electronics

MDPI

Article

Dual-Safety Knowledge Graph Completion for Process Industry

Lingzhi Wang [1,2,†], Haotian Li [1,2,†], Wei Wang [1,2], Guodong Xin [1,2] and Yuliang Wei [1,2,*]

[1] Research Institute of Cyberspace Security, Harbin Institute of Technology, Weihai 264200, China; 23s130410@stu.hit.edu.cn (L.W.); 22b903069@stu.hit.edu.cn (H.L.); wwhit@hit.edu.cn (W.W.); gdxin@hit.edu.cn (G.X.)

[2] School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264200, China

[*] Correspondence: wei.yl@hit.edu.cn

[†] These authors contributed equally to this work.

Abstract: With the rise of Industry 4.0, control systems have taken on increasing importance in industrial processes, and ensuring their security has become a pressing issue. While recent research has focused on cybersecurity threats, the security risks inherent to industrial processes themselves have been overlooked. Additionally, existing tools cannot simultaneously analyze both cyber vulnerabilities and processes anomaly in industrial settings. This paper aims to address these issues through two main contributions. First, we develop a knowledge graph to integrate information on security risks across cybersecurity and industrial processes, providing a foundation for comprehensively assessing threats. Second, we apply the link prediction task to the knowledge graph, introducing an embedding-based approach to unveil previously undiscovered knowledge. Our experiments demonstrate that the proposed method exhibits comparable performance on link prediction and is capable of mining valuable and diverse potential risks in industrial processes.

Keywords: industrial process security; cybersecurity; knowledge graph; link prediction

check for updates

Citation: Wang, L.; Li, H.; Wang, W.; Xin, G.; Wei, Y. Dual-Safety Knowledge Graph Completion for Process Industry. *Electronics* **2024**, *13*, 214. https://doi.org/10.3390/electronics13010214

Academic Editors: Aryya Gangopadhyay and Myung-Sup Kim

Received: 12 November 2023
Revised: 14 December 2023
Accepted: 26 December 2023
Published: 3 January 2024

1. Introduction

Control systems play a pivotal role in the process industry by ensuring the smooth operation of complex production processes. These systems typically consist of several key components, each serving a specific purpose. Monitors, for instance, provide real-time data visualization and process status information, giving operators critical insights into the ongoing processes. Another crucial component is the controller, which oversees and regulates individual devices and process variables, ensuring that they function optimally. Additionally, network switchers are instrumental in facilitating seamless communication and data exchange among the various components of the control system. The significance of control systems extends across various application domains, including the chemical industry [1], oil and gas production [2], and the power industry [3]. Consequently, the security of control systems has gained increased and more extensive attention.

Existing work on security insurance of control systems has focused on protecting them from cyberattacks. Jarmo Alnen et al. [4] introduced a cybersecurity risk analysis method that relies on the proposed ontology. This method provides a structured data repository aimed at bolstering the assessment process and establishing traceability between the generated artifacts. Angelo Corallo et al. [5] delved into the cybersecurity challenges encountered in the realm of the Internet of Everything during the Industry 4.0 era. Certainly, the increased efforts in terms of cybersecurity to protect control systems are indeed noteworthy; however, the broad issue of security throughout the entire production process often remains inadequately addressed. For instance, in 2018, Tesla faced significant challenges with its highly automated production line for the Model 3. The company employed an extensive fleet of robots equipped with sensors to aid in component assembly. Unfortunately, a few robots experienced sensor failures or misalignment issues, leading to production delays.

These disruptions stemming from sensor-related issues caused production bottlenecks, resulting in a slower output of Model 3 vehicles—a pivotal product for Tesla. Consequently, these setbacks led to delivery delays for customers and carried financial implications for the company. The reality is that sensor failures in industrial processes can result in substantial loss of both human life and property.

Therefore, a more comprehensive strategy should not only focus on strengthening control system defenses, but also prioritize the implementation of robust security measures throughout the entire process industry.

In order to aptly represent the comprehensive protection of critical infrastructure, we have coined the term 'dual-security' to encompass both the security of cyberspace and the industrial process. Since a substantial portion of prior research has concentrated on cybersecurity, the challenge of simultaneously addressing dual-security risks within the process industry has remained an uncharted territory.

Recently, knowledge graphs (KGs) have emerged as a fundamental technique for incorporating knowledge from diverse fields. It provides a structured representation of knowledge, i.e. factual triples in the form of (h, r, t), denoting the head and tail entities and the relationship between them [6]. For instance, in a KG, we may encounter entities like "Albert Einstein" and "Theory of Relativity" connected by the relationship "developed". These KGs are constructed by combining information from various sources, including structured databases, unstructured text, and linked open data [7]. The structured nature of KGs ensures semantic consistency, enabling machines to not only retrieve information, but also understand the context and perform sophisticated reasoning tasks. KGs have found applications in a wide range of domains, from natural language processing [8–10] and information retrieval to recommendation systems [11,12] and medical healthcare [13–15].

Link prediction is a fundamental subfield within knowledge representation and reasoning, dedicated to the task of inferring missing facts and relationships within the graph. Addressing the link prediction challenges has led to the development of various distinct approaches, among which the primary category is embedding-based methods. This kind of method is designed to acquire distributed low-dimension vector representations, commonly referred to as embeddings, for entities and relations within the KG. Most representative approaches within this research paradigm encompass TransE [16], DistMult [17], ConvE [18], TuckER [19], etc.

