Knowledge Graph on Cybersecurity: A Survey

Maman Sani ABOUBACAR  
University of Montpellier  
Montpellier, France  
maman-sani.aboubacar-djibo@lirmm.fr

Arnaud CASTELL TORT  
University of Montpellier  
Montpellier, France  
castelltort@lirmm.fr

Anne LAURENT  
University of Montpellier  
Montpellier, France  
laurent@lirmm.fr

ABSTRACT

Over the past decade, social networks and discussion forums have become valuable sources for open source cyber security intelligence. However, sorting out relevant and irrelevant information is very complex, as information comes from a variety of sources and in different formats. Different approaches exist for extracting and representing this information. In this review we present a comparative study on knowledge base approaches that consist in extracting relevant cyber security information and representing it in a graphical format for a better understanding and readability of the data. This comparative study reviews the various cyber security knowledge base construction works by identifying the potentials and limitations of each study.

Keywords:  
Cybersecurity, Information extraction, Ontology, Knowledge Graph.

1 INTRODUCTION

Cyber security is a fairly vast field that is evolving at a significant speed. Analysts and cyber security professionals have a critical need for the latest information to cope with this evolution. This information is usually generated by different tools, sensors or even available on the Web in a structured and unstructured format. The unification and organization of this information will allow cyber security professionals to have better visibility and situational awareness.

Data representation approaches based on graphs can be a solution to this problem. Indeed, they allow the integration, organization and representation of data in a readable and understandable way for both the machine and the analysts. Studies have been conducted on this issue in the field of cyber security [2, 10, 15, 16]. However, few of these studies [2] include a comparative study on the representation of graph-based data on cyber security.

The aim of this paper is to expose the various works dealing with the representation and integration of cyber security data. First, we present the need to use graph-based approaches, as well as the different challenges associated with these approaches. In a second step, we present the different steps of the creation of a knowledge base as well as the works dealing with this topic. At the end we present a synthetic comparison of these studies.

The rest of this paper is presented as follows: Section 2 presents the interest of using graphs. Section 3 presents the steps of the construction of a knowledge base. Section 4 presents the studies that have been conducted on the construction of knowledge graphs for cyber security data. Section 5 presents the conclusion and discussions.

2 WHY GRAPHS?

By definition a graph is a data structure made up of entities linked together by relationships. Entities are represented by nodes and relationships by arcs. The knowledge graph is a semantic knowledge base that allows to describe the semantics of information sources and thus make the content explicit. It is a term that was introduced by Google in 2012 as part of improving user experience. The goal is to allow users to solve their queries without having to navigate to other websites to access critical information [4]. Although knowledge representation is not new, it has gained popularity in recent decades through its use in artificial intelligence applications.

There are several advantages to using these approaches:

- Interdependent nature of the data: entities are often linked to each other and have dependencies;
- Powerful representation: graphs naturally represent the interdependencies by introducing links between entities. This allows to efficiently capture the correlation between them;
- Relational nature of the problem areas: the nature of the anomalies, vulnerabilities or even attacks can be relational. For example, a vulnerability can affect a piece of software, and at the same time the whole system.

2.1 Challenges specific to data and cyber security issues

Data challenges such as velocity, volume, variety and quantity can also be applied to graph-based data.

- Dynamism and scalability: with an explosion of user and machine-generated data in real time, dynamism and scalability are challenges in graph data management [1];
- Complexity: the available data is rich and complex in terms of content. A data can contain several pieces of information, the extraction of the totality of this information and its exploitation is often quite complex.

In addition to these challenges, there are also other challenges related to cyber security data.

- Lack and noise of labeled data: data labeling is a great challenge in extracting and representing data in graphical format. In cyber security there is a great lack of structured and publicly available data on attacks or incidents, unlike vulnerability databases[1]. Machine learning techniques for automatic data labeling exist, but produce noise or omit certain concepts of cyber security [3]. For example, the absence of true labeled data, i.e. field truth data, makes it difficult to evaluate techniques for detecting incidents or anomalies related to cyber security.

