanced automation and intelligence of network security. Cybersecurity ontologies will continue to play a crucial role in various application scenarios and provide robust support for research and practice in the cybersecurity domain. However, due to the continuous changes and evolutions in the security domain, the design and updates of ontologies need to be constantly followed up, otherwise, they may lose their proper value and usefulness. In addition, although ontologies can provide formal representation and reasoning of security knowledge, specific customization, and adaptation are still needed for some practical application scenarios to suit different needs and situations. Therefore, how to improve the scalability and practicality of ontology is a problem that needs to be further explored and solved.

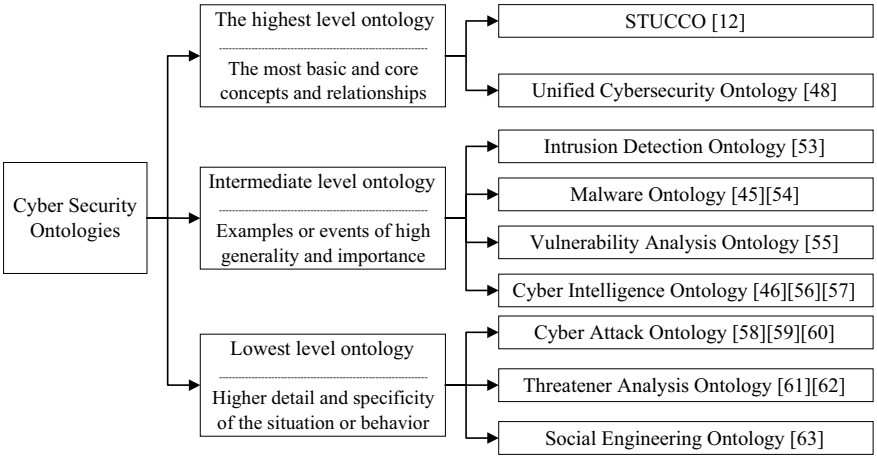

Fig. 6.3 Different levels of cybersecurity ontology

6.4.3 Knowledge Extraction for CyberSecurity Knowledge Graph

The construction of cybersecurity knowledge graphs needs to rely on knowledge extraction techniques. Currently, the core tasks of knowledge extraction include Named Entity Recognition (NER) and Relationship Extraction (RE). In the entity recognition task, traditional knowledge extraction methods can be categorized into three types: rule-based, statistical machine learning-based, and deep learning-based methods [64].

Entity extraction in the field of cybersecurity

Cybersecurity entity recognition is a type of named entity recognition (NER) specifically focused on identifying and classifying security-related entities within cybersecurity text data. These entities can include things like computers, domain names, hacker organizations, and vulnerabilities, among others. The goal of cybersecurity entity recognition is to extract and classify relevant cybersecurity vocabulary from unstructured text data, allowing for better analysis and understanding of cybersecurity threats and trends.

Rule-based approach

Most of the initial studies used rule-based approaches, whose main advantages are their high accuracy, close to the human way of thinking, intuitive presentation, and ease of machine reasoning. However, the disadvantage of the rule-based approach is that it is costly because most rules are only applicable to specific domains and cannot be extended to a wider range of domains.

In the field of intrusion detection, deep packet inspection techniques in products such as Snort, l7-filter, and Bro use a rule-based matching approach for attack type identification [65].

The paper [66] presents an improved Bootstrapping method for extracting secure entities from texts such as blogs and tweets. The traditional Bootstrapping method requires two full-text searches in one cycle, but PACE improves it by requiring only one full-text search in one cycle. First, an entity library with context is used instead of an entity library and a rule library (pattern library). When using the initial rules to extract entities, not only the entities are extracted, but also a certain number of words before and after them are selected as contextual words together to form the extraction result. Secondly, the rule generation process is changed from searching in the full text to generating in the entity database containing contextual words, thus reducing the time of searching the full text at a time. In addition, PACE relaxes the restrictions on rule generation and improves the selection of relevant contextual words to increase the recall rate and maintain a high accuracy rate.

The literature [67] proposes a combination of regular expressions and ontologies to extract entities from log files. The method first uses a support vector machine to determine whether a log file is security-relevant, then uses a separator to slice and dice paragraphs with the same format, and then uses a regular expression generated by a

genetic algorithm to tag the information in the paragraphs, and finally converts the tagged information into entities by ontology matching. The advantage of this method is that the format of semi-structured documents is used as features for type recognition and regular expression generation, and the accuracy of extraction is improved by the method of information extraction and ontology matching verification. However, this method cannot be applied to the extraction of unstructured documents. In contrast, the literature [68] proposes an approach that combines regular expressions and syntax trees to extract Indicators of Compromise (IoC) from blog text. The method first uses regular expressions and lexicons to locate web security entities and relations, and then uses grammar tree similarity to determine whether the content contains entities and relations. However, this method cannot be applied to the extraction of unstructured documents.

Statistical machine learning-based approach

Traditional machine learning-based entity extraction methods require a combination of a large number of manually designed features that are converted into a multiclassification or sequence labeling task, thus making full use of the contextual and internal features of the entities. This approach is more flexible and robust, but requires a large amount of feature engineering and manually labeled data, and also suffers from the problem of data sparsity. Several papers [69] also point out these problems.

The literature [70] proposes a weakly supervised approach based on security event extraction, but this approach relies heavily on the setting of seed samples while ignoring the entity information in the text itself.

Joshi et al. [71] developed a named entity recognition method for cybersecurity-related entities and relationships in web text data using Conditional Random Fields (CRF); Lal [72] proposed an SVM-based method for identifying cybersecurity-related entities and concepts in the unstructured text; Mulwad et al. [73] designed an SVM-based information extraction system for identifying vulnerability and attack information in web text.

The article [29] proposes an entity extraction method that uses structured data from the cybersecurity domain to automatically annotate the unstructured text. The method generates a large-scale annotated corpus by writing a script that matches entities from structured data sources with relevant text descriptions. The method then uses this corpus to train a supervised learning algorithm based on a maximum entropy model and an average perceptron to identify the desired entities from other cybersecurity documents. The method achieves a high level of accuracy, recall and efficiency. Similarly, Hanks et al. [37] present a new dataset for cybersecurity entity annotation, and the article also provides an online tool for visually and interactively annotating cybersecurity entities. This work has important implications for building artificial intelligence-based cyber defense systems and cybersecurity knowledge graphs.

Deep learning-based approach

Deep learning methods based on neural networks have shown promising results in named entity recognition and can be applied to cybersecurity as well. Compared with traditional entity recognition methods, deep learning methods are able to better

learn the feature and semantic combination capabilities of data to improve recognition accuracy by automatically discovering features and performing the potential representation and processing required for classification or detection. The current mainstream research direction is based on deep learning entity extraction methods. Deep neural networks are able to capture features and perform entity recognition automatically without excessive human intervention.

Neural networks are capable of feature extraction from data by building multilayer network structures. The Collobert [74] model based on Convolutional Neural Networks (CNN) was first used in the field of general named entity recognition and achieved good results. Since then, neural networks have been widely used for feature extraction in various domains, including cybersecurity. Hochreiter [75] proposed an LSTM model for filtering historical information using a threshold mechanism. Peng [76] uses a short text classification model based on convolutional neural network to extract features from microblogging text data and classify them by CRF, and achieves better results. Qin et al. [77] proposed a feature template-based FT-CNN-BiLSTM-CRF cybersecurity entity recognition algorithm based on a neural network model, which achieved an F-value of 86% on a large-scale cybersecurity dataset.