Link prediction on knowledge graphs serves as an effective and transparent approach for revealing hidden knowledge within knowledge graphs. In this paper, we study the problem of protecting the process industry in terms of dual-security. Specifically, we collect knowledge from both security aspects and build a dual-security knowledge graph. Then, we propose to discover potential security risks through inference on the KG with an embedding-based link prediction method. Empirical experiments substantiate the efficacy of our method, as it achieves good performance on the real-world dataset and mines meaningful dual-security risks. In addition to conducting comprehensive dataset experiments, we also performed an analysis of the risk prediction capability of existing models using experiments conducted on crack hydrogenation industrial processes. Our analysis showed that the employed approach is more accurate in predicting the potential risks of unitary industrial processes compared to other methods.

This paper contributes to two main aspects. Firstly, it innovatively integrates cybersecurity and industrial process security, extracting knowledge from both dimensions, into a knowledge graph. Secondly, it employs a relatively efficient method for link prediction, enabling it to mine undiscovered knowledge from the knowledge graph and effectively mitigate potential risks in industrial processes.

2. Related Work

Cybersecurity knowledge graphs (CSKG) [20] represent a specialized category of knowledge graphs tailored for the cybersecurity domain. They comprise entities and relations extracted from a myriad of attack and defense scenarios within the cybersecurity landscape [21,22]. Furthermore, abstract concepts such as 'attacker', 'attack pattern', and 'vulnerability' can be seamlessly represented as entities, while the paths of attacks and other connections are aptly captured as relations. CSKG adeptly employs knowledge graph construction techniques to extract and effectively integrate pre-existing knowledge from various security data sources, ensuring a comprehensive and unified understanding of the cybersecurity landscape. The link prediction of CSKG has three major applications in cybersecurity, namely the attack prediction [23], threat hunting [24] and intrusion detection [25–27]. However, there is an inadequate focus on process safety.

The application of knowledge graphs in the industry can be partitioned into two distinct phases, namely the construction and deduction periods, as indicated by Li et al. [28]. In the construction phase, the primary objective revolves around the integration of multiple text mining and machine learning tools to process raw data, yielding triples of the form (h, r, t) for the knowledge graph. As a result, Natural Language Processing (NLP) techniques [29–31] and associated toolkits are frequently harnessed to automatically extract entities from various unstructured knowledge resources. Notably, the utilization of knowledge graphs in industry transcends the conventional dissemination of existing knowledge items; instead, it caters to the elevated demands for synthesis and innovation within the industrial domain. Consequently, the deduction phase of industrial knowledge graphs strives to meet these requirements. Knowledge deduction can be further subcategorized into attribution prediction [32] and link prediction [33–35]. More specifically, the deduction process can be effectively reformulated into a series of matrix manipulations [36–38]. This transformation is achieved by vectorizing the entities and relations, aligning them according to their semantic and topological features within the KG. Such an approach not only enhances the computational efficiency, but also leverages the inherent structural properties of the KG for knowledge deduction. Nonetheless, the majority of the domain-specific knowledge graphs primarily focus on industrial products and services, and they often fall short in addressing the crucial aspect of industrial security.

Our work is related to the task of link prediction in dual-security KG; previous efforts in link prediction can be divided into two main aspects, embedding-based prediction and rule-based prediction. Former approaches focused on learning low-dimension representation for entities and relations. Related methods include TransE [16], TransH [39], ComplEx [40], etc., and detect facts by projecting entities and relations into a semantic space and conduct algebraic operations on the space. Specifically, TransE [16] represents the triples into d-dimensional space, $h, r, t \in \mathbb{R}^d$ and generates embeddings following the translational principle $h + r \approx t$. TransH [39] assigns a hyperplane for each relation to satisfy the N-to-N relationship scenarios. ComplEx [40] firstly uses complex vector space to generate embeddings. Through the space, we can capture both symmetric and antisymmetric relations, $h, r, t \in \mathbb{C}^d$, h can be formulated as $h = R(h) + iI(h)$ where $R(h)$ and $I(h)$ are real and imaginary parts of h, respectively. On the other hand, the latter(rule-based) approaches do not encounter these issues, including NeuralLP [41], DRUM [42]. They leverage first-order logic and rule-based approaches to infer missing relationships and facts. NeuralLP [41] combines neural networks with logical rules and utilizes a differentiable framework to learn the rules from data, allowing for the incorporation of prior domain knowledge. The key idea behind NeuralLP is to bridge the gap between symbolic reasoning and neural network learning, making it an attractive approach for rule-based knowledge graph completion. DRUM [42] integrates rules into an end-to-end differentiable neural network. This approach facilitates rule-based reasoning while benefiting from the learning capabilities of neural networks and extends the knowledge graph with rule-instantiated triples, enabling the incorporation of logical rules in the knowledge graph completion process.

3. Methodology

In this section, we will commence by presenting essential background knowledge, encompassing key concepts such as knowledge graphs, link prediction, and the fundamental steps involved in embedding-based methods for link prediction. Subsequently, we will delve into an extensive exposition of the best-performing model, namely PairRE.

3.1. Preliminaries

Knowledge graph $\mathcal{G} = \{\mathcal{E}, \mathcal{P}\}$ is a structured representation of knowledge, where $\mathcal{E}$ and $\mathcal{P}$ denote the set of entities and relations, respectively. It functions as a semantic network that organizes information into triples, denoted as (s, p, o), where $s, o \in \mathcal{E}$ and $p \in \mathcal{P}$. These triples take the form of $s \xrightarrow{p} o$, presenting a graph-like structure.

