Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8883696, 13 pages
https://doi.org/10.1155/2020/8883696

Research Article

Data-Driven Cybersecurity Knowledge Graph Construction for Industrial Control System Security

Guowei Shen,[1,2,3] **Wanling Wang**,[1] **Qilin Mu**,[2,3] **Yanhong Pu**,[2,3] **Ya Qin**,[1] **and Miao Yu**[4]

[1]*Guizhou Provincial Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China*
[2]*Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, Guiyang 550022, China*
[3]*CETC Big Data Research Institute Co., Ltd., Guiyang 550022, China*
[4]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China*

Correspondence should be addressed to Miao Yu; yumiao@iie.ac.cn

Received 14 July 2020; Revised 22 September 2020; Accepted 31 October 2020; Published 28 December 2020

Academic Editor: Ding Wang

Industrial control systems (ICS) involve many key industries, which once attacked will cause heavy losses. However, traditional passive defense methods of cybersecurity have difficulty effectively dealing with increasingly complex threats; a knowledge graph is a new idea to analyze and process data in cybersecurity analysis. We propose a novel overall framework of data-driven industrial control network security defense, which integrated fragmented multisource threat data with an industrial network layout by a cybersecurity knowledge graph. In order to better correlate data to construct a knowledge graph, we propose a distant supervised relation extraction model ResPCNN-ATT; it is based on a deep residual convolutional neural network and attention mechanism, reduces the influence of noisy data in distant supervision, and better extracts deep semantic features in sentences by using deep residuals. We empirically demonstrate the performance of the proposed method in the field of general cybersecurity by using dataset CSER; the model proposed in this paper achieves higher accuracy than other models. And then, the dataset ICSER was used to construct a cybersecurity knowledge graph (CSKG) on the basis of analyzing specific industrial control scenarios, visualizing the knowledge graph for further security analysis to the industrial control system.

1. Introduction

Industrial control systems (ICS), which involve key industries such as oil and gas production, electricity, chemical processing, transportation, and manufacturing, have seen increasing security problems and cyberattacks in recent years due to access to the Internet, such as Stuxnet. Stuxnet [1] infected and manipulated programmable logic controller (PLC) and caused serious physical damage to equipment which led to system failure. In 2016, the power system of Ukraine was attacked by a variant of the BlackEnergy malicious code [2], resulting in a large-scale power outage that affected 225,000 citizens. An industrial control network involves a lot of important infrastructure construction; in the event of a cyberattack, huge losses will be caused and endanger the economy, public safety, human life, and other aspects [3]. With the support of 5G technology, the industrial Internet will be integrated with the development of 5G [4], which promotes industrial development while introducing more security risks, so it is necessary to further improve the guarantee of industrial network security.

Data-driven prediction and analysis of cybersecurity incidents is a hot topic in current cybersecurity research; through mining correlations among industrial control network data, the asset equipment information of the industrial control system can be associated with corresponding vulnerabilities, to identify the potential internal and external threat relationship with fine granularity and construct the asset threat graph based on a specific industrial control network structure. It is more explicit to see threat situation

in security analysis of ICS by using visualization technology, which provides accurate support for industrial control network security protection decision-making. Currently, there are numerous open source threat intelligence sources periodically updating threat feeds fed into various analytical solutions. Security news, security forums, and vulnerability information are important data sources for cyberthreat intelligence. However, the above data is fragmented, and it is difficult to correlate such multisource data.

A cybersecurity knowledge graph (CSKG) is a powerful tool for data-driven thread intelligence computing. Researchers can intuitively know cybersecurity entities and relations between the entities through CSKG, such as utilization relation between malware and vulnerabilities, employment relation between attackers and organizations, and ownership between software and vulnerabilities. Relation extraction is a very important task in the construction of CSKG from unstructured data.

In relation extraction, the lack of labeled data for training is a challenge when constructing a network security knowledge graph. A common technique for coping with this difficulty is distant supervision in natural language processing. Distant supervision strategy is an effective method of automatically labeling training data. However, the assumption in the distant supervision method is too strong, leading to the wrong label problem.

In this paper, we first propose a novel overall framework of data-driven industrial control network security defense. In order to better mine entity relations in cybersecurity data, we propose a novel cybersecurity relation extraction model ResPCNN-ATT which combined Residual Learning, Piecewise Convolutional Neural Networks (PCNN), and multi-instance ATTention. The following list details the main contributions of the article:

(i) A novel data-driven industrial network security defense framework is proposed, which structures fragmented multisource data and integrates with industrial network layout

(ii) A distant supervised cybersecurity relation extraction model based on ResPCNN-ATT is proposed to reduce the impact of noise data in open source threat intelligence data sources

(iii) ResPCNN-ATT first uses the pretrained word vector and the position vector between cybersecurity entity pairs as the model input and then uses PCNN to extract the semantic features. Deep residual learning is used to solve the problem of gradient disappearance caused by noise data. A multi-instance attention mechanism is used to calculate the correlation between instance and the corresponding relation to reduce the impact of noise data

(iv) The datasets CSER and ICSER are constructed. We first empirically demonstrate the performance of the proposed method in the field of general cybersecurity by using dataset CSER. And then, we analyze asset information and network layout of Electric Power and Intelligent Control Testbed (EPIC) and use dataset ICSER to construct a cybersecurity knowledge graph for EPIC, visualizing the knowledge graph for further security analysis to the industrial control system

The rest of the paper is organized as follows. We describe related works in Section 2 and propose the overall framework in Section 3. The structure definition of CSKG is analyzed in Section 4. The cybersecurity relation extraction model and details are shown in Section 5, and performance evaluation of the model is discussed in Section 6. In Section 7, we construct and visualize a cybersecurity knowledge graph based on a specific industrial control scenario. Section 8 draws conclusions.

2. Related Work

Industrial control systems (ICS) consist of integrated hardware and software components for monitoring and controlling various industrial processes, often deployed in critical infrastructure such as water treatment plants, power grids, and gas pipelines [5]. In recent years, more and more components of ICS are connected to the Internet, exposing more and more security vulnerabilities that may be exploited by attackers [6]. Various vulnerabilities in Internet are important internal causes of network security risks. There are vulnerabilities in all levels and links of the information network; once exploited by malicious actors, they will affect normal operation of the system and its services [7]. Due to the increasing number of attack events and the serious consequences of attacking, and the many threats in the complex industrial network environment [8, 9], it is crucial to study industrial network security. Traditional passive defense measures of cybersecurity have the difficulty of effectively dealing with the increasingly complex threats; we must strengthen cybersecurity analysis capability based on vulnerabilities, threat intelligence, and other aspects and enhance the industrial network security active defense capability.

Structuring and organizing data can improve the efficiency and accuracy of cybersecurity analysis. Sadighian et al. [10] proposed ONTIDS, an ontology alarm association framework based on context information. By defining the ontology structure, security alarms are represented and stored, and the association between alarm information is regularized; on this basis, rules are set to filter alarms to reduce the false alarm rate and facilitate network security analysis. In order to further achieve cybersecurity information correlation and semantic analysis, many researches are devoted to improving the interpretation, feature correlation, and data processing of the alarm log, reducing the false alarm rate, and enhancing cybersecurity analysis capability [11–13].