The main content of the literature [78] is about cybersecurity named entity recognition using LSTM recurrent neural networks. The aim of this work is to convert cybersecurity information from unstructured online sources, such as blogs and articles, into a more formal representation, which is necessary for applications in many domains. Named Entity Recognition (NER) is one of the early stages in achieving this goal, and it involves detecting relevant information.

Related research is mainly based on deep learning methods [29, 79] for threat intelligence data extraction in cyberspace. In recent years, Ranade et al. [80] proposed a contextual word embedding method CyBERT for the cybersecurity domain, which has significant performance advantages in four cybersecurity tasks, namely named entity recognition, relationship extraction, sentiment analysis and threat intelligence generation, and can capture specific semantic and syntactic information in the cybersecurity domain. Chen et al. [81] proposed a BERT model-based cybersecurity named entity identification method for extracting relevant entity information from cyber threat intelligence texts and constructing cybersecurity knowledge graphs. The authors first define and classify six types of named entities in the cyber security domain, including attackers, attack techniques, attack targets, vulnerabilities, malware, etc. Then, the authors design a joint learning framework based on the BERT model to decompose the named entity identification task into two subtasks: entity boundary detection and entity type classification.

Relationship extraction in the field of cyber security

Cybersecurity entity relationship extraction involves extracting relationships between entities in a specific cybersecurity domain. While named entity recognition can identify discrete entities, it does not capture the relationships between them. Entity relationship extraction aims to address this issue. With the advancement of information extraction and big data technologies, entity relationship extraction is increasingly used in information retrieval, relationship mining, knowledge graphs,

and other domains. Relationships between entities are an essential component of knowledge graphs, and different relationships connect distinct entities to form a knowledge graph. Identifying relationships between entities from unstructured text is a critical task in knowledge graph construction [82]. Early research on relation extraction mainly used templates to discriminate semantic relations among entities in text, but it is impossible to exhaust all templates for multiple types of relations by manual methods.

Machine learning-based approach

With the development of machine learning, more and more researchers have adopted supervised learning methods to extract relationships among entities, such as feature and kernel function-based methods for supervised learning, bootstrap, collaborative training, label propagation methods for semi-supervised learning and clustering-centered methods for unsupervised methods. The model performance of traditional machine learning is very dependent on the size and quantity of manually labeled feature data, and therefore a method that can extract features automatically is needed.

Jones et al. [83] proposed a framework for semi-supervised security entity and relationship extraction for cybersecurity concept extraction. The goal of the framework is to help security analysts access relevant information, such as new vulnerabilities, attacks, or patches, for their networks. However, annotated text data in the cybersecurity domain is scarce and expensive. Therefore, the paper follows the development of semi-supervised natural language processing and implements a bootstrapping algorithm to extract security entities and their relationships from the text. The algorithm requires only a small amount of input data, specifically, some relations or patterns (heuristic rules for identifying relations), and contains an active learning component that prompts the user for feedback on the correctness of the automatically extracted relations. The paper describes a preliminary implementation of the algorithm and applies it to cybersecurity concepts such as vulnerabilities and attacks.

Deep learning-based approach

With the development of deep learning, neural network models have brought new breakthroughs for entity relationship extraction. Liu et al. [84] proposes to classify relationships based on CNN sentence semantic coding model, which has significant performance improvement compared with traditional statistical machine learning methods; [85, 86] proposes relationship extraction based on recurrent neural network (RNN) and long short-term memory neural network (LSTM); [87] proposes to use recurrent neural network for syntactic analysis tree modeling of sentences, which takes into account the lexical and syntactic features of sentences while extracting semantic features. However, manual annotation becomes expensive in the face of large-scale data. Subsequently, scholars proposed a relation extraction method based on far-supervised learning [88], which would introduce noise to the training set. Other scholars have further proposed relational extraction methods such as multiple example learning, sentence-level attention mechanism, adversarial training, and reinforcement learning mechanism [89, 90].

Many algorithms use deep learning models for entity relationship extraction, allowing for automatic feature learning without the need for manual feature templates. While these methods achieve excellent performance on entity relationship extraction tasks, they often require large amounts of annotated data and lack annotated datasets for specific fields. To address the problem of insufficient annotated data, researchers have begun to explore remotely supervised approaches to entity relationship extraction. Zeng [91] extended the relationship extraction model based on the segmented convolutional neural network PCNN to remotely supervised data, using a remote extraction strategy based on multi-instance learning to reduce the workload of manual data labeling. However, this method's performance on relationship extraction is not very high. To address this issue, Lin et al. [92] added an attention mechanism to address noise during encoding. In 2018, a few researchers [93] found that reinforcement learning can address the noise problem during remote monitoring with good results, and this method may become a trend for future research.

For cyberspace entity relationship extraction, literature [94] extracts cyberspace knowledge from malware action reports based on Stanford extractor; literature [95] defines entity relationship extractor based on deep learning method to determine which pre-defined relationship between two entities belongs to; literature [96] proposes a CASIE system based on BERT pre-training model to achieve classification and extraction of cyber security event-related elements, and similarly, Li et al. [36] use a BERT-based sequence annotation model to identify attack techniques from each CTI report and use a GCN-based relationship classification model to predict the relationship type from each pair of neighboring techniques. The literature [26] proposes an ETL serial knowledge extraction-based approach to transforming existing public cyberspace knowledge such as CWE, CVE, CAPEC, and Common Vulnerability Scoring System (CVSS) into the triples needed for knowledge mapping.

In a recent study, Agrawal et al. [97] proposed a method for constructing knowledge graphs in the cybersecurity domain from unlabeled unstructured text and applied it to cybersecurity education. The method uses pre-trained language models and graph neural networks to extract entities and relations and uses a rule-based approach to filter and correct erroneous entities and relations. Table 6.2 summarizes and compares the methods and extraction content of representative cybersecurity knowledge extraction efforts.

6.4.4 CyberSecurity Knowledge Graph Reasoning

Regarding knowledge graph inference in cyberspace, several studies have been conducted. For example, [98] constructed a knowledge graph using MITRE's CWE knowledge base and utilized the TransE model to learn the representation of structural and textual description information in the knowledge graph, which enables inference applications such as CWE link prediction and vulnerability damage prediction. In addition, [99] developed a vulnerability knowledge graph based on the Unified Cybersecurity Ontology (UCO) ontology and implemented the inference of vulnerability

Table 6.2 Cybersecurity knowledge extraction method

Literature and year	Method	Extraction content
[65], 2011	Rule matching	Attack type identification
[66], 2013	Improved Bootstrapping method	Security entities in the blog text
[67], 2015	Regular expressions and ontologies combined	Entities in the log text
[68], 2016	Combining regular expressions and syntax trees	Missing indicators in the blog text
[70], 2015	Weak supervision method	Security events
[71], 2013	CRF	Web text data entity
[72], 2013	SVM	Entities in unstructured text
[73], 2011	SVM	Vulnerability and attack information
[29], 2013	Automatic labeling based on machine learning	Name of person, organization, etc
[37], 2022	Automatic labeling based on machine learning	Vulnerabilities, attack categories, etc
[77], 2019	Feature templates and CNN-BiLSTM-CRF	Cybersecurity entities
[78], 2018	LSTM	Security entities in blogs and articles
[80], 2021	BERT	Threat intelligence entities
[81], 2021	BERT	Threat intelligence entities
[83], 2015	Bootstrapping method	Correlation between vulnerabilities and attacks, etc
[95], 2019	Feed-forward neural networks	Malware relationships
[96], 2020	BERT	Cyber security events
[36], 2022	BERT, GCN	Attack technology relationships

hiding relationships. Moreover, [100] proposed an embedding and prediction method for software security entities and relationships. This study constructed a knowledge graph based on public knowledge bases, such as CWE, CVE, and CAPEC, and proposed a knowledge graph embedding method to embed software security entities, relationships, and description information into a continuous vector space. The study performed knowledge inference based on the open-world assumption to discover hidden relationships among software security entities.