Link prediction task refers to a key task in knowledge graph completion. It involves predicting missing triples by identifying the most plausible tail entity t^* from the set of entities $\mathcal{E}$ for a given incomplete triple $(h, r, ?)$. During inference, when provided with h and r as inputs, the model computes a ranked list of all entities, with higher scores indicating higher rankings. In essence, link prediction is a supervised learning technique that aims to minimize the dissimilarity between the predicted entity and the ground truth provided in the available data.

Embedding-based methods for link prediction generally consist of three key steps. In the first step, we initialize embeddings for all entities and relations. To achieve this, we establish an index table, wherein the indexes are associated with entity and relation IDs, and the table's entries comprise stochastic vectors that represent the embeddings. The second step involves the formulation of an objective function designed to assess the score of a triple (h, r, t) within a low-dimensional space. This objective function can be realized through various methods, such as TransE, ComplEX, TuckER, and others. The final step is training and inference. During training, the model is trained, often using negative samples to ensure its robustness and accuracy. Negative samples are essentially fictitious triples constructed at random. These fictive triples share the same h and r with the factual triple, but the t is substituted with a different entity. We anticipate that the scores of true triples will be higher than those of the negative samples. Adhering to this principle, the model iteratively optimizes the embedding tables until convergence is achieved. During inference, the process typically entails taking the embeddings of h and r as inputs, followed by the computation of the score for (h, r, t') across all potential entities $t' \in \mathcal{E}$. Higher scores indicate a higher probability of a connection between entities h and t' through the relation r. This approach enables effective predictions and assessments.

3.2. The Link Prediction Model

We now advocate the utilization of PairRE [43] for augmenting the dual-security knowledge graph. This choice is motivated by two primary factors: Firstly, the data in the graph encompass diverse complex relations, including N-to-1, 1-to-N, and N-to-N relationships. Secondly, the data exhibit various relation patterns, including symmetry and antisymmetry. These intricacies pose a unique challenge that other existing methods are ill-equipped to address simultaneously. However, PairRE stands out by employing paired embedding vectors for each relation representation, allowing for adaptable margin adjustments in the loss function. This adaptability is pivotal in effectively accommodating complex relations and addressing the challenges presented by the dual-security knowledge graph.

PairRE adopts a unique approach by acquiring paired representations for each relation. For a given training triple (h, r, t), the model learns vector embeddings for entities and relations within real space. Specifically, it treats relation embedding as a pair of vectors, denoted as (r_h, r_t). These paired vectors are responsible for projecting the head entity h and the tail entity t into the Euclidean space, with the projection operation being an element-wise product between the two vectors. Furthermore, the method calculates the distance between these two embeddings to gauge the plausibility of the triple. The desired outcome is that $h \circ r_h \approx t \circ r_t$ when the triple (h, r, t) holds true. According to the calculated measure, the scoring function is defined as follows:

$$\mathcal{S}_r(h,t) = -||h \circ r_h - t \circ r_t|| \tag{1}$$

where $h, r_h, r_t, t \in \mathbb{R}^d$ with h and t subject to specific constraints, such as $||h||^2 = ||t||^2 = 1$.

To further optimize the model, we employ a contrastive learning measure, specifically adopting the self-adversarial negative sampling loss [44] as the training objective:

$$\mathcal{L} = -log(\sigma(\gamma - \mathcal{S}_r(h,t))) - \sum_{i=1}^{n} p(h'_i, r, t'_i) log(\sigma(\mathcal{S}_r(h'_i, t'_i) - \gamma)) \tag{2}$$

In this equation, γ represents a fixed margin, and σ denotes the sigmoid function. (h'_i, r, t'_i) corresponds to the ith negative triple, while $p(h'_i, r, t'_i)$ signifies the weight assigned to this negative sample. The weight, $p(h'_i, r, t'_i)$, is calculated using the softmax function:

$$p((h'_i, r, t'_i)|(h,r,t)) = \frac{exp(\mathcal{S}_r(h'_i, t'_i))}{\sum_{j=1}^{n} exp(\mathcal{S}_r(h'_j, t'_j))} \tag{3}$$

An illustration depicting the structure of this paper is presented in Figure 1. First, we extract entities and relations from the extensive documentation of industrial processes through manual annotation. Next, we employ this information to construct a dual-security knowledge graph, incorporating data from a vulnerability database. Finally, we utilize the PairRE method for the link prediction task, enabling the prediction of potential risks in the dual-security domain.

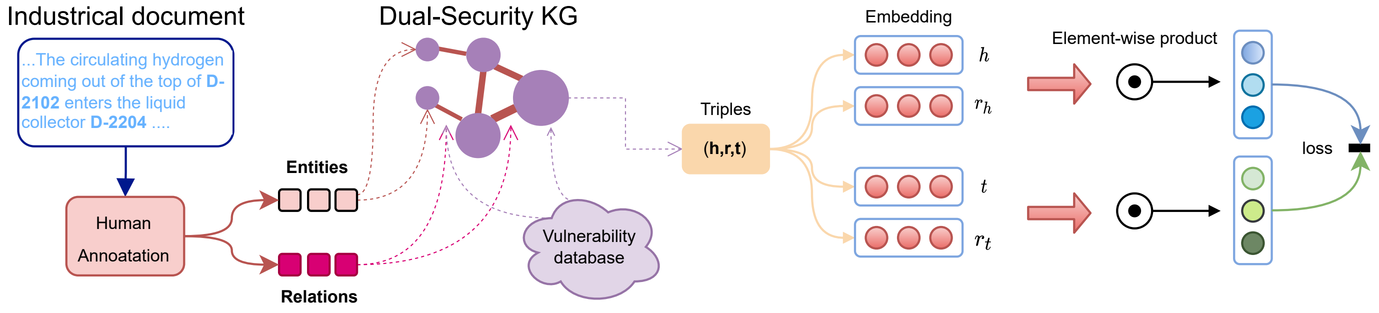

Figure 1. The construction of dual-KG and PairRE model.