Data-driven cybersecurity event prediction and analysis are hot topics in the current cybersecurity research [14]. Shu et al. introduced a new methodology that models threat discovery as a graph computation problem for threat intelligence [15]. Yu et al. proposed a relation extraction method for the construction of a knowledge graph in the food field [16]. As a semantic knowledge base, a knowledge

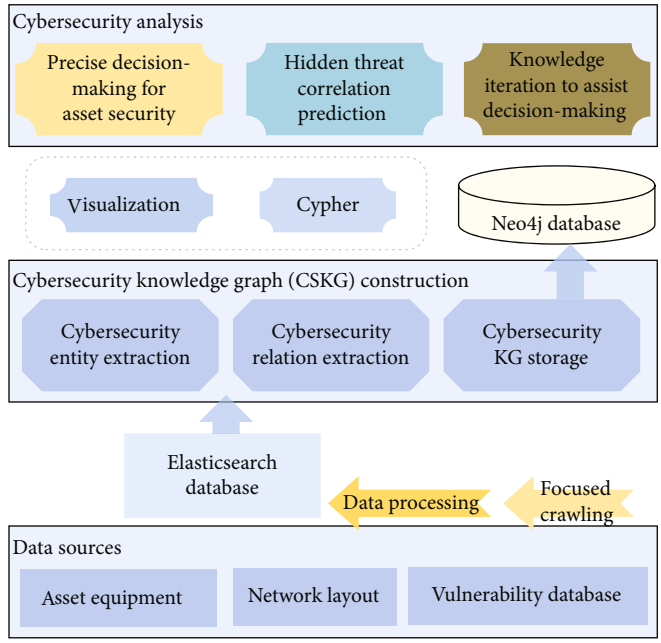

Figure 1: The overall framework of data-driven industrial control network security analysis.

graph is a powerful tool for managing large-scale knowledge consisting of entities and relations between them. Using a knowledge graph to analyze and process data provides a new idea for cybersecurity analysis, integrates open source fragmented data, identifies its correlation, associates asset equipment in ICS with corresponding vulnerability information, excavates the internal and external potential threat relation, and further conducts more accurate analysis on industrial control network security. It is crucial to mine the association of data resources efficiently and accurately.

Natural language processing technology [17–19] tends to only consider the domain name and IP address when analyzing the relation between malicious entities, both of which have very simple relation definitions. Pingle et al. proposed the RelExt [20] system, which strives to improve various cyberthreat representation schemes, especially cybersecurity knowledge graphs (CSKG), by predicting the relations between cybersecurity entities identified by cybersecurity named entity recognizer. VIEM [21] analyzed a large number of inconsistencies by extracting software names and software versions in public security vulnerability reports, so the extraction of relations is more complicated.

Relation extraction (RE) is one of the most important topics in NLP. Many relation extraction methods have been proposed [22–24], such as bootstrapping, unsupervised relation discovery, and supervised classification. Most existing supervised RE methods require a large amount of labeled relation-specific training data, which is very time-consuming and labor-intensive. Distant supervision is proposed to automatically generate training data. Under the framework of distance supervised learning, some recent work [25–28] attempts to use deep neural networks in relation prediction. Although distant supervision is an effective strategy to automatically label training data, it always suffers from the wrong label problem.

3. Overall Framework

There are numerous open source threat intelligence sources periodically updating threat feeds fed into various analytical solutions; it is significant for cybersecurity analysis that structures these data and applies them to specific scenarios. As shown in Figure 1, we propose a data-driven industrial control network security analysis framework based on a cybersecurity knowledge graph. We combine threat intelligence such as third party attack reports and vulnerability libraries with asset network layouts, and so, internal network layout and threat information corresponding to assets in networks are integrated with external threat intelligence. A knowledge graph extends the problem of cybersecurity analysis to the study of the graph structure; graph-based analysis is conducive to the development of effective system protection, detection, and response mechanisms.

We first analyze ICS scenarios to identify asset equipment and communication layout. On this basis, we mine external vulnerability information from vulnerability libraries such as Cybersecurity and Infrastructure Security Agency (CISA) (https://www.us-cert.gov/ics), National Vulnerability Database (NVD) (https://nvd.nist.gov/), Common Weakness Enumeration (CWE) (https://cwe.mitre.org/), and Common Vulnerabilities and Exposures (CVE) (http://cve.mitre.org/). We collect data by the way of focused crawling and obtain the key corpus for constructing a knowledge graph after processing. And then, we utilize cybersecurity entity identification and relation extraction technology to form a cybersecurity knowledge graph (CSKG), offering structured analysis data for specific cybersecurity scenarios. Based on the constructed CSKG, we can use visualization technology to show the connection between assets and threats clearly; it becomes easier to query entities, relations, and path. We further research on the basis of the knowledge graph,

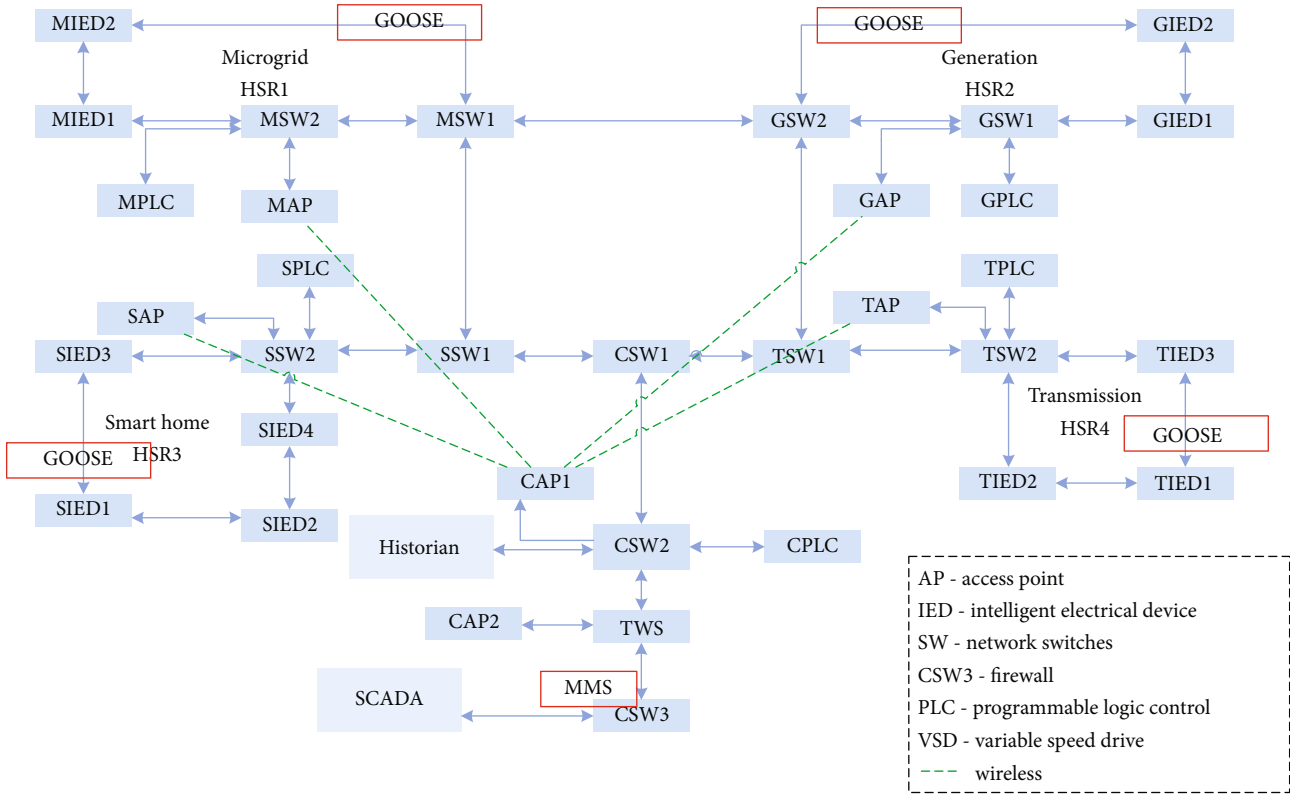

Figure 2: The communication layout of EPIC.

utilizing knowledge reasoning technology to forecast correlation of threats and assets, to more comprehensively analyze industrial control network security.