The existing inference methods include rule-based methods, representation learning-based inference and neural network-based inference methods, etc. Graph inference mainly involves graph association retrieval, graph data mining algorithms, graph representation learning methods, relational inference, etc. Graph association retrieval provides responses to a specified entity, relationship, and attribute feature queries through the shortest path, similarity analysis, and other methods. Graph data

mining algorithms include node clustering on graphs, association discovery, significant node discovery, path mining, and so on, to provide in-depth data insight for knowledge graphs. Graph representation learning acquires vectorized representations of key elements of the knowledge graph through learning methods of structure, attributes, and other dimensions, which can be used to support knowledge retrieval, knowledge inference, and other types of technical implementations. Relational reasoning provides inference results such as knowledge semantic derivation and relational link prediction based on representation learning results or through end-to-end graph neural network model design.

Rule-based approach

Rule-based reasoning methods achieve deductive reasoning of knowledge with the help of rules, axioms and other logical forms. In security research, rule-based inference methods mainly include first-order logic rule-based methods and ontology rule-based methods [101]. Among them, the first-order logic-based approach achieves knowledge inference by constructing predicate logic formulas, which has a long research history, but the use of predicate logic formulas is narrow and the application process is more complicated. The ontology rule-based approach has been more fully researched in recent years. Rule constraints defined in OWL and SWRL or other formal languages are used to build inference relations on the basis of ontology, which has the characteristics of concise definition and rich description [102].

The literature [103] proposes a method using SQL as a rule for determining whether an access policy is misconfigured in Android. The method uses access patterns to associate specific access behaviors with explicitly defined access policies and identifies access policies that do not follow the minimization principle. Although SQL as a rule carrier can achieve simple truth-value judgment, it cannot carry complex semantics and has limited reasoning power. In contrast, SWRL is a language for Semantic Web reasoning that enables rich logical reasoning based on ontology-based OWL representations.

The literature [104] proposes a UCO ontology-based approach that uses SWRL rules to combine security content in Twitter with internal asset intelligence to generate targeted alerts. The method automatically discovers security information on social media by matching security content on the ontology and generating alerts based on SWRL rules for a specific system portrait. However, the rule defined in this method is too flat and not conducive to management. In order to effectively manage a large number of rules, rules can be defined in multiple dimensions such as time and space to improve management efficiency.

The literature [105] proposes an SWRL rule approach based on multiple processes to determine the threat risk from the time dimension. The method divides the asset risk analysis process into multiple steps, defines SWRL rules through different processes, and finally obtains asset analysis results in series. In the specific implementation, the asset analysis process is divided into four processes: element association, calculation of threat likelihood, identification of affected assets and their degree, and analysis of threat propagation paths, which has the characteristics of clear organization and rigorous logic. The literature [106] proposes a method to construct SWRL rules

based on multiple levels to identify erroneous IoT security configurations from the spatial dimension. The method checks configuration information through different levels of constraints to achieve multiple inference judgments on the correctness of configuration information and detect configurations with risks. In the specific implementation, the configuration constraints are divided into two aspects: foundational constraints and user-driven constraints, and SWRL rules are constructed at different levels. Foundational constraints include constraints for internal information such as reachability, sampling, resources, etc.; user-driven constraints include constraints for external information such as capabilities and conditions, and the rules can be extended according to specific attack behaviors.

Recently, Yi et al. [107] used knowledge graph and rule-based inference techniques to model and infer entities, attributes, relationships, and events in satellite networks to enable understanding and assessment of the security posture of satellite networks.

Representation-based learning approach

Representation learning-based inference learns a representation in vector space for each element by mapping the elements of the knowledge graph containing entities and relationships into a continuous vector space, where the representation in vector space can be one or more vectors or matrices. Representation learning allows the algorithm to automatically capture the information required for inference in the process of learning vector representations, and encodes the information of discrete symbolic representations in the knowledge graph in different vector space representations through training learning, enabling the inference of the knowledge graph to be automatically implemented through the computation between predefined vector space representations, without the need for a displayed inference step. Relational inference based on knowledge graph representation learning consists of knowledge graph representation learning and potential relationship prediction. The commonly used representation learning methods for relational inference are TransE (Translating Embedding) family of algorithms, RESCAL, DistMult, etc.

Bordes et al. [108] proposed TransE, the first transfer-based representation model for knowledge graph representation learning. The main idea of TransE is that the sum of the head entity vector and the relation vector is similar to the tail entity vector if the triple (h, r, t) holds, otherwise, it is far away. This basic transfer assumption is the foundation of subsequent research works. The score function is obtained based on this assumption, measuring the distance in terms of L1 or L2 parametrization. During the learning process, negative examples are obtained by replacing the head or tail entity, and a Margin-based loss is minimized so that the score of positive examples is at least one Margin higher than that of negative examples, similar to support vector machines. The candidate entity/relationship with the larger value of the score function is the inference result during inference. However, TransE has some limitations, such as not considering rich semantic information, lack of further adjustment of vector distribution positions in space, and not considering rich relationships, which limit its application in security scenarios.

The knowledge graph representation-based learning method is a simple and efficient relational reasoning method due to its effective and reasonable vector space assumption. For the problem of the security knowledge graph, knowledge graph representation-based learning methods have been applied in security knowledge complementation and attack path investigation. In a recent study, Wang et al. proposed a cybersecurity knowledge graph complementation method CSEA based on integrated learning and adversarial training, which integrates multiple projection and rotation operations to model the relationships between entities and uses angular information to distinguish entities. The method achieves better results than existing methods on knowledge graphs in cybersecurity and is robust to noisy data.

Neural network-based approach

Graph representation learning enables single-step relational reasoning, which can be applied to investigate individual attack events and fix broken chains. However, complex multi-step attacks require multi-step relational inference. The PRA algorithm (path ranking algorithm) is a classic multi-step relational inference algorithm that uses paths as features to predict the existence of specified relationships between entities. PATH-RNN [109] is an example of a PRA-based algorithm for multi-step relational inference. In PATH-RNN, the input is a path between two entities, and the output is the inferred new relationship between the two. The connections between relationships are represented by RNN for inference, and the representation of the path is given by the final hidden state of the RNN after processing all the relations in the path.

Ren et al. [10] proposed a deep learning-based APT knowledge graph inference method for extracting feature vectors of APT organizations from the knowledge graph and using cosine similarity and clustering algorithms to calculate the similarity and attribution relationships among APT organizations. This method can effectively utilize the structured and unstructured information in the knowledge graph to improve the accuracy and interpretability of APT attribution analysis.