4. Experiments

This section is organized into three subheadings. Firstly, we provide a detailed overview of our experimental setup, which includes a description of the dataset, evaluation metrics, and other relevant details. Secondly, we present a concise and accurate analysis of the experimental results. Finally, we draw insightful interpretations from these results, offering our conclusions based on the experiments.

4.1. Experiment Setting

Datasets, we culminated in the construction of our knowledge graph, which is founded on real-world resources and encompasses impressive 37,000 triples. We can find detailed statistics about the dataset in Table 1. Additionally, to showcase the model's proficiency in predicting potential risks within specific industrial processes, such as crack hydrogenation, we generated a focused sub-KG. This sub-KG encapsulates the knowledge related to the crack hydrogenation process alongside all relevant cybersecurity information. Refer to Table 2 for detailed statistics about this sub-KG. To optimize the data for our link prediction task, we meticulously partitioned it into three distinct subsets, ensuring a balanced division. Specifically, we allocated the data into a test set, a training set, and a validation set, maintaining a proportional split ratio of 1:8:1.

Table 1. Statistics of dual-Security knowledge graph dataset.

Entity	Relation	Triples	Train	Validation	Test
2247	45	37,000	29,600	3700	3700

Table 2. Statistics of sub-KG about crack hydrogenation and cybersecurity information.

Entity	Relation	Triples	Train	Validation	Test
104	25	10,686	8544	1068	1074

Evaluation metric, in the case of each query of the form $(h, r, ?)$ or $(?, r, t)$, we calculate a score for every entity while simultaneously assessing the rank of the correct answer. To evaluate the performance of our models, we employ two widely-used metrics, namely the Mean Reciprocal Rank (MRR) and Hit@k. MRR is computed as the mean of the reciprocal ranks for the answer entities. Hit@k, on the other hand, measures the percentage of the desired entities correctly ranked within the top k positions. Furthermore, it is important to note that these metrics exclude the scores of all known true triples in the training, validation, and testing sets. This exclusion allows for a more robust evaluation of our models, providing valuable insights into their effectiveness.

Comparison of models, in our experiments, we applied our dataset to various models, primarily focusing on comparing their performance. Given the substantial scale of our dataset, we employed embedding-based methods to conduct our experiments. These methods employ various embedding techniques to represent entities and relations in low-dimensional spaces, which include methods such as TransE [16], TransR [45], TransH [39], HAKE [46], PairRE [43]. We finally compared their performance with each other.

Hyperparameters, we maintained consistent hyperparameter settings during training. These settings entailed a dimension of embedding vectors set to 512 and a maximum of 1000 training epochs. Concurrently, other methods like TuckER [19], which decomposes a tensor into a core tensor with three-factor matrices, and ConvE [18], founded on the graph neural networks framework, we carefully selected the most suitable hyperparameters through multiple experiments. In this scenario, we set the number of graph convolutional network layers to 1, fixed the dimension of embedding vectors to 200, and established the maximum number of training epochs as 2000, among other key parameters.

4.2. Results on Link Prediction

All methods were evaluated across evaluation metrics on the dual-security dataset, and the results are presented in Table 3. Notably, these evaluations yield clear observations on the performance of these methods. It is apparent that the graph convolutional network method (ConvE) excels and outperforms some of the traditional embedding-based methods, including TransE, TransR, TransH, and TruckER. This observed superiority can be attributed to the incorporation of graph structures, allowing these models to harness a more comprehensive understanding of the training data—a critical aspect for effective link prediction tasks. However, it is worth mentioning that recently proposed methods, such as HAKE and PairRE, manage to outperform even the graph convolutional network methods. This superior performance may be attributed to the adoption of more intricate embedding techniques and measures, underscoring the benefits of sophisticated embedding approaches in addressing dual-security link prediction challenges.

While PairRE demonstrates outstanding overall performance on the dual-security dataset, achieving a commendable level of accuracy, its direct hit rate remains relatively modest at just 22.8%. This observation prompts a vital question regarding the ability of existing models to effectively absorb and adapt to the security data within the dual-security dataset. The dual nature of the KG, incorporating both cybersecurity and industrial process security knowledge, results in a semantic space that is notably vast, rendering it challenging to learn comprehensively. It is imperative to acknowledge that this dataset inherently represents the confluence of two distinct dimensions. Therefore, we deduce that the existing models' deficiencies arise from their failure to adequately learn the knowledge encompassed by both aspects in the dual-KG.

Table 3. The performance of the dominant model on the link prediction. Hit@k is in %.

	MMR	**Hit@1**	**Hit@3**	**Hit@10**
TransE	0.164	0.059	0.204	0.374
TransR	0.199	0.087	0.234	0.439
TransH	0.270	0.159	0.312	0.492
TuckER	0.219	0.131	0.200	0.242
ConvE	0.326	0.218	0.368	0.545
HAKE	0.331	0.215	0.381	0.569
PairRE	**0.352**	**0.228**	**0.414**	**0.593**

The bold in the table indicates the best performance of the corresponding item.