We have done a lot of research on the key technologies of the knowledge graph. Information extraction, as a key technology of CSKG, is of great significance in the entire architecture. Cybersecurity entities have the characteristics of mixed Chinese and English, confusing classification, and unclear features, and the existing related datasets are also very few, leading to difficulties in cybersecurity entity relation extraction.

For the lack of related datasets, we construct dataset CSER for general cybersecurity relation extraction and dataset ICSER for industrial control network relation extraction. First, the cybersecurity entity recognition model based on FT-CNN-BiLSTM-CRF proposed by Qin et al. [29] is used to extract cybersecurity entity pairs. This method uses artificial feature templates to extract local context features and further uses a neural network to automatically extract character features and global text features. Cybersecurity entity pairs were used to manually annotate some of the relation extraction corpora and match entity pairs with text data from vulnerability databases to form final datasets. Finally, the cybersecurity relation extraction dataset CSER and industrial control network relation extraction dataset ICSER are constructed.

4. CSKG Structure Definition

4.1. Scenario Analysis. In this paper, we take Electric Power and Intelligent Control Testbed (EPIC) from iTrust Labs (https://itrust.sutd.edu.sg/itrust-labs-home/itrust-labs_epic/) as a specific industrial control network scenario. We analyze the network layout and list the key asset equipment and resources in EPIC.

EPIC is a power testbed that maps a small smart grid system in real life, including four stages of generation, transmission, microgrid, and smart home; each stage is controlled by its own PLC/controller. There are communication channels between SCADA, distributed control system (DCS), and energy management system (EMS) and each PLC/controller. Attackers can exploit vulnerabilities to enter the communication network and maliciously manipulate the control flow and launch DDos attack on the PLC control flow, and then, the system cannot work normally. Attackers can also utilize the communication channel to enter the SCADA workstation and operate on the SMA portal to launch more attacks.

According to [30], the communication layout of EPIC is shown in Figure 2, which is composed of a SCADA workstation, historian, programmable logic controller (PLC), intelligent electrical devices (IEDs), access points (APs), and switches (SWs), and redundancy in the ring network is achieved using high availability seamless redundancy (HSR) and media redundancy protocol (MRP).

EPIC uses the IEC 61850 standard as the communication protocol for automation systems. There are two main protocols: Manufacturing Message Specification (MMS) and General Object-Oriented Substation Event (GOOSE). It allows data communication between IED, PLC, and SCADA workstations. PLC uses MMS to communicate with SCADA

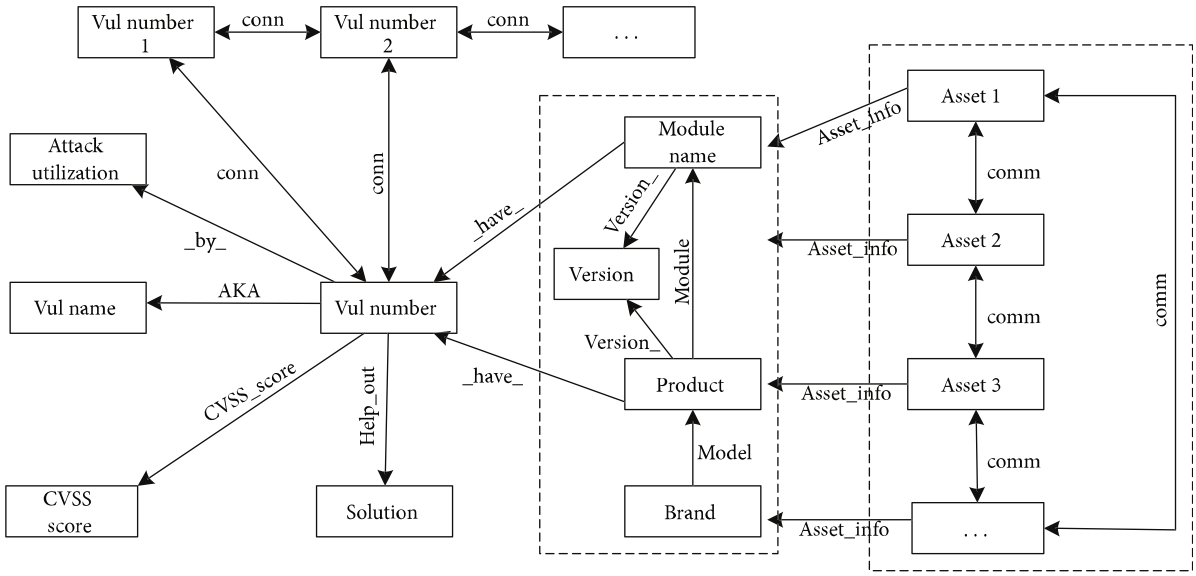

Figure 3: Ontology structure.

workstations and IEDs and communicate through GOOSE in four stages. The fieldbus communication between physical process and PLC, Master PLC, and SCADA of each stage is achieved through optional wired and wireless channels.

The key asset resources in EPIC [31] mainly include the following: SCADA system, which uses Pcvue in EPIC and runs on a personal computer equipped with the Windows operating system; PLCs, which use WAGO's PLC series PFC200 perform logic control in EPIC, located on control and network panel, and work based on firmware and control logic programs, and in a few cases, use Modbus TCP/IP communication; Codesys (Codesys v3), which is the programming standard of PLC; IEDs, SIPROTEC Relays from Siemens for protection and control which is used in EPIC, located in the control center and uses IEC61850 standard to communicate with the rest of the system, and maintains the entire process by firmware and control logic; VSD, SEW Eurodrive and the corresponding motor which are used as VSD in EPIC, located in the motor/generator room; and network switches and access points located in the network control panel which adopt HIRSCHMANN products.

4.2. Ontology Structure. Mining EPIC-related vulnerabilities to form a knowledge graph correspond to network layout and asset information of EPIC. For the convenience of research, the study mainly considers assets involved in the communication layout of EPIC. In this paper, we use assets as keywords to collect strong correlation information from vulnerability databases and form a relation extraction corpus with common vulnerabilities in ICS. The communication layout in EPIC is mapped into multiple groups of bidirectional communication relation between nodes and represented by triples. The connection between internal network layout and external threat information is established through the matching between nodes and specific asset information, thus forming the final industrial control network security knowledge graph. The ontology structure we define in this paper is shown in Figure 3.

We define 9 relations including model, _have_, version_, AKA, version, _by_, CVSS_score, module, help_out, and conn and additionally define two relations, comm and asset_info, to represent the connection relation in the EPIC communication network and asset information. There are 11 relations in total. Use <head, tail, relation> to identify the head entity, tail entity, and the relation between them. In this paper, the information of the network layout is mapped into triples <asset1, asset2, comm>, such as <MIED1, MIED2, comm>. Furthermore, <asset, Product, asset_info> combines the internal network layout and external threat intelligence through connecting asset nodes with the product information used by them. Through analysis of vulnerability databases, the vulnerability number is associated with CVSS score, solution, attack vector, and other relevant vulnerability numbers, making vulnerability analysis more multidimensional.

5. The Proposed Model

In this section, we describe the architecture of the proposed cybersecurity entity relation extraction model and then introduce each component of the model in detail.

Under the framework of distant supervised learning, the problem of insufficient label data in deep learning can be solved, but at the same time, it also brings some problems, such as the low-quality label data and the wrong label data. This would have a great impact on subsequent tasks of entity relation extraction. In view of the above problems, we propose a distant supervised relation extraction model ResPCNN-ATT based on the deep residual neural network and attention mechanism. The framework is shown in Figure 4. The model is mainly composed of a vector representation layer, a deep residual convolutional network layer, and a multi-instance attention layer.