Yi et al. [107] proposed a knowledge graph inference method combining rules and neural networks, which is applicable to satellite network anomaly detection and threat assessment. First, a knowledge graph of the satellite network is constructed, including entity types, attribute types and relationship types, as well as the association relationships among entities. Secondly, anomaly detection rules and threat assessment rules are defined and trained and optimized based on historical data, and finally, a neural network model is designed. The model can transform the rules into trainable parameters and use historical data for training optimization. Finally, the trained neural network model is used to reason about the new input satellite network data and output anomaly detection results, threat assessment results and response suggestion results. This approach can help the satellite network detect anomalies and take effective response measures in a timely manner.

In recent years, graph neural networks have attracted a lot of attention, and Garrido et al. [110] proposed a method for context-aware security monitoring using knowledge graphs to represent data and background knowledge related to network security and combining graph neural networks and anomaly detection algorithms. The method

is effective in detecting potential attacks and reducing false positives and misses. Yin et al. [111] proposed a knowledge graph inference method for discovering software vulnerability co-exploitation behavior, which is a type of cyber attack that exploits multiple vulnerabilities at the same time. The method uses a graph neural network model to learn embedding representations of vulnerability entities and relationships from the cybersecurity knowledge graph and then predicts co-exploitation links between vulnerabilities based on these embedding representations. The method also combines attention mechanisms and graph convolutional networks to enhance the learning process. The article claims that the method can achieve better performance than existing methods and provide interpretable results for cybersecurity intelligence.

6.5 Application Scenarios of Network Security Knowledge Graph

Currently, knowledge graph has been widely adopted in many fields, including big data analysis and natural language processing. With the development of knowledge graph technology, there are also more and more applications of network security knowledge graph in various scenarios, such as threat detection, intelligent security decision-making, vulnerability management, and attack path analysis. In this paper, the application scenarios are classified from the perspective of application objectives into the following categories.

6.5.1 CyberSecurity Knowledge Graph Defense Class Application Scenarios

The defense class application scenario aims to improve network defense capability by using network security knowledge mapping. Among them, situational awareness and security assessment can realize visual display and analysis of multi-dimensional data such as network environment, threat intelligence, and attack events to assess network security risks and threats by constructing a network security situational knowledge graph. Vulnerability management and prediction, on the other hand, can collect, organize, analyze and push vulnerability information, predict potential vulnerability exploitation and attack methods, and formulate effective protective measures by constructing a vulnerability knowledge map. Intrusion detection, on the other hand, realizes the description, representation and reasoning of intrusion behavior by constructing an intrusion detection knowledge graph, and improves the accuracy and efficiency of intrusion detection.

The use of knowledge graph-based network security situational awareness models has various practical applications. For example, Chen et al. [112] proposed a KG-based attack situational detection scheme to detect network security threats by

abstracting attack events, which improves the accuracy of network security situational detection. Similarly, Wang et al. [113] proposed a KG-NSSA model that utilizes similarity estimation and attribute graph mining methods to effectively reflect network attack scenarios in the case of asset nodes. Pang et al. [114] developed a KG-based security assessment method for power IoT terminals based on the specific application scenarios and security threat characteristics of power IoT terminals. Moreover, Chen et al. [115] generated extended attack graphs to obtain the maximum probability vulnerability paths, providing insights into attack success rates and losses.

Garrido et al. [110] propose a knowledge graph-based machine learning approach for context-aware security monitoring. The approach uses knowledge graphs to represent entities and relationships in cyberspace, as well as attackers' behavior patterns and uses graph neural networks and anomaly detection algorithms to generate meaningful and interpretable security alerts. Recently, Yin et al. [111] proposed a model based on graph neural networks and attention mechanisms that can effectively capture semantic and structural information among vulnerabilities to predict potential co-option behaviors. The model is also capable of generating interpretable co-exploitation paths to help security experts understand and prevent cyber attacks. Li et al. [116] proposed a knowledge graph-based approach for automated cyber threat intelligence analysis (K-CTIAA) that can extract threat behaviors by parsing the semantic information of cybersecurity terms. They introduced a visibility matrix and modified the formula for self-attention to reduce the negative impact of knowledge insertion, i.e., the knowledge noise problem. They also used the mapping relationship between ATT&CK and D3fend (cybersecurity knowledge graph) to provide countermeasures for the extracted threat behaviors, which can help security experts respond quickly to upcoming threats.

6.5.2 CyberSecurity Knowledge Graph Attack Class Application Scenarios

The attack class application scenarios aim to improve attack capabilities or simulate attack behaviors by using network security knowledge graphs. Among them, attack investigation can realize the tracing, analysis and attribution of attack events and reveal the identity, motive and target of attackers by constructing the attack investigation knowledge graph. Attack prediction is to achieve the prediction and early warning of possible future attacks by constructing the attack prediction knowledge map, so as to prepare for prevention or countermeasures in advance. Attack strategy generation is to realize deep analysis and mining of target systems or organizations to generate effective and covert attack strategies by building an attack strategy generation knowledge map.

For different types of attacks, Sun et al. proposed a 0-day attack path prediction method based on network defense KG, which can accurately predict the potential attack paths of 0-day attacks by using graph neural networks and knowledge graph

embeddings [117]. For distributed DDoS attacks, Liu et al. constructed a malicious behavior knowledge base, which includes a malicious traffic detection database and a network security knowledge base [118]. The authors of [119] proposed a method for generating optimal penetration paths that consider both insider and unknown attacks. The method utilizes a two-layer threat penetration graph (TLTPG), consisting of a host threat penetration graph (HTPG) in the lower layer and a network threat penetration graph (NTPG) in the upper layer. This approach can effectively generate optimal attack paths and improve the efficiency of security response. The method uses genetic algorithms and heuristic search algorithms to find the optimal host path in the HTPG and the optimal network path in the NTPG and combines the two into an optimal penetration path.

Kurniawan et al. [120] proposed a knowledge graph-based framework for discovering and analyzing the tactical behavior of cyber attacks from audit data. The authors used knowledge graph techniques to formally represent concepts, relationships, and rules in the cybersecurity domain and combined the ATT&CK knowledge base and security event log data to construct a cyber attack knowledge graph (ATT&CK-KG). The authors design a graph pattern matching-based algorithm for detecting subgraphs associated with known attack tactics from ATT&CK-KG and calculate the confidence level of tactical behavior based on the node and edge attributes in the subgraphs. The authors evaluate the framework on a real cybersecurity audit dataset and compare it with other approaches. The results show that the framework has high accuracy, efficiency, and interpretability in discovering and analyzing the tactical behavior of cyber attacks.

6.5.3 CyberSecurity Knowledge Graph Optimization Class Application Scenarios

Using cybersecurity knowledge graphs, optimization of cybersecurity operations or decision processes can be achieved. Among them, the optimization categories include intelligent operation, security policy verification, and result prediction. By constructing intelligent operation knowledge graphs, resources such as data, tools, and tasks in the process of network security operations can be managed and scheduled, thus improving the efficiency and quality of operations. By constructing the security policy validation knowledge map, the validation and evaluation of existing or newly formulated security policies in terms of legitimacy, effectiveness, and consistency can be verified and evaluated. By constructing the result prediction knowledge map, it realizes the simulation and prediction of possible results under different scenarios and gives corresponding optimization suggestions.