The link prediction results for the sub-KG are presented in Table 4. Notably, PairRE demonstrates efficient prediction of potential risks, with over half of the test data producing accurate predictions. In addition, 77.4% of the test data achieves a top-three ranking, and nearly all test data predictions secure a spot within the top ten. Consequently, focusing on the crack hydrogenation process, obtaining the top three predicted results for all devices would thwart three quarters of potential attacks. Moreover, achieving the top ten predicted results would provide comprehensive protection for almost all devices.

Table 4. The performance of the dominant embedding-based model on sub-KG. Hit@k is in %.

	MMR	**Hit@1**	**Hit@3**	**Hit@10**
TransE	0.249	0.087	0.32	0.567
TransR	0.367	0.195	0.477	0.679
TransH	0.26	0.123	0.305	0.551
TuckER	0.539	0.392	0.615	0.854
ConvE	0.649	0.505	0.744	0.934
HAKE	0.502	0.343	0.577	0.858
PairRE	**0.673**	**0.533**	**0.774**	**0.938**

The bold in the table indicates the best performance of the corresponding item.

After comparing the sub-KG performance with the whole dataset, we found that PairRE is excellent at predicting potential risks in a single scenario. This approach enables full absorption of knowledge within the sub-KG. So, we could generate suitable sub-KG for various specific industrial processes that can be safeguarded against potential security risks by applying this approach.

4.3. Mined Knowledge

By utilizing the best-performing model, PairRE, we have achieved successful predictions of various cybersecurity attacks and industrial process anomalies, as evidenced in Table 5. This approach allows us not only to identify diverse cyber virus attack patterns, but also to access various methods for mitigating cyber risks in the field of cybersecurity. Additionally, we can extend our capabilities to complete the industrial process knowledge graph, enabling us to uncover potential risks within this domain.

To demonstrate the risk prediction ability of our approach, we generate a sub-KG representing the crack hydrogenation process and all cybersecurity knowledge. Furthermore, we present the mined knowledge regarding the crack hydrogenation process and Industroyer attacks, as depicted in Figure 2 and summarized in Table 6. The approach successfully identified vulnerable devices, specifically D-2106, D-2104, and K-2102A/B, under Industroyer attacks. These devices are classified as power supply equipment. Given that Industroyer primarily targets power systems, our approach can precisely identify potentially vulnerable devices. In addition, Industroyer primarily employs remote system discovery to detect a list of devices. Hence, remote system discovery, which includes malicious emails, the office network system, and external O&M terminals, has been predicted for all devices. Subsequently, it identifies target devices and executes various attacks. Depending on the function of different devices, the focus of Industroyer attacks varies. Therefore, the potential risks predicted by the approach differ, as shown in Table 6. However, these identified risks strongly correlate with the devices' functionality. For example, the K-2102A/B serves as the backbone node of the power network. Consequently, attacks against it primarily concentrate on the control aspect, encompassing actions like blocking serial COM, stopping services, and obstructing command messages. In summary, the approach excels in not only identifying vulnerable devices in the industrial process, but also mining potential risks based on the functions of the devices and the adversaries' targets.

Table 5. The mined knowledge from the dual-security by PairRE.

Pattern	Mined Triples
$Virus \xrightarrow{\texttt{Use}} AttackPattern$	$Stuxnet \xrightarrow{\texttt{Use}} I/OImage$
	$LockerGoga \xrightarrow{\texttt{Use}} LossOfView$
	$Industroyer \xrightarrow{\texttt{Use}} DataDestruction$
	$Nodeproperties \xrightarrow{\texttt{Use}} LateralToolTransfer$
$Action \xrightarrow{\texttt{Mitigate}} AttackPattern$	$Encrypt \xrightarrow{\texttt{Mitigate}} Network$
	$NetworkSegmentation \xrightarrow{\texttt{Mitigate}} BruteForceI/O$
	$NetworkAllowlists \xrightarrow{\texttt{Mitigate}} BlockCommandMessage$
$Controller \xrightarrow{\texttt{control}} Device$	$1021FV0101 \xrightarrow{\texttt{control}} 1021PdIC0101$
	$1031HV0051 \xrightarrow{\texttt{control}} GasifierLevel$
	$Loadcontroller \xrightarrow{\texttt{control}} GasifierPressure$
	$Levelcontroller \xrightarrow{\texttt{control}} GasifierWaterFlow$
	$Circuit1031LIC0001 \xrightarrow{\texttt{control}} 1031LV0001A$
$Device \xrightarrow{\texttt{reaction}} ReactionType$	$Reboiler(F\text{-}2102) \xrightarrow{\texttt{reaction}} Heat$
	$Pump(P\text{-}2103A/B) \xrightarrow{\texttt{reaction}} Supercharge$
	$Tower(T\text{-}2568A/B) \xrightarrow{\texttt{reaction}} Desulfurization$

Table 6. Potential risks in crack hydrogenation and Industroyer attacks.