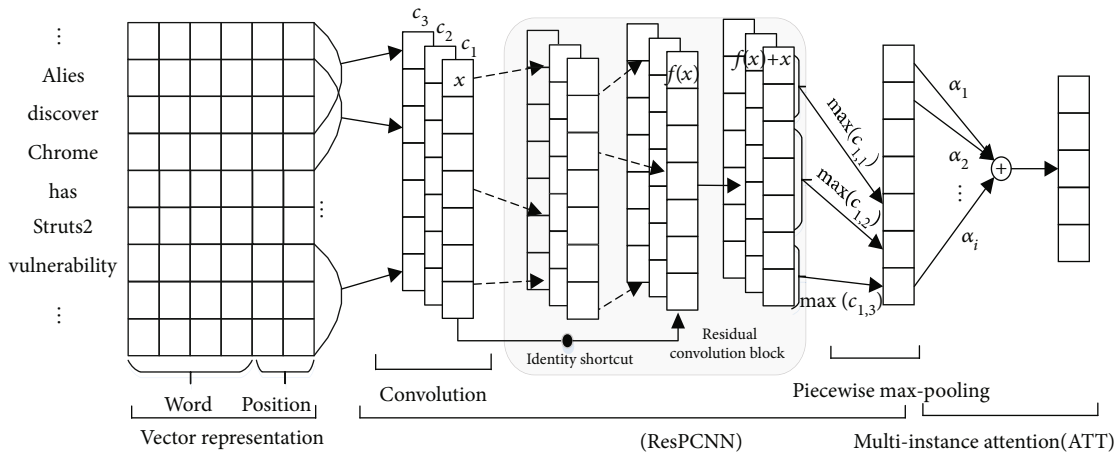

Figure 4: Cybersecurity relation extraction model based on ResPCNN-ATT.

The model first uses the pretrained word vector and the position vector between entity pairs as input, which can highlight the role of the two entities, and then uses the piecewise convolutional neural networks to extract semantic features. At the same time, deep residual learning is introduced to solve the problem of gradient disappearance caused by noise data, so as to extract more effective semantic features. Finally, in order to better capture the more important semantic features in sentences, the multi-instance attention mechanism is used to calculate the correlation between instances and corresponding relation, so as to reduce the impact of noise data and improve the performance of relation extraction.

5.1. Vector Representation. The vector representation layer in the model mainly includes word embedding and position embedding.

5.1.1. Word Embedding. Before training the relation extraction model, the text data needs to be vectorized so that the model can read the data. Compared with traditional one-hot coding, word vector mapping can represent more semantic and syntactic information. Word vector mapping is to map each word in the text to a k-dimensional real-valued vector. It is a distributed representation of words. When training a neural network model, the most common method is to randomly initialize all parameters and then use an optimization algorithm to optimize the parameters. Research shows that when a neural network is initialized with a pretrained word vector, the parameters can be converged to a better local minimum.

For a given sentence $X = \{x_1, x_2, \cdots, x_n\}$ consisting of n words, use word2vec to map each word to a low-dimensional real-valued vector space, then perform word vector processing on the sentence, and finally get a vector representation of each word in the sentence, to form a word vector query matrix D^c. Each input training sequence can be mapped by the word vector query matrix D^c to obtain the corresponding real-valued vector $x_t = \{w_1, w_2, \cdots, w_n\}$.

5.1.2. Position Embedding. In the relation extraction task, we focus on finding the relation of entity pairs. Words that are often close to the entity are more able to highlight the relation between the two entities, such as some verbs: attack, use, etc. Therefore, in order to make full use of the information in the sentence, the position of each word in the sentence for two entities is an important feature in the relation extraction task. This paper uses the position vector (position embeddings (PE)) mapping representation method proposed by Zeng et al.; that is, the relative distance between the current word, entity e_1 and entity e_2, is stitched and converted into a vector representation through embedding. In sentence position vectorization, if the dimension of the word vector is d^c and the dimension of the position vector is d^p, then the dimension of the sentence vector is

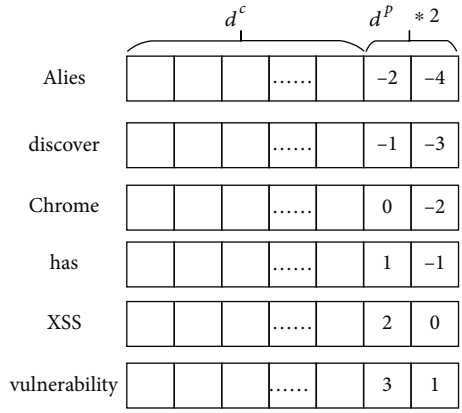

Figure 5: Position embedding.

$$d^s = d^c + d^p * 2. \tag{1}$$

For example, the vectorized representation of "Alies discover Chrome has XSS vulnerabilities" is shown in Figure 5, "Chrome" and "XSS" in the sentence correspond to entities e_1 and entities e_2, respectively. Then, the distance

from "Alies" to "Chrome" is 2, the distance from "Alies" to "XSS" is 4, the distance from "vulnerability" to "Chrome" is -3, and the distance from "vulnerability" to "XSS" is -1.

5.2. Deep Residual Neural Network. In cybersecurity relation extraction tasks, the main challenge is that the length of the input sentence is variable and not fixed, and important feature information may appear in any area of the sentence. Therefore, in order to be able to use all local features and predict relations globally, this paper uses a piecewise convolutional neural network PCNN model to extract semantic features in sentences.

In this paper, a residual convolution block is designed for residual learning. Each residual convolution block is a sequence composed of two convolution layers. After each convolution layer, the activation function ReLU is used for nonlinear mapping, and features are then extracted using a local maximum pool. The kernel size of all convolution operations in the residual convolution module is w, and the newly generated features are guaranteed to be the same size as the original ones through the border padding operation. The convolution kernels of the two-layer convolution are $W_1, W_2 \in R^{w*1}$. The first layer of the residual convolution block is

$$c_{i,1} = f(W_1 \cdot c_{i,i+w-1} + b_1). \tag{2}$$

The second layer is

$$c_{i,2} = f(W_2 \cdot c_{i,i+w-1} + b_2), \tag{3}$$

where b_1, b_2 are bias vectors. In this paper, we optimize the residual learning to get the output vector c of the residual convolution block [32, 33].

After the semantic feature is acquired by the convolution layer, the most representative local feature is further extracted by the pooling layer. In order to capture characteristic information of different sentence structures, a piecewise max pooling process is used.

5.3. Multi-Instance Attention. In the relational extraction model, sentence-level attention is built on multiple instances, dynamically reducing the weight of noisy instances, and making full use of semantic information in these sentences to obtain final sentence vector representation.

For the instance set $S = (g_1, g_2, g_3, \cdots, g_n)$ describing the same entity pair $<e_i, e_j>$, g_i is the instance vector output by the convolution layer and n is the number of instances contained in the set S. This paper will calculate the correlation degree between the instance vector g_i and the relation r. In order to reduce the impact of noise data and make full use of the semantic information contained in each instance in the set, the calculation of instance set vector S will depend on each instance g_i in the set:

$$S = \sum_i \alpha_i g_i, \tag{4}$$

where a_i is the weight of the input instance vector g_i, which measures the correlation of the corresponding relation r. The calculation formula of α_i is as follows:

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}. \tag{5}$$

e_i is a query-based function, which indicates the degree of matching between the input instance vector g_i and the prediction relation r.

Conditional probability of prediction relation $p(R \mid S)$ is calculated by softmax function:

$$p(R \mid S) = \text{soft max}(\tilde{r}S + b), \tag{6}$$

where $\tilde{r}$ is the relation matrix and b represents the bias vector. $p(R \mid S)$ is used to predict the relation between pairs of cybersecurity entities:

$$\tilde{R} = \arg\max p(R \mid S). \tag{7}$$

6. Performance Evaluation

In this section, we empirically demonstrate the performance of the proposed method on datasets CSER and ICSER. Commonly used Precision-Recall (*P-R*) curve, AUC value, and average accuracy (*P@N*) are used to evaluate the model. The *P-R* curve is a curve drawn with the recall rate R as the abscissa and the accuracy rate P as the ordinate, using P and R at different confidence levels. The AUC value is the area included under the *P-R* curve. Generally, the larger the AUC value is, the better the model performs. *P@N* is the accuracy rate calculated by comparing the first N relation instances.