1https://nvd.nist.gov/
• Class imbalance: incidents, vulnerabilities or anomalies are often very rare in the data collected. Also with labeling errors, some classes (concepts) may be under-represented, which can make it difficult for machine learning approaches to predict this class. This is why this issue must be carefully taken into account.

3 STEPS OF CREATING KNOWLEDGE GRAPHS

The creation of knowledge graphs can be done by several approaches. However, in recent decades, the ontology-based approaches are much more illustrated by the results they provide. This is why, in this review, we will present only those works oriented towards this approach. For the creation of a knowledge base, a fairly generic pipeline is most often used. This pipeline is shown in the figure 1.

3.1 Data sources

Cyber security data is usually generated by different tools, sensors or even available on the Web in a structured and unstructured format. Indeed, over the last decade, social networks and discussion forums have become valuable sources for cyber security information. With the right tools and methods, these information sources can be identified, explored and then exploited to obtain actionable information about cyber threats.

To this end, various studies have been conducted on the collection of cyber security data. This is the case, for example, of the work of Macdonal et al. [9] which focuses on analyzing the communications of hackers on forums to identify potential threats against critical infrastructures using automated analysis tools.

3.2 Named entities extraction

Entity extraction consists in extracting from annotated data cyber security concepts for subsequent exploitation. The process consists of extracting cyber security concepts from the extracted data and linking them together through an ontology by relationships so as to form a knowledge graph on which it is possible to carry out reasoning.

It is also possible to link these concepts to external knowledge bases such as: DBpedia, Wikidata, etc. in order to enrich the data.

3.3 Ontologies mapping

Ontologies are used in this pipeline to enable the creation of relationships between entities. Indeed, ontologies are semantic data models that define the types of things that exist in a domain and the properties that can be used to describe them. They are an exclusive means of representing and communicating facts and relationships. An ontology consists of a set of classes with attributes and relationships between instances of different classes. Figure 2 presents a STIX [14] cyber security ontology that models the relationship between an attack campaign, the instigator of the campaign, the target of the attack, and the tools used to carry out the attack.

3.4 Knowledge base

The knowledge base is a semantic knowledge graph that describes the semantics of information sources and thus makes the content more explicit for analysts and professionals in the field. A knowledge graph is the result of associating the concepts of a domain with a data representation model, namely here, ontologies. In our case, the knowledge graph can be defined as the instantiation of cyber security concepts extracted in a dedicated cyber security ontology.

4 COMPARATIVE STUDY OF WORK ON THE CONSTRUCTION OF KNOWLEDGE BASE RELATED TO CYBER SECURITY

Several works in the literature have focused on the creation of cyber security knowledge bases using ontologies. This is the case of Undercoffer et al. [16] who introduced the first cyber security ontology for the modeling of the Intrusion Detection System (IDS). Syed et al [15] proposed the Unified Cyber Security Ontology, an extension of the IDS ontology. This ontology has the advantage of being linked to several external cyber security knowledge bases such as: CVE\(^2\), CWE\(^3\), STUCCO\(^4\), STIX\(^5\).

Systems that create or integrate data into cyber security knowledge graphs have been proposed in the literature. This is the case, for example, of the recent work of Taneeya et al. [12] which deals

---

\(^2\)https://cve.mitre.org/
\(^3\)https://cwe.mitre.org/
\(^4\)https://github.com/stucco-archive/ontology
\(^5\)https://oasis-open.github.io/cti-documentation/
with the extraction of cyber security information from cyber security blogs, to integrate them into a knowledge base and reason on the data to create new knowledge.

Pingle et al. [11] proposed a system to create semantic triplets from cyber security data, using deep learning approaches to extract possible relationships. The semantic triplets generated by the system can be used to make assertions in a cyber security knowledge graph. They can also be extracted from the knowledge graph to be exploited by analysts in their decision making on cyber attacks.

Iannacone et al. [5] have developed an ontology for a Cyber Security Knowledge Graph database. This ontology integrates information from various structured and unstructured data sources and includes all relevant concepts in the field of cyber security.