Device	Mined Risk
D-2106	$Industroyer \xrightarrow{\texttt{Use}} Remote\ System\ Discovery \xrightarrow{\texttt{Attack}} D\text{-}2106$
	$Industroyer \xrightarrow{\texttt{Use}} Device\ Restart/Shutdown \xrightarrow{\texttt{Attack}} D\text{-}2106$
	$Industroyer \xrightarrow{\texttt{Use}} Manipulation\ of\ Control \xrightarrow{\texttt{Attack}} D\text{-}2106$
D-2104	$Industroyer \xrightarrow{\texttt{Use}} Remote\ System\ Discovery \xrightarrow{\texttt{Attack}} D\text{-}2104$
	$Industroyer \xrightarrow{\texttt{Use}} Connection\ Proxy \xrightarrow{\texttt{Attack}} D\text{-}2104$
	$Industroyer \xrightarrow{\texttt{Use}} Denial\ of\ Control \xrightarrow{\texttt{Attack}} D\text{-}2104$
	$Industroyer \xrightarrow{\texttt{Use}} Network\ Connection\ Enumeration \xrightarrow{\texttt{Attack}} D\text{-}2104$
K-2102A/B	$Industroyer \xrightarrow{\texttt{Use}} Remote\ System\ Discovery \xrightarrow{\texttt{Attack}} K\text{-}2102A/B$
	$Industroyer \xrightarrow{\texttt{Use}} Block\ Serial\ COM \xrightarrow{\texttt{Attack}} K\text{-}2102A/B$
	$Industroyer \xrightarrow{\texttt{Use}} Service\ Stop \xrightarrow{\texttt{Attack}} K\text{-}2102A/B$
	$Industroyer \xrightarrow{\texttt{Use}} Block\ Command\ Message \xrightarrow{\texttt{Attack}} K\text{-}2102A/B$
	$Industroyer \xrightarrow{\texttt{Use}} Loss\ of\ Protection \xrightarrow{\texttt{Attack}} K\text{-}2102A/B$

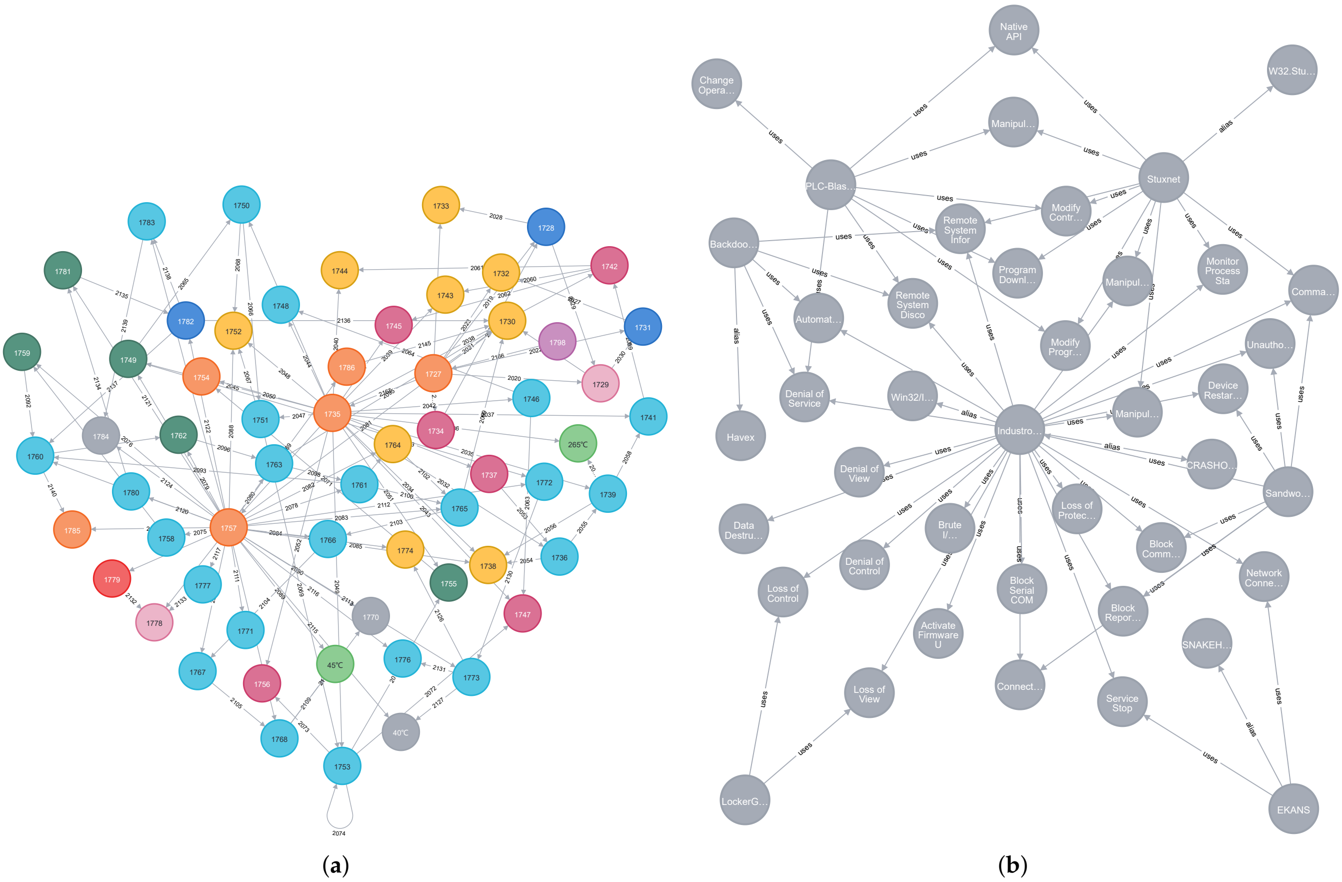

Figure 2. The part of dual-security KG. (**a**) Crack hydrogenation process KG. (**b**) Industroyer attacks KG.

5. Conclusions

In this paper, we begin with constructing a dual-security knowledge graph, which innovatively integrates industrial process security information, a domain often overlooked by researchers, into the cybersecurity knowledge graph, a domain that has garnered widespread attention. Subsequently, we employ an embedding-based method (RairRE) capable of learning the underlying semantics in the knowledge graph, allowing us to predict the occurrence of novel cybersecurity attack patterns or potential security issues in the industrial process. The experimental results suggest that this method can effectively mine potential risks in some industrial process. However, the overall potential risk prediction accuracy on the dual-KG is not very satisfactory, as existing models struggle to comprehensively grasp the knowledge embedded in both aspects of the whole dataset.