6.1. Datasets and Parameters. In order to verify the performance of our proposed model, we build a cybersecurity entity relation (CSER) dataset. 10 types of relations were labeled. The dataset CSER is clawed from the Freebuf (https://www.freebuf.com/) website and wooyun vulnerability database, which includes network text data such as technology sharing, network security, and vulnerability information.

The set of dimensions of the word vector is $\{50, 60, \cdots, 300\}$. The set of dimensions of the position vector is $\{1, 2, \cdots, 10\}$. During the training process, the Adam optimizer performs optimization training. The value set of the learning rate is {0.01, 0.001, 0.0001}. The set of batch sizes processed in one iteration is {40, 160, 640, 1280}. In order to prevent the model from overfitting, the dropout method is used in CNN. Other parameters are shown in Table 1.

6.2. Results and Analysis. The experimental comparison in this paper mainly compares two aspects of the models.

On the one hand, it uses the CNN algorithm with different performances to encode the training data and extract the semantic features in the sentence, mainly including the traditional models: CNN, PCNN, and ResPCNN.

The second aspect is based on how CNN/PCNN/ResPCNN uses the information in the packaging bag for

Table 1: Parameters.

Parameters	Value
CNN window size	3
CNN hidden size	230
Learning rate	0.01
Batch size	160
Epoch	60
Dimension of the position vector	5
Dropout rate	0.5
Dimension of the word vector	50

experimental comparison. Three different methods were used to process the information in the bag, namely, AVE, ONE, and ATT. AVE assigns the same weight to all the sentences in the packet as the entity pair, that is, $\alpha_i = 1/n$. ONE means to take the instance vector with the highest confidence and find a sentence with the highest score from each bag to represent the entire bag. All models in this paper have been trained and tested on the dataset CSER. Figures 6–8 show the *P-R* curves of the results on different bag models. AVE can introduce more information of sentences, but since it has the same evaluation on each sentence, it will also introduce noise from the wrong label data, which reduces the performance of relation extraction, so AVE has the lowest performance of relation extraction among the bag models. The AUC value difference between ONE and ATT on model PCNN is 0.12%, which refers that the performance of relation extraction does not differ much. On model ResPCNN and CNN, the performance of relation extraction of ATT is slightly higher than that of ONE; ATT can achieve a higher accuracy rate throughout the recall scope.

From Figure 9, the AUC value of the model ResPCNN-ATT is the highest value on the dataset CSER, which reaches 12.68%. The model ResPCNN-ATT proposed in this paper can better extract the deep semantic information of sentences, indicating that the introduction of the ATT method can effectively reduce the redundant data in distant supervised learning.

As can be seen from Table 2, comparing the accuracy of the first 100, 200, and 300 relation instances on the dataset CSER, the relation extraction accuracy of ResPCNN-ATT is the highest, which reaches 32.67%. However, the accuracy of the CSER dataset is lower than other datasets. This is because the sentences in the CSER dataset are mixed with Chinese and English; the more complicated the sentence structure is, the less obvious the entity relation characteristics are, and the less the corpus data is.

In order to further analyze the relation extraction model proposed in this paper, by adding the depth of the ResPCNN-ATT model to verify the effectiveness of the introduction of residual learning, comparative experiments of convolutional layers with different depths are designed. In this paper, the number of convolutional layers is increased by increasing the number of residual convolution blocks, and the experimental comparison is performed on the CSER dataset.

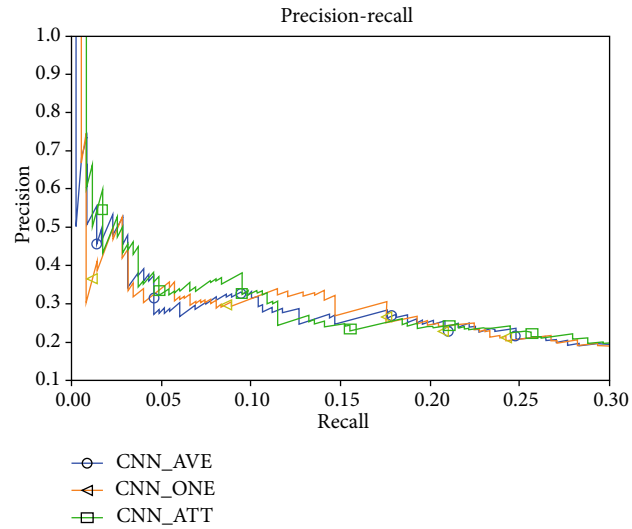

Figure 6: The results of different bag methods AVE/ONE/ATT based on CNN.

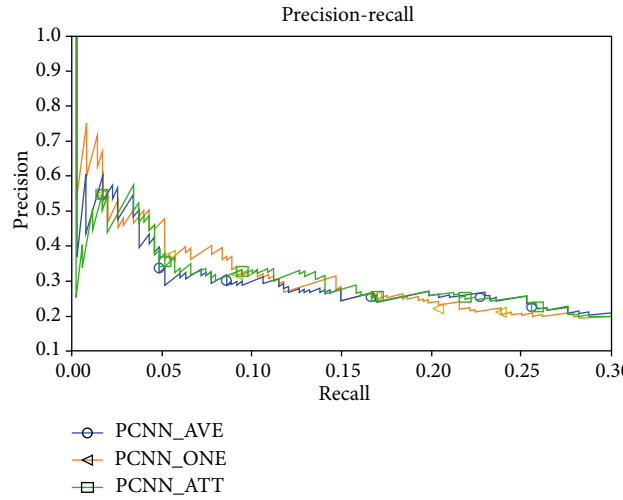

Figure 7: The results of different bag methods AVE/ONE/ATT based on PCNN.

Figure 10 shows the *P-R* curves on models with different depths.

7. CSKG Construction and Visualization for ICS

The proposed model ResPCNN-ATT performs well on the dataset CSER, and further, we apply ResPCNN-ATT to the relation extraction task in the construction of a knowledge graph for EPIC.

7.1. Relation Extraction. We analyze key assets and the communication relation between the assets in EPIC and obtained datasets through labeling in distant supervision. Due to the need for strong data correlation, after filtering and cleaning, 19,838 examples of industrial control network security entity relations were finally formed. 15,937 sentences were randomly selected as training data, which included 3838

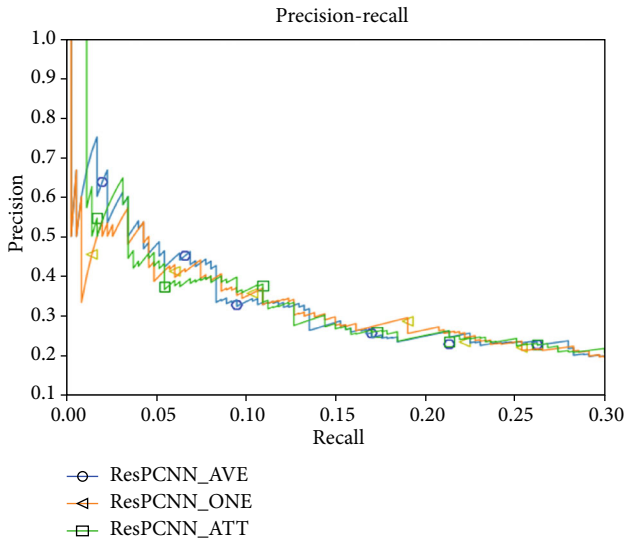

Figure 8: The results of different bag methods AVE/ONE/ATT based on ResPCNN.