Yi et al. [107] used a domain knowledge-based reasoning approach to achieve automatic correlation analysis of multi-source intelligence to understand the status of satellite networks. The authors analyze the needs and challenges of satellite network

situational awareness, design a framework for satellite network situational awareness based on knowledge graphs, and give a concrete application case to demonstrate the advantages of the approach in improving satellite network security defense capabilities.

Gao et al. [121] proposed ThreatRaptor, a log-based cyber threat-hunting system that leverages external threat knowledge from OSCTI. It features an unsupervised natural language processing method that extracts structured threat behaviors from unstructured OSCTI text. ThreatRaptor also includes a domain-specific query language TBQL for hunting malicious system activities, a query synthesis mechanism that automatically generates TBQL queries, and an efficient query execution engine for searching large-scale audit log data.

Existing intrusion detection methods often suffer from generating an excessive number or poor quality of alerts. To address these issues, Garrido et al. applied machine learning to knowledge graphs to detect unexpected activities in industrial automation systems integrating IT and OT elements [110]. In the area of cyber intelligence support, Wang et al. propose a method to build a cyber attack KG based on CAPCE and CWE and implement it in the graph database Neo4j and present the directional aspects of KG in the area of secure operations, challenges faced by intelligent operations, and technical prospects [122, 123].

The cybersecurity knowledge graph also acts in logical analysis of security policies, filtering false information: Vassilev et al. propose a four-layer framework and use it for validation of the most common security threat scenarios in digital banking and implement a prototype event-driven engine for intelligent graph navigation [124]. Mitra et al. proposed a system that captures and integrates source information with Cyber Threat Intelligence (CTI) by enhancing the existing Cyber Security Knowledge Graph (CSKG) model. The system incorporates an information source graph with CSKG to improve its inference capability, enforcing rules that preserve trusted information and discard the rest [125]. Table 6.3 summarizes the latest applications of knowledge graphs in security.

6.6 Challenges and Future Trends

6.6.1 Challenges to the Cybersecurity Knowledge Graph

Cybersecurity knowledge graph is a technology for representing and managing knowledge in the cybersecurity domain, which can improve the intelligence, automation and visualization of cybersecurity. However, there are some challenges and problems with cybersecurity knowledge graphs, mainly in the following aspects.

Cybersecurity knowledge acquisition and representation

How to extract, integrate and standardize cybersecurity knowledge from multiple sources of heterogeneous data, how to select appropriate models and methods to

Table 6.3 Cybersecurity knowledge application scenarios

Year	Ref.	Application purpose
2020	[112]	Knowledge graph-based attack posture detection
2020	[113]	Network security situational awareness model (KG-NSSA)
2021	[114]	Power IoT terminal security assessment
2021	[110]	Security monitoring
2022	[111]	Preventing cyber attacks
2023	[116]	Automated cyber threat intelligence analysis (K-CTIAA)
2022	[117]	0day attack path prediction
2019	[119]	Generate optimal penetration paths
2022	[120]	Analyzing the tactical behavior of cyber attacks
2022	[107]	Satellite network security defense
2021	[121]	Log-based network threat hunting
2021	[110]	Detection of unexpected activities in industrial automation systems
2021	[123]	DDos flood attacks and multi-stage attacks
2021	[124]	Validation of the most common security threat scenarios in digital banking
2021	[125]	Filtering Cybersecurity Intelligence

represent cybersecurity knowledge, and how to deal with incomplete, inconsistent, uncertain and dynamically changing cybersecurity knowledge. Maintaining data sources related to cybersecurity research is a prerequisite for achieving efficient access to information. Due to the specialized nature of the security field, it is important for the information sources to both cover a wide range of security information and reduce the presence of security irrelevant information.

Construction and update of cybersecurity knowledge graph

How to construct a cybersecurity knowledge graph quickly, accurately and effectively, how to realize real-time updates and maintenance of cybersecurity knowledge graph, how to solve the problems of scale, complexity, and openness of cybersecurity knowledge graph and how to better improve the quality and trustworthiness of the knowledge graph. Recently, there has been related research in improving the quality of cybersecurity knowledge graphs, such as the problem of co-referential disambiguation of cybersecurity entities [126].

Reasoning and analysis of cybersecurity knowledge graph

How to use cybersecurity knowledge graph for deep-level semantic reasoning and analysis, how to support multiple types of query, retrieval, matching and recommendation functions, and how to improve the reasoning efficiency and accuracy of cybersecurity knowledge graph.

Evaluation and application of cybersecurity knowledge graph

How to evaluate the quality, reliability, validity and impact of cybersecurity knowledge graph, how to apply cybersecurity knowledge graph to various scenarios, such as threat intelligence analysis, attack traceability, risk assessment, etc., and how to improve user experience and satisfaction.

6.6.2 Future Trends of Knowledge Graph in Cybersecurity

Knowledge graph is a knowledge representation and management method based on Semantic Web technology, which can integrate structured and unstructured data into a unified, queryable, reasonable and visualized knowledge network. Knowledge graphs have a wide application prospect in the field of cybersecurity and can improve the efficiency and effectiveness of cybersecurity analysis and decision-making. This paper summarizes the development trend of the knowledge graph in the field of cybersecurity in the following aspects.

Knowledge graphs will be combined with other artificial intelligence techniques, such as machine learning, natural language processing, and computer vision, to enable deep mining and intelligent analysis of cybersecurity data. For example, Sleeman et al. [127] use more advanced graph neural network (GNN) methods to extract deeper information and knowledge from cybersecurity knowledge graphs to support more complex cybersecurity tasks.

Moreover, there will be a greater emphasis on interoperability and integration between different knowledge graphs, allowing for the creation of larger, more comprehensive knowledge repositories that can support more complex and sophisticated cybersecurity analysis. This will require the development of standardized data models and knowledge representation formats that can be used across different domains and contexts.

Finally, there will be a growing focus on the development of knowledge graphs that are tailored to specific cybersecurity domains, such as critical infrastructure protection, Internet of Things (IoT) security, and cloud security. This will require the integration of domain-specific knowledge and expertise into the knowledge graph, as well as the development of specialized algorithms and analytical techniques that are optimized for the specific cybersecurity domain.

6.7 Conclusion

In recent years, the research and development of cybersecurity have received wide attention from academia and industry, but its development process faces problems such as discrete data distribution, inaccurate information content, and difficulties in comprehensive intelligence analysis, and the emergence of knowledge graph

provides an effective solution to the above problems. This review comprehensively reviews the key technologies and application scenarios of cybersecurity knowledge graphs. We first outline the basic concepts of cybersecurity knowledge graphs, analyze the basic datasets required to build knowledge graphs, including general-purpose and specialized datasets, and summarize a framework for building them. Then we detail the key techniques for building cybersecurity knowledge graphs, including ontology construction, information extraction, and knowledge inference. Finally, we review the applications of knowledge graphs in cybersecurity and point out the challenges and future development trends of cybersecurity knowledge graphs.