In light of these insights, our proposed model currently used has the potential for improved performance in fully tapping into the rich, interrelated data contained within the knowledge graph.

Our future endeavors will be dedicated to the development of a more tailored dual-security mining model, specifically designed to extract and integrate all knowledge with greater precision from both the realms of cyber and industrial process security. The ultimate goal is to enhance the accuracy of predicting potential risks in industrial processes.

Author Contributions: Conceptualization, L.W. and H.L.; methodology, W.W.; software, L.W. and H.L; validation, L.W., G.X. and Y.W.; formal analysis, W.W.; investigation, G.X.; resources, Y.W.; data curation, Y.W.; writing—original draft preparation, L.W.; writing—review and editing, H.L.; visualization, L.W.; supervision, Y.W.; project administration, W.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC No.2021YFB2012400).

Data Availability Statement: Source codes and desensitized datasets are available in the GitHub repository, and the link is https://github.com/Faker-lz/Dual-safety-knowledge-graph-completion-for-process-industry.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee, W.; Weekman, V.W., Jr. Advanced control practice in the chemical process industry: A view from industry. *AIChE J.* **1976**, *22*, 27–38. [CrossRef]
2. Lu, H.; Guo, L.; Azimi, M.; Huang, K. Oil and Gas 4.0 era: A systematic review and outlook. *Comput. Ind.* **2019**, *111*, 68–90. [CrossRef]
3. Schrotenboer, A.H.; Veenstra, A.A.; uit het Broek, M.A.; Ursavas, E. A Green Hydrogen Energy System: Optimal control strategies for integrated hydrogen storage and power generation with wind energy. *Renew. Sustain. Energy Rev.* **2022**, *168*, 112744. [CrossRef]
4. Alanen, J.; Linnosmaa, J.; Malm, T.; Papakonstantinou, N.; Ahonen, T.; Heikkilä, E.; Tiusanen, R. Hybrid ontology for safety, security, and dependability risk assessments and Security Threat Analysis (STA) method for industrial control systems. *Reliab. Eng. Syst. Saf.* **2022**, *220*, 108270. [CrossRef]
5. Corallo, A.; Lazoi, M.; Lezzi, M. Cybersecurity in the context of industry 4.0: A structured classification of critical assets and business impacts. *Comput. Ind.* **2020**, *114*, 103165. [CrossRef]
6. Noy, N.F.; McGuinness, D.L. *Ontology Development 101: A Guide to Creating Your First Ontology*; Stanford Knowledge Systems Laboratory Technical Report KSL-01-05; Knowledge Systems Laboratory: Stanford, CA, USA, 2001.
7. Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data—The Story So Far. *Int. J. Semant. Web Inf. Syst.* **2009**, *79*, 637–638. [CrossRef]
8. Yu, D.; Zhu, C.; Yang, Y.; Zeng, M. Jaket: Joint pre-training of knowledge graph and language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February–1 March 1 2022; Volume 36, pp. 11630–11638.
9. Lin, Q.; Mao, R.; Liu, J.; Xu, F.; Cambria, E. Fusing topology contexts and logical rules in language models for knowledge graph completion. *Inf. Fusion* **2023**, *90*, 253–264. [CrossRef]
10. Bakhshi, M.; Nematbakhsh, M.; Mohsenzadeh, M.; Rahmani, A.M. SParseQA: Sequential word reordering and parsing for answering complex natural language questions over knowledge graphs. *Knowl.-Based Syst.* **2022**, *235*, 107626. [CrossRef]
11. Gogleva, A.; Polychronopoulos, D.; Pfeifer, M.; Poroshin, V.; Ughetto, M.; Martin, M.J.; Thorpe, H.; Bornot, A.; Smith, P.D.; Sidders, B.; et al. Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat. Commun.* **2022**, *13*, 1667. [CrossRef]
12. Yang, Y.; Huang, C.; Xia, L.; Li, C. Knowledge graph contrastive learning for recommendation. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 1434–1443.
13. Wu, X.; Duan, J.; Pan, Y.; Li, M. Medical knowledge graph: Data sources, construction, reasoning, and applications. *Big Data Min. Anal.* **2023**, *6*, 201–217. [CrossRef]
14. Santos, A.; Colaço, A.R.; Nielsen, A.B.; Niu, L.; Strauss, M.; Geyer, P.E.; Coscia, F.; Albrechtsen, N.J.W.; Mundt, F.; Jensen, L.J.; et al. A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.* **2022**, *40*, 692–702. [CrossRef]
15. Li, M.M.; Huang, K.; Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nat. Biomed. Eng.* **2022**, *6*, 1353–1369. [CrossRef] [PubMed]
16. Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. In Proceedings of the Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013.
17. Yang, B.; tau Yih, W.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. *arXiv* **2014**, arXiv:1412.6575.
18. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2d knowledge graph embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
19. Balazevic, I.; Allen, C.; Hospedales, T. TuckER: Tensor Factorization for Knowledge Graph Completion. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5185–5194. [CrossRef]
20. Liu, K.; Wang, F.; Ding, Z.; Liang, S.; Yu, Z.; Zhou, Y. Recent Progress of Using Knowledge Graph for Cybersecurity. *Electronics* **2022**, *11*, 2287. [CrossRef]
21. Rastogi, N.; Dutta, S.; Gittens, A.; Zaki, M.J.; Aggarwal, C. TINKER: A framework for Open source Cyberthreat Intelligence. In Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Wuhan, China, 9–11 December 2022; pp. 1569–1574. [CrossRef]
22. Lin, S.C.; Tseng, S.S. Constructing detection knowledge for DDoS intrusion tolerance. *Expert Syst. Appl.* **2004**, *27*, 379–390. [CrossRef]