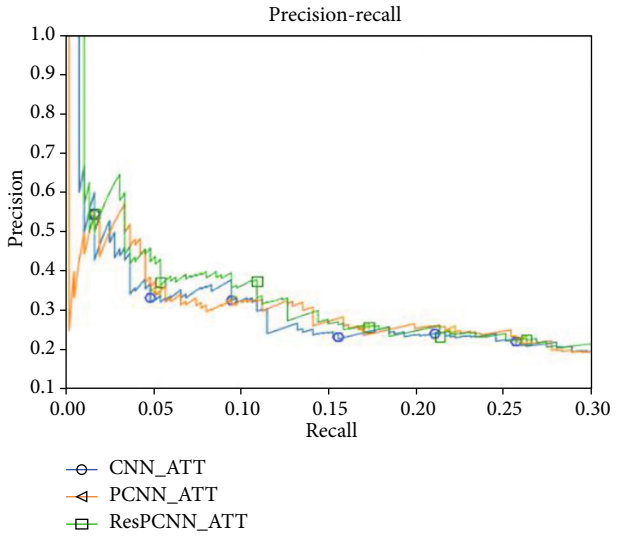

Figure 9: The results of different sentence semantic feature extraction models CNN/PCNN/ResPCNN.

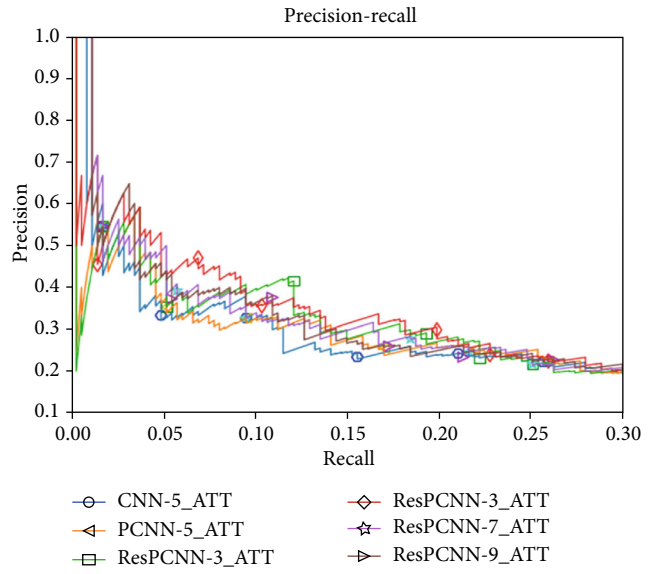

Figure 10: The results on models with different depths.

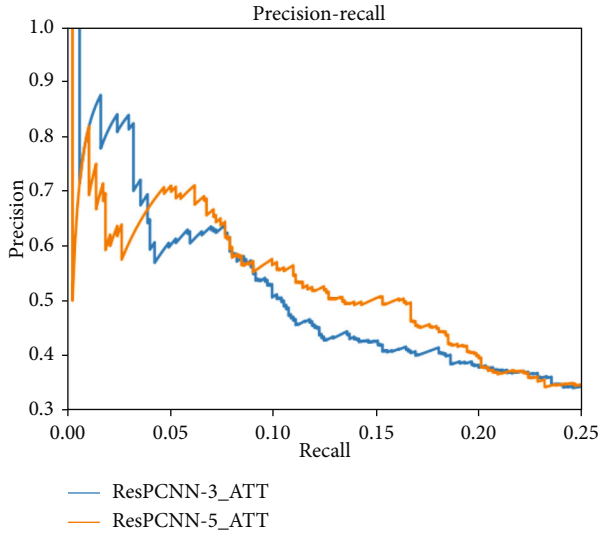

Figure 11: The results of ResPCNN-ATT with different depths on dataset ICSER.

Table 2: Results for the first 100, 200, and 300 extracted relation instances upon manual evaluation.

Models	*P*@100	*P*@200	*P*@300	Mean	AUC
CNN+AVE	0.3267	0.2537	0.2452	0.2743	0.1062
CNN+ONE	0.2971	0.3035	0.2392	0.2799	0.1096
CNN+ATT	0.3267	0.2437	0.2425	0.2710	0.1121
PCNN+AVE	0.2971	0.2587	0.2645	0.2727	0.1096
PCNN+ONE	0.3168	0.2587	0.2358	0.2705	0.1109
PCNN+ATT	0.3267	0.2736	0.2525	0.2842	0.1121
ResPCNN+AVE	0.3267	0.2686	0.2458	0.2804	0.1205
ResPCNN+ONE	0.3564	0.2786	0.2558	0.2969	0.1184
ResPCNN+ATT	0.4158	0.3084	0.2558	0.3267	0.1268

entity pairs, and 4001 sentences were selected as test data, which included 876 entity pairs.

In this paper, when the depth of the ResPCNN-ATT model is 3 and 5, respectively, an experiment is carried out on dataset ICSER, corresponding to different layers of convolution layers. Figure 11 shows the *P*-*R* curves at different depths. The *P*-*R* curves above show the effectiveness of introducing residual learning when the model depth is shallow such as 3 and 5.

Table 3 shows the prediction accuracy and AUC values of the test set in the first 100, 200, and 300 relation instances of the model at two depths. Based on the complex industrial control network security dataset, the model has performed well.

Table 3: Results for the first 100, 200, and 300 extracted relation instances.

Models	*P*@100	*P*@200	*P*@300	Mean	AUC
ResPCNN-3_ATT	0.6237	0.4726	0.4252	0.5072	0.2277
ResPCNN-5_ATT	0.6435	0.5174	0.4850	0.5486	0.2343