References

1. Kaur, J., Ramkumar, K.R.: The recent trends in cyber security: a review. J. King Saud Univ.-Comput. Inform. Sci. **34**(8), 5766–5781 (2022)
2. Sani, M.: Knowledge graph on cybersecurity: a survey (2020)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 34–43 (2001)
4. Singhal, A.: Introducing the knowledge graph: things, not strings. Official Google Blog **5**(16), 3 (2012)
5. Chen, X., Jia, S., Xiang, Y.: A review: knowledge reasoning over knowledge graph. Exp. Syst. Appl. **141**, 112948 (2020)
6. Bollacker, K., Evans, C., Paritosh, P., et al.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. ACM Press, New York, NY, USA
7. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a large ontology from Wikipedia and WordNet. J. Web Semant. **6**(3), 203–217 (2008)
8. Auer, S., Bizer, C., Kobilarov, G., et al.: Dbpedia: a nucleus for a web of open data. In: Cruz, I.F., Decker, S., Allemang, D., et al. (eds.) The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, 11–15 Nov. 2007, Proceedings, pp. 722–735. Springer, Berlin, Heidelberg
9. Liu, K., Wang, F., Ding, Z., et al.: A review of knowledge graph application scenarios in cyber security (2022). arXiv preprint arXiv:2204.04769
10. Ren, Y., Xiao, Y., Zhou, Y., et al.: CSKG4APT: a cybersecurity knowledge graph for advanced persistent threat organization attribution. IEEE Trans. Knowl. Data Eng. (2022)
11. CyberSecurity Knowledge graph. Available at https://github.com/HoloLen/CyberSecurity_Knowledge_graph
12. Iannacone, M., Bohn, S., Nakamura, G., et al.: Developing an ontology for cyber security knowledge graphs. In: Proceedings of the 10th Annual Cyber and Information Security Research Conference, pp. 1–4 (2015)
13. Noel, S., Harley, E., Tam, K.H., et al.: CyGraph: graph-based analytics and visualization for cybersecurity. In: Handbook of Statistics, vol. 35, pp. 117–167. Elsevier (2016)
14. Zhang, K., Liu, J.: Review on the application of knowledge graph in cyber security assessment. IOP Conf. Ser. Mater. Sci. Eng. **768**(5), 052103 (2020). IOP Publishing
15. Li, X., Lian, Y., Zhang, H., Huang, K.: Key technologies of cyber security knowledge graph. Frontiers Data Comput. **3**(3), 9–18 (2021)
16. Ding, Z., Liu, K., Liu, B., et al.: Survey of cyber security knowledge graph. J. Huazhong Univ. Sci. Tech. (Natural Science Edition) **49**(07), 79–91 (2021)
17. Noel, S.: A review of graph approaches to network security analytics. In: From Database to Cyber Security, pp. 300–323 (2018)
18. Dong, C., Jiang, B., Lu, Z.G., et al.: Knowledge graph for cyberspace security intelligence: a survey. J. Cyber Sec. **5**, 56–76 (2020)

19. Liu, H., Yao, W.J., Che, S., et al.: Classification and application of cyberspace surveying and mapping system. Inform. Technol. Netw. Sec. **40**(10), 16–21+28 (2021)
20. MITRE: CTI for MITRE in GitHub (2023). Available at https://github.com/mitre/cti
21. CNNVD: CNNVD list (2023). Available at https://www.cnnvd.org.cn/home/childHome
22. Kaspersky: Vulnerability (2023). Available at https://threats.kaspersky.com/en/vulnerability/
23. Verizon Security Research & Cyber Intelligence Center: The VERIS framework (2023). Available at http://veriscommunity.net/
24. TALOS: Talos threat source newsletters (2023). Available at https://talosintelligence.com
25. CyberMonitor: APT cybercriminal campaign collections (2022). Available at https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections
26. Kiesling, E., Ekelhart, A., Kurniawan, K., et al.: The SEPSES knowledge graph: an integrated resource for cybersecurity. In: International Semantic Web Conference, pp. 198–214. Springer (2019)
27. Wang, D.: CyberSecurity Knowledge graph (2020). Available at https://github.com/HoloLen/CyberSecurity_Knowledge_graph
28. Lal, R.: Information Extraction of Security related entities and concepts from unstructured text (2013)
29. Bridges, R.A., Jones, C.L., Iannacone, M.D., et al.: Automatic labeling for entity extraction in cyber security (2013). arXiv preprint arXiv:1308.4941
30. Lim, S.K., Muis, A.O., Lu, W., et al.: Malwaretextdb: a database for annotated malware articles. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1557–1567 (2017)
31. Sun, N.: CWE-knowledge-graph-based-Twitter-data-analysis-for-cybersecurity (2019). Available at https://github.com/nansunsun/CWE-Knowledge-Graph-Based-Twitter-Data-Analysis-for-Cybersecurity
32. Kim, G., Lee, C., Jo, J., et al.: Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. Int. J. Mach. Learn. Cybern. **11**(10), 2341–2355 (2020)
33. Sarhan, I., Spruit, M.: Open-cykg: an open cyber threat intelligence knowledge graph. Knowl.-Based Syst. **233**, 107524 (2021)
34. Rastogi, N., Dutta, S., Christian, R., et al.: Predicting malware threat intelligence using KGs (2021). arXiv preprint arXiv:2102.05571
35. Kurniawan, K., Ekelhart, A., Kiesling, E.: An ATT&CK-KG for linking cybersecurity attacks to adversary tactics and techniques (2021)
36. Li, Z., Zeng, J., Chen, Y., et al.: AttacKG: constructing technique knowledge graph from cyber threat intelligence reports. In: Computer Security Copenhagen, Denmark, 26–30 Sept. 2022, Proceedings, Part I. Springer International Publishing, Cham (2022)
37. Hanks, C., Maiden, M., Ranade, P., et al.: Recognizing and extracting cybersecurity entities from text. In: International Conference on Machine Learning Workshop on Machine Learning for Cybersecurity (2022)
38. Yang, Y.J., Xu, B., Hu, J.W., Tong, M.H., Zhang, P., Zheng, L.: Accurate and efficient method for constructing domain knowledge graph. Ruan Jian Xue Bao/J. Softw. **29**(10), 2931–2947 (2018)
39. Wikipedia: Ontology (2023). Available at https://en.wikipedia.org/wiki/Ontology
40. Khadir, A.C., Aliane, H., Guessoum, A.: Ontology learning: grand tour and challenges. Comput. Sci. Rev. **39**, 100339 (2021)
41. Manola, F., Miller, E., McBride, B.: RDF primer. w3c recommendation **10**(1–107), 6 (2004)
42. McBride, B.: The resource description framework (RDF) and its vocabulary description language RDFS. In: Handbook Ontologies, pp. 51–65 (2004)
43. McGuinness, D.L., Van Harmelen, F.: OWL web ontology language overview. w3C recommendation **10**(10), 2004 (2004)
44. Mavroeidis, V., Bromander, S.: Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In: Proceedings of the 2017 European Intelligence and Security Informatics Conference (EISIC), pp. 91–98. IEEE, Athens (2017)