23. Narayanan, S.N.; Ganesan, A.; Joshi, K.; Oates, T.; Joshi, A.; Finin, T. Early Detection of Cybersecurity Threats Using Collaborative Cognition. In Proceedings of the 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), Philadelphia, PA, USA, 18–20 October 2018; pp. 354–363. [CrossRef]
24. Gao, P.; Shao, F.; Liu, X.; Xiao, X.; Qin, Z.; Xu, F.; Mittal, P.; Kulkarni, S.R.; Song, D. Enabling efficient cyber threat hunting with cyber threat intelligence. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; IEEE: New York, NY, USA, 2021; pp. 193–204.
25. Kiesling, E.; Ekelhart, A.; Kurniawan, K.; Ekaputra, F. The SEPSES knowledge graph: An integrated resource for cybersecurity. In Proceedings of the International Semantic Web Conference, Auckland, New Zealand, 26–30 October 2019; Springer: Cham, Switzerland, 2019; pp. 198–214.
26. Garrido, J.S.; Dold, D.; Frank, J. Machine learning on knowledge graphs for context-aware security monitoring. In Proceedings of the 2021 IEEE International Conference on Cyber Security and Resilience (CSR), Virtual Conference, 26–28 July 2021; IEEE: New York, NY, USA, 2021; pp. 55–60.
27. Li, J.; Liu, Y.; Gu, L. DDoS attack detection based on neural network. In Proceedings of the 2010 2nd International Symposium on Aware Computing, Tainan, Taiwan, 1–4 November 2010; IEEE: New York, NY, USA, 2010; pp. 196–199.
28. Li, X.; Lyu, M.; Wang, Z.; Chen, C.H.; Zheng, P. Exploiting knowledge graphs in industrial products and services: A survey of key aspects, challenges, and future perspectives. *Comput. Ind.* **2021**, *129*, 103449. [CrossRef]
29. Chen, Y.; Qian, T. Relation constrained attributed network embedding. *Inf. Sci.* **2020**, *515*, 341–351. [CrossRef]
30. Gao, J.; Li, X.; Xu, Y.E.; Sisman, B.; Dong, X.L.; Yang, J. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* **2019**, *12*, 1679–1691. [CrossRef]
31. Liu, W.; Liu, J.; Wu, M.; Abbas, S.; Hu, W.; Wei, B.; Zheng, Q. Representation learning over multiple knowledge graphs for knowledge graphs alignment. *Neurocomputing* **2018**, *320*, 12–24. [CrossRef]
32. Tay, Y.; Tuan, L.A.; Phan, M.C.; Hui, S.C. Multi-task neural network for non-discrete attribute prediction in knowledge graphs. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1029–1038.
33. Long, J.; Chen, Z.; He, W.; Wu, T.; Ren, J. An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. *Appl. Soft Comput.* **2020**, *91*, 106205. [CrossRef]
34. Wang, Q.; Zou, D.; Ge, L. Multi-integrated Reform for the Course of Data Structure. In Proceedings of the 15th International Conference on Computer Science & Education (ICCSE), Delft, The Netherlands, 18–22 August 2020; pp. 9–13. [CrossRef]
35. Zhang, Y.; Xu, H.; Zhang, X.; Wu, X.; Yang, Z. TRFR: A ternary relation link prediction framework on Knowledge graphs. *Ad Hoc Netw.* **2021**, *113*, 102402. [CrossRef]
36. Wang, Q.; Hao, Y. ALSTM: An attention-based long short-term memory framework for knowledge base reasoning. *Neurocomputing* **2020**, *399*, 342–351. [CrossRef]
37. Ai, Q.; Zhang, Y.; Bi, K.; Croft, W.B. Explainable product search with a dynamic relation embedding model. *ACM Trans. Inf. Syst. (TOIS)* **2019**, *38*, 1–29. [CrossRef]
38. Wang, Q.; Ji, Y.; Hao, Y.; Cao, J. GRL: Knowledge graph completion with GAN-based reinforcement learning. *Knowl.-Based Syst.* **2020**, *209*, 106421. [CrossRef]
39. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; Volume 28.
40. Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; Bouchard, G. Complex embeddings for simple link prediction. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 2071–2080.
41. Yang, F.; Yang, Z.; Cohen, W.W. Differentiable learning of logical rules for knowledge base reasoning. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
42. Sadeghian, A.R.; Armandpour, M.; Ding, P.; Wang, D.Z. DRUM: End-to-End Differentiable Rule Mining on Knowledge Graphs. *arXiv* **2019**, arXiv:1911.00055.
43. Chao, L.; He, J.; Wang, T.; Chu, W. Pairre: Knowledge graph embeddings via paired relation vectors. *arXiv* **2020**, arXiv:2011.03798.
44. Sun, Z.; Deng, Z.; Nie, J.Y.; Tang, J. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. *arXiv* **2018**, arXiv:1902.10197.
45. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
46. Zhang, Z.; Cai, J.; Zhang, Y.; Wang, J. Learning hierarchy-aware knowledge graph embeddings for link prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 3065–3072.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.