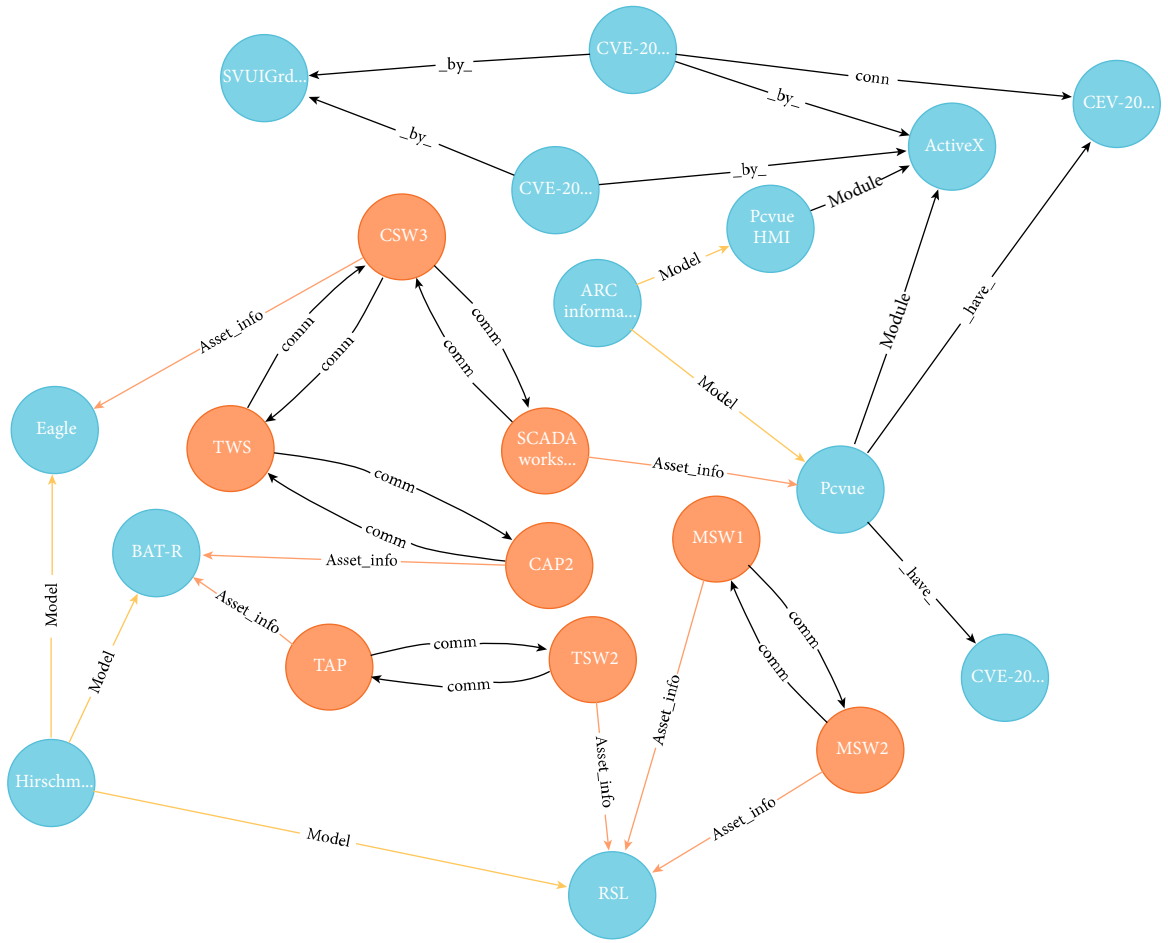

Figure 12: Part of relations of asset node SCADA workstation.

7.2. Visualization and Analysis. Finally, 3878 relationships are extracted and stored. Asset as an entity has the communication relation between other assets in network layout. One specific asset node matches one asset equipment at least; through brands, models, or components used by asset equipment, the corresponding vulnerability information can be connected with the asset. A part of the relations of asset node SCADA workstation is shown in Figure 12.

The versions, components, and vulnerabilities of WAGO RFC200 series of products used by PLC in EPIC can be seen in Figure 13. The correlation between different vulnerabilities is defined, such as the correlation between vulnerabilities from CVE and CWE, which enables the network analysis to locate the source code faster and more accurately.

As shown in Figure 14, the CVSS score can quantify the vulnerability threat level; information such as vulnerability solutions, patch links, and security recommendations is structurally related to the corresponding vulnerability, which can help to troubleshoot equipment failures and strengthen security status. The asset vulnerability corresponding to the vulnerability, such as the port number used, is associated with the exploit relationship.

The preliminary construction of the EPIC industrial control network security knowledge graph not only facilitates daily management, daily maintenance, and network security analysis but also supports the completion of downstream tasks of the knowledge graph. The knowledge expression form in the knowledge graph is simple, intuitive, flexible, and rich. Based on the existing knowledge graph structure, we can deepen the industrial control network security defense at a deeper level and make network security defense research more diversified. Further, through knowledge reasoning, we can link to hidden entities and predict new relationships. It helps find out new attack behaviors and improve the richness and accuracy of the knowledge graph. The mining of entities and relationships offers constant

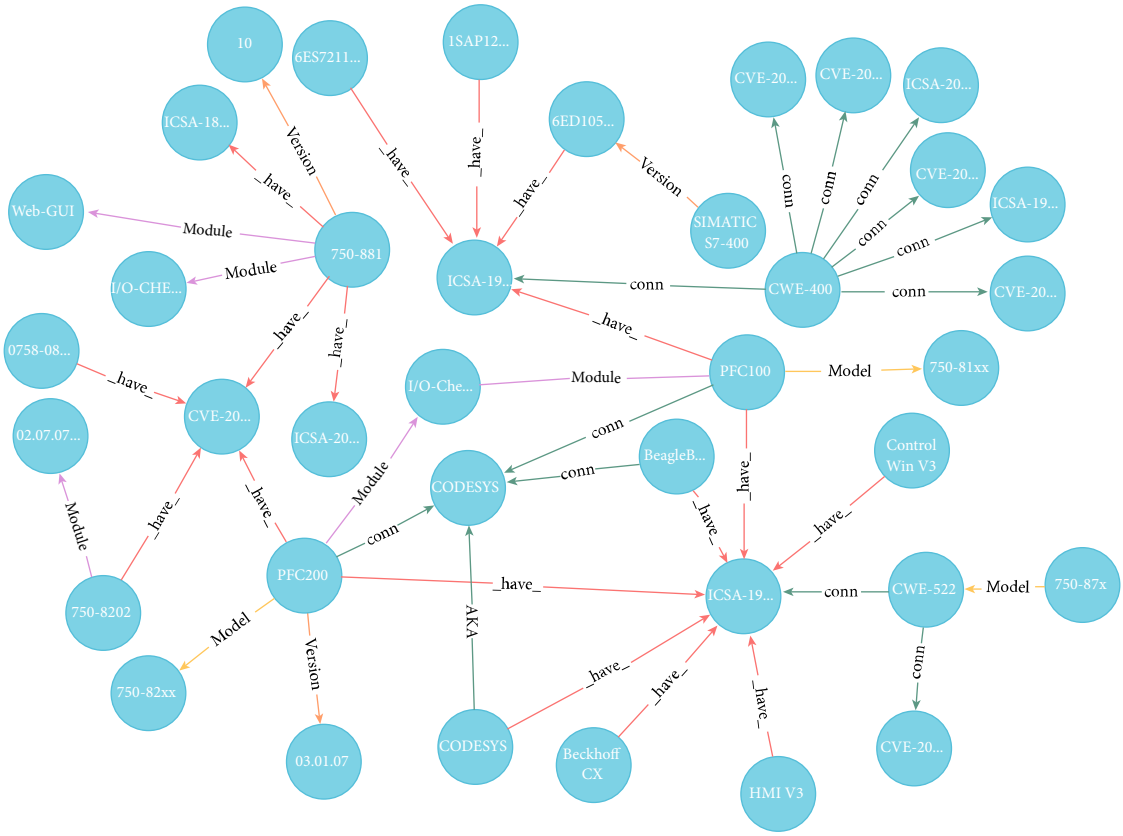

Figure 13: Part of relations of WAGO RFC200.

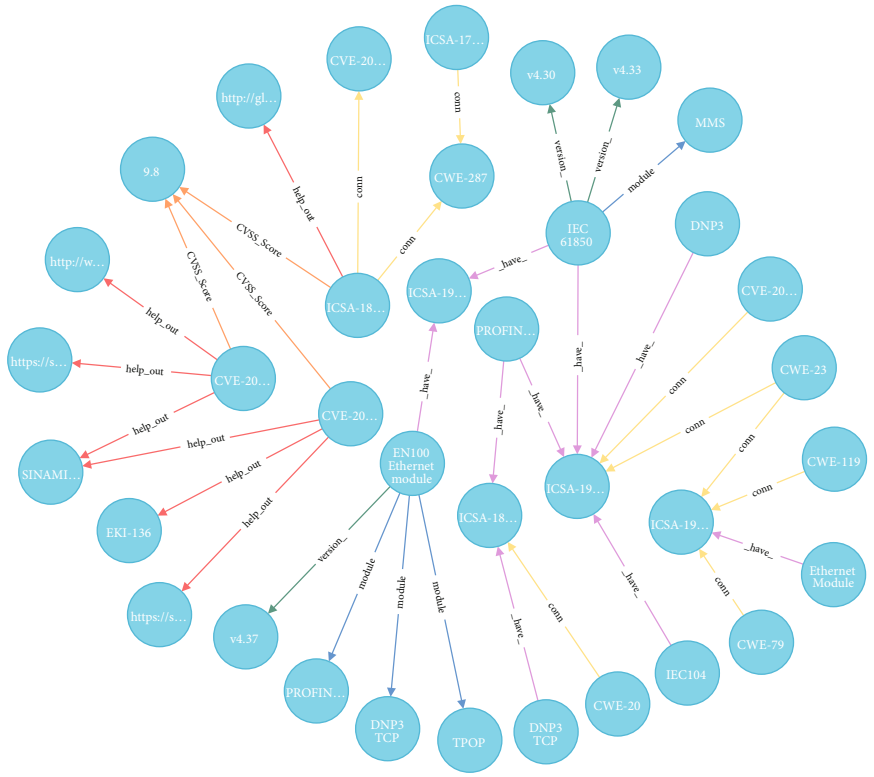

Figure 14: Example of vulnerability information.

supplement for the existing knowledge graph and makes sense in decision-making, to enhance the active defense capability of industrial control network security.