Shang et al. [13] proposed a framework that allows the integration of cyber security information extracted from texts into a knowledge base. To do so, they created a vulnerability-centered ontology and formed a model for extracting named entities related to cyber security using statistical rules and models such as conditional random fields.

Mittal et al. proposed CyberTwitter [10] a system that uses Twitter as a data source to study vulnerabilities related to cyber security published on the social network. After data collection, concepts related to cyber security are extracted through the Security Vulnerability Concept Extractor (SVCE) [7] tool they have developed for this purpose. The extracted concepts are then presented in RDF format through the Unified Cyber Security Ontology (UCO) [15]. Subsequently SWRL (Semantic Web Rule Language) rules are used to reason about the extracted concepts to issue alerts to security analysts.

Jia et al. [6] proposed an approach for building a knowledge base on cyber security using inference rules based on a path sorting algorithm. The basic principle of the path sorting algorithm is to use the path connecting two entities as a characteristic to predict the relationship between the two entities. By using path sorting algorithms for a given relationship, it is possible to determine whether a relationship exists between the two entities.

### 4.1 Synthetic comparison of the work presented

The table presents a comparative study of the work presented. For each study, five arguments are considered. The data source, which specifies whether the data being exploited comes from a single or multiple source. The NER (Named Entity Recognition) to specify if the study to use entity extraction approaches. Ontology, if the study used ontologies for the creation of its knowledge base. External KG, if the study used external knowledge bases such as: DBpedia, Wikidata, etc. to enrich its knowledge base. Finally, the field of application of the study.

From this table, it appears that all the studies presented, except the work of Syed et al. [15] and Undercoffer et al. [16], use named entity extraction approaches for the extraction of cyber security concepts. Indeed, Syed et al. use cyber security concepts already existing in the literature. It also appears that only the work of Syed et al. and Mittal et al. use external knowledge to link the extracted concepts to this knowledge base in order to enrich the data.

However, none of this work has dealt with scalability and the instantaneous aspect of the data stored in the knowledge base. In addition, none of the works presented dealt with the verification of the quality of the collected data. Indeed, it is more than important in a field such as cyber security to ensure the quality of the information to be fed back to analysts and professionals in the field at the risk of making bad decisions.

### 5 CONCLUSION

In this review, we have provided a critical overview of the various works on building a knowledge base for cyber security. We have clearly defined the main challenges as well as the different stages of setting up such a database. We then carried out a comparative study of the different works that have dealt with this issue.

The analysis of the studies presented allowed us to note that only the work of Mittal et al. ticked all the boxes. Indeed, the use of external knowledge allowed the authors to enrich their data and obtain a better result. However, none of this work deals with data quality verification. Moreover, since data sources are most often informal sources (blogs, social networks, etc.), it will be important to take this issue into account. The latter may be a lead for future work, in addition to the knowledge base scalability lead and the snapshot data collection lead. For the latter, the work of Sceller et al. [8] can serve as a basis.

### REFERENCES


Table 1: Comparative table of studies on building a knowledge base for cyber security

<table>
<thead>
<tr>
<th>Paper</th>
<th>Data source</th>
<th>NER</th>
<th>Ontology</th>
<th>External KG</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undercoffer et al. [16]</td>
<td>Multiple</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>All cybersecurity domain</td>
</tr>
<tr>
<td>Syed et al. [15]</td>
<td>Multiple</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>All cybersecurity domain</td>
</tr>
<tr>
<td>Taneeya et al. [12]</td>
<td>blogs</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>Cybersecurity concepts extraction</td>
</tr>
<tr>
<td>Pingle et al. [11]</td>
<td>Multiples</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>All cybersecurity domain</td>
</tr>
<tr>
<td>Iannacone et al. [5]</td>
<td>Multiples</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>All cybersecurity domain</td>
</tr>
<tr>
<td>Shang et al. [13]</td>
<td>Multiples</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>Vulnerability detection</td>
</tr>
<tr>
<td>Jia et al. [6]</td>
<td>Multiples</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>Vulnerability detection</td>
</tr>
<tr>
<td>Mittal et al. [10]</td>
<td>Multiples</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Cybersecurity event detection on Twitter</td>
</tr>
</tbody>
</table>