45. Ding, Y., Wu, R., Zhang, X.: Ontology-based knowledge representation for malware individuals and families. Comput. Secur. **87**, 101574 (2019)
46. Gao, J., Wang, A.: Research on ontology-based network threat intelligence analysis technology. Comput. Eng. Appl. **56**(11), 112–117 (2020)
47. Liu, J., Li, Y., Duan, H., et al.: Knowledge graph construction techniques. J. Comput. Res. Dev. **53**(3), 582–600 (2016)
48. Syed, Z., Padia, A., Finin, T., et al.: UCO: a unified cybersecurity ontology. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (2016)
49. Cyber Threat Intelligence (2023). Available at https://oasis-open.github.io/cti-documentation/
50. MITRE ATT&CK (2023). Available at https://attack.mitre.org
51. Unified-Cybersecurity-Ontology (2019). Available at https://github.com/Ebiquity/Unified-Cybersecurity-Ontology
52. Cyber Intelligence Ontology (2015). Available at https://github.com/daedafusion/cyber-ontology
53. Jeffrey, U., John, P., Anupam, J., et al.: A target centric ontology for intrusion detection. In: The IJCAI-03 Workshop on Ontologies and Distributed Systems, pp. 47–58. IJCAI, Acapulco (2004)
54. Grégio, A., Bonacin, R., Nabuco, O., et al.: Ontology for malware behavior: a core model proposal. In: 2014 IEEE 23rd International WETICE Conference, pp. 453–458. IEEE (2014)
55. Qin, S., Chow, K.P.: Automatic analysis and reasoning based on vulnerability knowledge graph. In: Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health, pp. 3–19. Springer, Singapore (2019)
56. Philpot, M.: Cyber Intelligence Ontology (2015). https://github.com/daedafusion/cyber-ontology, 18 Oct.
57. Gao, J.B.: Research on ontology model and its application in information security evaluation. Shanghai Jiao Tong University (2015)
58. Simmonds, A., Sandilands, P., van Ekert, L.: An ontology for network security attacks. In: Applied Computing, pp. 317–323. Springer (2004)
59. Razzaq, A., Anwar, Z., Ahmad, H.F., et al.: Ontology for attack detection: an intelligent approach to web application security. Comput. Secur. **45**(S1), 124–146 (2014)
60. Grigoriadis, C., Berzovitis, A.M., Stellios, I., et al.: A cybersecurity ontology to support risk information gathering in cyber-physical systems. In: Computer Security. ESORICS 2021 International Workshops: CyberICPS, SECPRE, ADIoT, SPOSE, CPS4CIP, and CDT&SECOMANE, pp. 23–39, Darmstadt, Germany, 4–8 Oct. Springer International Publishing, Cham (2022)
61. Hooi, E.K.J., Zainal, A., Maarof, M.A., et al.: TAGraph: knowledge graph of threat actor. In: International Conference on Cybersecurity (ICoCSec), pp. 76–80. IEEE (2019)
62. Kaloroumakis, P.E., Smith, M.J.: Toward a knowledge graph of cybersecurity countermeasures, p. 11. The MITRE Corporation (2021)
63. Wang, Z., Zhu, H., Liu, P., et al.: Social engineering in cybersecurity: a domain ontology and knowledge graph application examples. Cybersecurity **4**, 1–21 (2021)
64. Li, J., Sun, A., Han, J., et al.: A survey on deep learning for named entity recognition. IEEE Trans. Knowl. Data Eng. **99**, 1–11 (2020)
65. Zhang, S.Z., Luo, H., Fang, B.X.: Regular expressions matching for network security. J. Softw. **22**(8), 1838–1854 (2011)
66. McNeil, N., Bridges, R.A., Iannacone, M.D., et al.: Pace: pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts. In: Machine Learning and Applications (ICMLA), pp. 60–65 (2013)
67. Kushner, S.: Ontology-driven data semantics discovery for CyberSecurity. In: Practical Aspects of Declarative Languages (PADL), pp. 1–16 (2015)
68. Liao, X., Yuan, K., Li, Z., et al.: Acing the IOC game: toward automatic discovery and analysis of open-source cyber threat intelligence. In: ACM Sigsac Conference on Computer and Communications Security (ACM SIGSAC), pp. 755–766 (2016)

69. Georgescu, T.M.: Natural language processing model for automatic analysis of cybersecurity-related documents. Symmetry **12**(3), 20–35 (2020)
70. Ritter, A., Wright, E., Casey, W., et al.: Weakly supervised extraction of computer security events from Twitter. In: The 24th International Conference on World Wide Web, pp. 896–905 (2015)
71. Joshi, A., Lal, R., Finin, T., et al.: Extracting cybersecurity related linked data from tex. In: 2013 IEEE Seventh International Conference on Semantic Computing, pp. 252–259 (2013)
72. Lal, R.: Information extraction of cyber security related terms and concepts from unstructured text. University of Maryland, Baltimore County (2013)
73. Mulwad, V., Li, W., Joshi, A., et al.: Extracting information about security vulnerabilities from web text. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 257–260 (2011)
74. Collobert, R., Weston, J., Bottou, L., et al.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(ARTICLE), 2493–2537 (2011)
75. Graves, A., Graves, A.: Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks, pp. 37–45 (2012)
76. Peng, N., Dredze, M.: Named entity recognition for Chinese social media with jointly trained embeddings. In: The 2015 Conference on Empirical Methods in Natural Language Processing, pp. 548–554 (2015)
77. Qin, Y., Shen, G., Zhao, W., et al.: A network security entity recognition method based on feature template and CNN-BiLSTM-CRF. Frontiers Inform. Technol. Electron. Eng. **20**(6), 872–884 (2019)
78. Gasmi, H., Bouras, A., Laval, J.: LSTM recurrent neural networks for cybersecurity named entity recognition. ICSEA **11**, 2018 (2018)
79. Yu, H., Zhang, N., Deng, S., et al.: Bridging text and knowledge with multi-prototype embedding for few-shot relational triple extraction (2020). arXiv preprint arXiv:2010.16059
80. Ranade, P., Piplai, A., Joshi, A., et al.: Cybert: contextualized embeddings for the cybersecurity domain. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 3334–3342. IEEE (2021)
81. Chen, Y.X., Ding, J., Li, D., et al.: Joint BERT model based cybersecurity named entity recognition. In: 2021 The 4th International Conference on Software Engineering and Information Management, pp. 236–242 (2021)
82. Fisher, J., Vlachos, A.: Merge and label: a novel neural network architecture for nested NER (2019). arXiv preprint arXiv:1907.00464
83. Jones, C.L., Bridges, R.A., Huffer, K.M.T., et al.: Towards a relation extraction framework for cyber-security concepts. In: the 10th Annual Cyber and Information Security Research Conference, pp. 1–4 (2015)
84. Liu, C.Y., Sun, W.B., Chao, W.H., et al.: Convolution neural network for relation extraction. In: 9th International Conference on Advanced Data Mining and Applications (ADMA), China, pp. 231–242, Hangzhou (2013)
85. Zhang, D., Wang, D.: Relation classification via recurrent neural network (2015). arXiv preprint arXiv:1508.01006
86. Zhou, P., Shi, W., Tian, J., et al.: Attention-based bidirectional long short-term memory networks for relation classification. In: The 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short papers), pp. 207–212 (2016)
87. Socher, R., Huval, B., Manning, C.D., et al.: Semantic compositionality through recursive matrix-vector spaces. In: The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1201–1211 (2012)
88. Mintz, M., Bills, S., Snow, R., et al.: Distant super-vision for relation extraction without labeled data. In: The International Joint Conference on ACL Association for Computational Linguistics, pp. 1003–1011. Association for Computational Linguistics, Singapore (2009)
89. Feng, J.: Reinforcement learning for relation classification from noisy data. In: The Thirty-Second AAAI Conference on Artificial Intelligence, pp. 5779–5786. Louisiana, New Orleans (2018)