8. Conclusions

In this paper, we propose a novel data-driven industrial network security defense framework, which structures fragmented multisource data and integrates these threat data with the industrial network structure. In order to better mine entity relations in cybersecurity data, we introduce a novel distant supervised cybersecurity relation extraction model ResPCNN-ATT. The experimental results show that the model proposed in this paper has the highest accuracy of relation extraction compared with other model methods on cybersecurity datasets. Further, based on specific industrial control network security scenarios, we constructed an ICS security knowledge graph by applying ResPCNN-ATT, which strengthens the cybersecurity analysis capabilities. In the future, we intend to introduce reinforcement learning to the model to further reduce the impact of noise and study the downstream application tasks of the industrial control network security knowledge graph to strengthen the industrial control network security defense capabilities.

Data Availability

All the data used to support this study were supplied by Guowei Shen under license and so cannot be made freely available. Requests for access to these data should be made to Guowei Shen (gwshen@gzu.edu.cn).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61802081 and Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory Open Fund Project (No.W-2018023).

References

[1] N. Falliere, L. O. Murchu, and E. Chien, "W32. Stuxnet dossier," *White paper, Symantec Corporation Security Response*, vol. 5, no. 6, p. 29, 2011.

[2] I. C. S. C. Alert, *Cyber-attack against Ukrainian critical infrastructure. Cybersecurity Infrastructure Security Agency*, Technical Report ICS Alert (IR-ALERT-H-16-056-01), Washington, DC, USA, 2016.

[3] K. Coffey, R. Smith, L. Maglaras, and H. Janicke, "Vulnerability analysis of network scanning on SCADA systems," *Security and Communication Networks*, vol. 2018, Article ID 3794603, 21 pages, 2018.

[4] L. Zhen, "Cultivate the 5G+ industrial internet to promote mutual progress-interpretation of "5G+ industrial internet" 512 project promotion program," *Network Security and Informatization*, vol. 1, pp. 23-24, 2020.

[5] C. Feng, V. R. Palleti, A. Mathur, and D. Chana, "A systematic framework to generate invariants for anomaly detection in industrial control systems," in *Proceedings 2019 Network and Distributed System Security Symposium*, San Diego, CA, USA, 2019.

[6] S. McLaughlin, C. Konstantinou, X. Wang et al., "The cybersecurity landscape in industrial control systems," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1039–1057, 2016.

[7] H. Holm, M. Karresand, A. Vidström, and E. Westring, "*A survey of industrial control system testbeds*," in *Secure IT Systems*, pp. 11–26, Springer International Publishing, Cham, 2015.

[8] C. Wang, D. Wang, Y. Tu, G. Xu, and H. Wang, "Understanding node capture attacks in user authentication schemes for wireless sensor networks," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2020.

[9] D. Wang, W. Li, and P. Wang, "Measuring two-factor authentication schemes for real-time data access in industrial wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4081–4092, 2018.

[10] A. Sadighian, J. M. Fernandez, A. Lemay, and S. T. Zargar, "Ontids: A highly flexible context-aware and ontology-based alert correlation framework," in *Foundations and Practice of Security. FPS 2013*, J. Danger, M. Debbabi, J. Y. Marion, J. Garcia-Alfaro, and N. Zincir Heywood, Eds., vol. 8352 of Lecture Notes in Computer Science, pp. 161–177, Springer, Cham, 2014.

[11] R. Shittu, A. Healing, R. Ghanea-Hercock, R. Bloomfield, and M. Rajarajan, "Intrusion alert prioritisation and attack detection using post-correlation analysis," *Computers & Security*, vol. 50, pp. 1–15, 2015.

[12] Y. Yao, Z. Wang, C. Gan et al., "Multi-source alert data understanding for security semantic discovery based on rough set theory," *Neurocomputing*, vol. 208, pp. 39–45, 2016.

[13] A. A. Ramaki, A. Rasoolzadegan, and A. G. Bafghi, "A systematic mapping study on intrusion alert analysis in intrusion detection systems," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1–41, 2018.

[14] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xiang, "Data-driven cybersecurity incident prediction: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1744–1772, 2019.

[15] X. Shu, F. Araujo, D. L. Schales et al., "Threat intelligence computing," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1883–1898, Toronto, Canada, 2018.

[16] H. Yu, H. Li, D. Mao, and Q. Cai, "A relationship extraction method for domain knowledge graph construction," *World Wide Web*, vol. 23, no. 2, pp. 735–753, 2020.

[17] X. Liao, K. Yuan, X. F. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the ioc game: toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 755–766, Vienna, Austria, 2016.

[18] G. Siracusano, M. Trevisan, R. Gonzalez, and R. Bifulco, "Poster: on the application of NLP to discover relationships between malicious network entities," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2641–2643, London, United Kingdom, 2019.

[19] Z. Zhu and T. Dumitras, "Chainsmith: automatically learning the semantics of malicious campaigns by mining threat intelligence reports," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 458–472, London, UK, 2018.

[20] A. Pingle, A. Piplai, S. Mittal, A. Joshi, J. Holt, and R. Zak, "RelExt: relation extraction using deep learning approaches for cybersecurity knowledge graph improvement," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 879–886, Vancouver, British Columbia, Canada, 2019.

[21] Y. Dong, W. Guo, Y. Chen, X. Xing, Y. Zhang, and G. Wang, "Towards the detection of inconsistencies in public security vulnerability reports," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 869–885, Santa Clara, CA, USA, 2019.

[22] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1201–1211, Jeju Island, Korea, 2012.

[23] Z. Daojian, L. Kang, L. Siwei, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344, Dublin, Ireland, 2014.

[24] P. Zhou, W. Shi, J. Tian et al., "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 207–212, Berlin, Germany, 2016.

[25] C. N. D. Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," *Computer Science*, vol. 86, no. 86, pp. 132–137, 2015.

[26] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2124–2133, Berlin, Germany, 2016.

[27] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1753–1762, Lisbon, Portugal, 2015.

[28] P. Qin, W. Xu, and W. Y. Wang, "Robust distant supervision relation extraction via deep reinforcement learning," 2018, https://arxiv.org/abs/1805.09927.

[29] Y. Qin, G. Shen, W. Zhao, Y. P. Chen, M. Yu, and X. Jin, "A network security entity recognition method based on feature template and CNN-BiLSTM-CRF," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 6, pp. 872–884, 2019.

[30] S. Adepu, N. K. Kandasamy, and A. Mathur, "Epic: An electric power testbed for research and training in cyber physical systems security," in *Computer Security, SECPRE 2018, CyberICPS 2018*, S. Katsikas, Ed., vol. 11387 of Lecture Notes in Computer Science, pp. 37–52, Springer, Cham, 2018.

[31] S. Adepu, N. K. Kandasamy, J. Zhou, and A. Mathur, "Attacks on smart grid: power supply interruption and malicious power generation," *International Journal of Information Security*, vol. 19, no. 2, pp. 189–211, 2020.

[32] Y. Y. Huang and W. Y. Wang, "Deep residual learning for weakly-supervised relation extraction," 2017, https://arxiv.org/abs/1707.08866.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.