90. Han, X., Zhu, H., Yu, P., et al.: FewRel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: The 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4803–4809. Association for Computational Linguistics, Brussels (2018)
91. Zeng, D., Liu, K., Chen, Y., et al.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: The 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1753–1762 (2015)
92. Lin, Y., Shen, S., Liu, Z., et al.: Neural relation extraction with selective attention over instances. In: The 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2124–2133 (2016)
93. Qin, P., Xu, W.Y., Wang, W.Y.: Robust distant supervision relation extraction via deep reinforcement learning (2018). arXiv preprint arXiv:1805.09927
94. Gupta, M., Abdelsalam, M., Khorsandroo, S., et al.: Security and privacy in smart farming: challenges and opportunities. IEEE Access **8**, 34564–34584 (2020)
95. Pingle, A., Piplai, A., Mittal, S., et al.: Relext: relation extraction using deep learning approaches for cybersecurity knowledge graph improvement. In: The 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 879–886 (2019)
96. Satyapanich, T., Ferraro, F., Finin, T.: Casie: extracting cybersecurity event information from text. In: The AAAI Conference on Artificial Intelligence, vol. 34(05), pp. 8749–8757 (2020)
97. Agrawal, G., Deng, Y., Park, J., et al.: Building knowledge graphs from unstructured texts: applications and impact analyses in cybersecurity education. Information **13**(11), 526 (2022)
98. Han, Z., Li, X., Liu, H., et al.: Deepweak: reasoning common software weaknesses via knowledge graph embedding. In: 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), pp. 456–466. IEEE (2018)
99. Qin, S., Chow, K.P.: Automatic analysis and reasoning based on vulnerability knowledge graph. In: Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health: International 2019 Cyberspace Congress, CyberDI and CyberLife, Beijing, China, 16–18 Dec. 2019, Proceedings, Part I 3, pp. 3–19. Springer, Singapore (2019)
100. van Gerven, M.A.J., Bohte, S.M.: Artificial neural networks as models of neural information processing. Frontiers Comput. Neurosci. (2017)
101. Saiping, G., Xiaolong, J., Yantao, J., et al.: Knowledge graph oriented knowledge inference methods: a survey. J. Softw. **29**(10), 2966–2994 (2018)
102. Yu, L., Yu, L.: OWL: web ontology language. In: A Developer's Guide to the Semantic Web, pp. 155–239 (2011)
103. Wang, R., Azab, A.M., Enck, W., et al.: Spoke: scalable knowledge collection and attack surface analysis of access control policy for security enhanced android. In: The 2017 ACM on Asia Conference on Computer and Communications Security, pp. 612–624 (2017)
104. Mittal, S., Das, P.K., Mulwad, V., et al.: Cybertwitter: using Twitter to generate alerts for cybersecurity threats and vulnerabilities. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 860–867. IEEE (2016)
105. Qamar, S., Anwar, Z., Rahman, M.A., et al.: Data-driven analytics for cyber-threat intelligence and information sharing. Comput. Secur. **67**, 35–58 (2017)
106. Mohsin, M., Anwar, Z., Zaman, F., et al.: IoTChecker: a data-driven framework for security analytics of Internet of Things configurations. Comput. Secur. **70**, 199–223 (2017)
107. Yi, J., Liu, B., Yao, L.: Satellite cyber situational understanding based on knowledge reasoning. Syst. Eng. Electron. (2022)
108. Bordes, A., Usunier, N., Garcia-Duran, A., et al.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, p. 26 (2013)
109. Das, R., Zaheer, M., Reddy, S.: Chains of reasoning over entities, relations, and text using recurrent neural networks. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 132–141 (2017)
110. Garrido, J.S., Dold, D., Frank, J.: Machine learning on knowledge graphs for context-aware security monitoring. In: 2021 IEEE International Conference on Cyber Security and Resilience (CSR), pp. 55–60. IEEE (2021)

111. Yin, J., Tang, M.J., Cao, J., et al.: Knowledge-driven cybersecurity intelligence: software vulnerability co-exploitation behaviour discovery. IEEE Trans. Ind. Inform. (2022)
112. Chen, J.: Design and implementation of network attack situation detection system based on knowledge graph. Beijing University of Posts and Telecommunications (2020)
113. Wang, Y.: Research and implementation of NSSA technology based on knowledge graph. University of Electronic Science and Technology of China (2020)
114. Pang, T., Song, Y., Shen, Q.: Research on security threat assessment for power IOT term terminal based knowledge graph. In: 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 1717–1721. IEEE (2021)
115. Chen, Z., Dong, N., Zhong, S., et al.: Research on the power network security vulnerability expansion attack graph based on knowledge map. Inform. Technol. **46**(02), 30–35 (2022)
116. Li, Z.X., Li, Y.J., Liu, Y.W., et al.: K-CTIAA: automatic analysis of cyber threat intelligence based on a knowledge graph. Symmetry **15**(2), 337 (2023)
117. Sun, C., Hu, H., Yang, Y., et al.: Prediction method of 0 day attack path based on cyber defense knowledge graph. Chin. J. Netw. Inform. Sec. **8**(01), 151–166 (2022)
118. Liu, F., Li, K., Song, F.: Distributed DDoS attacks malicious behavior knowledge base construction. Telecommun. Sci. **37**(11), 17–32 (2021). 111
119. Wang, S., Wang, J.H., Tang, G.M., et al.: Intelligent and efficient method for optimal penetration path generation. J. Comput. Res. Dev. **56**, 929–941 (2019)
120. Kurniawan, K., Ekelhart, A., Kiesling, E., et al.: KRYSTAL: knowledge graph-based framework for tactical attack discovery in audit data. Comput. Secur. **121**, 102828 (2022)
121. Gao, P., Shao, F., Liu, X., et al.: Enabling efficient cyber threat hunting with cyber threat intelligence. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 193–204. IEEE (2021)
122. NEFOCUS: Security Knowledge Graph Technology White Paper (2022). https://www.nsfocus.com.cn/html/2022/92_0105/166.html
123. Wang, W., Zhou, H., Li, K., et al.: Cyber-attack behavior knowledge graph based on CAPEC and CWE towards 6G. In: International Symposium on Mobile Internet Security, pp. 352–364. Springer (2021)
124. Vassilev, V., Sowinski-Mydlarz, V., Gasiorowski, P., et al.: Intelligence graphs for threat intelligence and security policy validation of cyber systems. In: Proceedings of International Conference on Artificial Intelligence and Applications, pp. 125–139. Springer (2021)
125. Mitra, S., Piplai, A., Mittal, S., et al.: Combating fake cyber threat intelligence using provenance in cybersecurity knowledge graphs. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 3316–3323. IEEE (2021)
126. Liu, Z., Su, H., Wang, N., et al.: Coreference resolution for cybersecurity entity: towards explicit, comprehensive cybersecurity knowledge graph with low redundancy. In: 18th EAI International Conference on Security and Privacy in Communication Networks (SecureComm 2022), pp. 89–108. virtual Event, October 2022, Proceedings. Springer Nature Switzerland, Cham
127. Sleeman, J., Finin, T., Halem, M.: Understanding cybersecurity threat trends through dynamic topic modeling. Frontiers Big Data **4**, 601529 (2021)